Academic Peer Effects with Different Group Assignment Rules: Residential Tracking versus Random Assignment[*]

Robert Garlick[†]

December 29, 2013

Abstract

I study the relative academic performance of students tracked or randomly assigned to South African university dormitories. This advances the literature on peer effects under different peer group assignment policies and on optimal group design. I find that tracking reduces low-scoring students' GPAs but has little effect on high-scoring students. The net effect is to reduce mean GPA and increase the spread or inequality of GPA. I also directly estimate peer effects using random variation in dormitory peer groups. I find that own and peer characteristics are substitutes in GPA production and that peer effects are considerably stronger within than across race groups. I finally explore whether peer effects estimated under random assignment can predict

[*] This paper is a revised version of the first chapter of my dissertation. I am extremely grateful to my advisors David Lam, Jeff Smith, John DiNardo, Manuela Angelucci, and Brian Jacob for their extensive guidance and support. I thank Raj Arunachalam, Emily Beam, John Bound, Tanya Byker, Julian Cristia, Susan Godlonton, Andrew Goodman-Bacon, Italo Gutierrez, Claudia Martinez, David Slusky, Rebecca Thornton, and Dean Yang for helpful comments on earlier drafts of the paper, as well as conference and seminar participants at Chicago Harris School, Columbia, Columbia Teachers College, CSAE 2012, Cornell, Duke, EconCon 2012, ESSA 2011, Harvard Business School, LSE, Michigan, MIEDC 2012, Michigan State, NEUDC 2012, Northeastern, Notre Dame, PacDev 2011, SALDRU, Stanford SIEPR, SOLE 2012, UC Davis, the World Bank, and Yale School of Management, I received invaluable assistance with student data and institutional information from several staff members at the University of Cape Town, particularly Jane Hendry, Josiah Mavundla, and Charmaine January. I acknowledge financial support from the Gerald R. Ford School of Public Policy and Horace H. Rackham School of Graduate Studies at the University of Michigan. All errors are my own.

[†] Development Research Group at The World Bank and Duke University Department of Economics. E-mail: rob.garlick@gmail.com

the effects of tracking. The quantitative predictions are sensitive to model specification choices over which neither economic theory nor model selection tests provide clear guidance.

Keywords: education, peer effects, tracking

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.[1] Academic peer effects play a role in both empirical and theoretical research on alternative ways of organizing students into classrooms and schools.[2] Most studies focus on the effect of assignment to or selection into different peer groups for a given group assignment or selection process.[3]

This paper advances the peer effects literature by asking a subtly different question: What are the relative effects of different group assignment policies, tracking and randomization, on the distribution of student outcomes? This contributes to a small but growing empirical literature on

[1] Manski (1993) lays out the identification challenge in studying peer effects: do correlated outcomes within peer groups reflect peer effects – causal relationships between students' outcomes and their peers' outcomes or pre-determined characteristics – or correlated unobserved pre-determined characteristics or institutional factors. Many papers address this challenge using randomized or controlled variation in peer group composition; peer effect have been documented on standardized test scores (Hoxby, 2000), college GPAs (Sacerdote, 2001), college entrance examination scores (Ding and Lehrer, 2007), cheating (Carrell, Malmstrom, and West, 2008), job search (Marmaros and Sacerdote, 2002), and major choices (Di Giorgi, Pellizzari, and Redaelli, 2010). Estimated peer effects may be sensitive to the definition of peer groups (Foster, 2006) and the measurement of peer characteristics (Stinebrickner and Stinebrickner, 2006).

[2] See Arnott (1987) and Duflo, Dupas, and Kremer (2011) on classroom tracking, Benabou (1996) and Kling, Liebman, and Katz (2007) on neighborhood segregation, Epple and Romano (1998) and Hsieh and Urquiola (2006) school choice and vouchers, and Angrist and Lang (2004) on school integration.

[3] See Sacerdote (2011) for a recent review that reaches a similar conclusion.

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 with few pecuniary costs. Such low cost education interventions are particularly attractive for developing country education systems that often face serious resource shortages.

Studying peer effects under one group assignment policy provides limited information about predicted outcomes under a different group assignment policy. Consider the comparison between random group assignment and academic tracking, in which students are assigned to academically homogeneous groups. First, tracking generates groups consisting of only high- or only low-performing students, which are unlikely to be observed under random assignment. Strong assumptions are required to extrapolate outcomes under tracking from small observed cross-group differences in mean scores under random assignment.[4] Second, student outcomes may depend on multiple dimensions of their peer group characteristics. Econometric models estimated under the status quo assignment policy may omit characteristics that would be important under other assignment policies. For example, within-group variance in peer characteristics may appear unimportant in homogeneous groups under tracking but matter in heterogeneous groups under random assignment. Third, peer effects will not be policy-invariant if students' interaction patterns change with group assignment policies. If, for example, students prefer homogeneous social groups, then the intensity of within-group interaction will be higher under tracking than random assignment. Peer effects estimated in "low-intensity" randomly assigned groups will then understate the strength of peer effects in "high-intensity" tracked groups.

I study peer effects under two different group assignment policies, using

[4] Random assignment may generate all possible types of groups if the groups are sufficiently small and group composition can be captured by a small number of summary statistics. I thank Todd Stinebrickner for this observation.

a natural experiment at the University of Cape Town in South Africa. First year students at the university were tracked into dormitories up to 2005 and randomly assigned from 2006 onward. This generated spatial peer groups that were respectively homogeneous or heterogeneous in baseline academic performance. I contrast the distribution of academic outcomes in the first year of university under the two policies. I use non-dormitory students as a control group in a difference-in-differences design to remove time trends and cohort effects.

I show that tracking leads to lower and more unequally distributed grade point averages (GPAs) than random assignment. Mean GPA is 0.12 standard deviations lower under tracking. Low-scoring students perform substantially worse under tracking than random assignment, while high-scoring students' GPAs are approximately equal under the two policies. I adapt results from the econometric theory literature to estimate the effect of tracking on academic inequality. Standard measures of inequality are substantially higher under tracking than random assignment. I explore a variety of alternative explanations for these results: time-varying student selection into dormitory or non-dormitory status, differential time trends in student performance between dormitory and non-dormitory students, limitations of GPA as an outcome measure, and direct effects of dormitory assignment on GPAs. I conclude that peer effects continue to play an important role after accounting for these factors.

I then use randomly assigned dormitory-level peer groups to estimate directly the effect of living with marginally higher- or lower-scoring peers. I find that students' GPAs are increasing in the mean prior academic performance of their peers. Low-scoring students benefit more than high-scoring students from living with high-scoring peers. Equivalently, own and peer academic performance are substitutes in GPA production. The results from the cross-dormitory and cross-policy analyses are qualitatively consistent. Peer effects estimated under random assignment can quantitatively predict features of the GPA distribution under tracking. However, the predictions are very sensitive to model specification choices over which economic theory and statistical model selection criteria provide little guidance. This prediction challenge reinforces the value of cross-policy evidence on peer effects.

I go on to explore the mechanisms driving the estimated peer effects. I find evidence that dormitory-level peer effects exist between students only if they are also socially proximate. Direct interaction between students thus appears necessary to generate peer effects. However, the relevant form of the interaction does not appear to be direct academic collaboration. Peer effects may therefore operate through spillovers on time use or through transfers of tacit knowledge such as study skills or norms about how to interact with faculty.

This paper makes four contributions. First, I contribute to the literature on optimal group design in the presence of peer effects or spillovers. Early work by Arnott (1987) and Benabou (1996) showed that the effect of peers' characteristics on agents' outcomes influence the optimal assignment policies to classroom or neighborhood peer groups.[5] Empirical evidence on this topic is extremely limited. My paper most closely relates to Carrell, Sacerdote, and West (2013), who use peer effects estimated under random group assignment to derive an "optimal" assignment policy. Mean outcomes are, however, worse under this policy than under random assignment. They ascribe this result to changes in the structure of within-group student interaction induced by the policy change. Bhattacharya (2009) and Graham, Imbens, and Ridder (2013) establish assumptions under which peer effects estimated under random group assignment can predict outcomes under a new group assignment policy. The assumptions are strong: that peer effects are policy-invariant, that no out-of-sample extrapolation is required, and that relevant peer characteristics have low dimention. These results emphasize the difficulty of using peer effects estimated under one group assignment policy to learn about peer effects under other group assignment policies.

Second, I contribute to the literature on peer effects in education.[6] I show that student outcomes are affected by randomly living with higher-scoring peers and by changes in the peer group assignment policy. Both

[5] A closely related literature studies the efficiency implications of private schools and vouchers in the presence of peer effects (Epple and Romano, 1998; Nechyba, 2000).

[6] My work most closely relates to the empirical literature studying randomized or controlled group assignments. There are also related literatures on the theoretical foundations of peer effects models and on identifying peer effects in endogeneously formed peer groups (Blume, Brock, Durlauf, and Ioannides, 2011; Brock and Durlauf, 2001).

analyses show that low-scoring students are more sensitive to changes in peer group composition, implying that own and peer academic performance are substitutes in GPA production. This is the first finding of substitutability in the peer effects literature of which I am aware.[7] I find that peer effects operate almost entirely within race groups, suggesting that spatial proximity generates peer effects only between socially proximate students.[8] I also find that dormitory peer effects are not stronger within than across classes. An economics student, for example, is no more strongly affected by other economics students in her dormitory than by non-economics students in her dormitory. This suggests that peer effects do not operate mainly through direct academic collaboration but rather through channels such as time use or transfer of soft skills, consistent with Stinebrickner and Stinebrickner (2006).

Third, I contribute to the literature on academic tracking by isolating a peer effects mechanism. Most existing papers estimate the effect of school or classroom tracking relative to another assignment policy or of assignment to different tracks.[9] However, tracked and untracked units may differ on multiple dimensions: peer group composition, instructor behavior, and school resources (Betts, 2011; Brunello and Checchi, 2007; Figlio and Page, 2002). Isolating the causal effect of tracking on student outcomes via peer group composition, net of these other factors, requires strong assumptions in standard research designs. I study a setting where instruction does not differ across tracked and untracked students or across students in different tracks. Students living in different dormitories take classes together from the same instructors. While variation in dormitory-level characteristics might in prin-

[7] Hoxby and Weingarth (2006) provide a general taxonomy of peer effects other than the linear-in-means model studied by Manski (1993). Burke and Sass (2013), Cooley (2013), Hoxby and Weingarth (2006), Imberman, Kugler, and Sacerdote (2012) and Lavy, Silva, and Weinhardt (2012) find evidence of nonlinear peer effects.

[8] Hanushek, Kain, and Rivkin (2009) and Hoxby (2000) document stronger within- than across-race classroom peer effects.

[9] Betts (2011) reviews the economic evidence regarding the effects of tracking. Key papers include Argys, Rees, and Brewer (1996), Hanushek and Woessmann (2006), Meghir and Palme (2005), Pischke and Manning (2006), and Slavin (1987, 1990). A smaller literature studies the effect of assignment to different tracks in an academic tracking system (Abdulkadiroglu, Angrist, and Pathak, 2011; Ding and Lehrer, 2007; Pop-Eleches and Urquiola, 2013).

ciple affect student outcomes, my results are entirely robust to conditioning on these characteristics. I therefore interpret the negative average treatment effect of tracking, particularly on low-scoring students' outcomes, as operating through peer effects. Studying dormitories as assignment units limits the generalizability of my results but allows me to focus on one mechanism at work in school or classroom tracking. My findings are consistent with the results reported in Duflo, Dupas, and Kremer (2011). They find that tracked Kenyan students in first grade classrooms obtain higher average test scores than untracked students. They ascribe this to a combination of targeted instruction (positive effect for all students) and peer effects (positive and negative effects for high- and low-track students respectively).

Fourth, I make a methodological contribution to the study of peer effects and of academic tracking. These literatures strongly emphasize inequality considerations but generally do not measure the effect of different group assignment policies on inequality (Betts, 2011; Epple and Romano, 2011). I note that an inequality treatment effect of tracking can be obtained by comparing inequality measures for the observed distribution of outcomes under tracking and the counterfactual distribution of outcomes that would have been obtained in the absence of tracking. This counterfactual distribution can be estimated using standard methods for quantile treatment effects.[10] Firpo (2010) and, in a different context, Rothe (2010) establish formal identification, estimation, and inference results for inequality treatment effects. I use a difference-in-differences design to calculate the treatment effects of tracking net of time trends and cohort effects. I therefore integrate a nonlinear difference-in-differences model (Athey and Imbens, 2006) with an inequality treatment effects framework (Firpo, 2010). I also propose a conditional nonlinear difference-in-differences model in the online appendix that extends the original Athey-Imbens model. This extension accounts flexibly for time trends or cohort effects using inverse probability weighting (DiNardo, Fortin, and Lemiuex, 1996; Hirano, Imbens, and Ridder, 2003).

I outline the setting, research design, and data in section 2. I discuss the negative average effect of tracking on GPAs in section 3 and show how

[10] See Firpo (2007) and Heckman, Smith, and Clements (1997) for discussion on quantile treatment effects estimators.

this in concentrated on low-scoring students in high school graduation tests. In section 4, I discuss the effects of tracking on the entire distribution of GPAs, showing that the lower tail of the distribution is negatively affected while the upper tail is largely unaffected. I show in section 5 that this implies higher academic inequality. I go on to show in section 6 that students' GPAs are increasing in the mean high school graduation test score in their dormitory. Students with low test scores are particularly sensitive to peer group composition, while students with high test scores are largely unaffected. I present a framework to unify these results in section 7. The cross-policy and cross-dormitory results both indicate that own and peer characteristics are substitutes in GPA production. However, the effects of tracking predicted by the cross-dormitory peer effects are sensitive to specification choices over which economic theory and statistical model selection criteria provide weak guidance. In section 8 I report a variety of robustness checks to verify the validity of the research design used to identify the effects of tracking. I confirm that main results are robust to accounting for potential violations of the identifying assumptions. I outline the conditional nonlinear difference-in-differences model in appendix A.

2 Research Design

I study a natural experiment at the University of Cape Town (UCT) in South Africa, where first-year students are allocated to dormitories using either random assignment or academic tracking. UCT is a selective research university. During the time period I study, admissions decisions included an affirmative action component favouring students from low-income schools. The student population is thus relatively heterogeneous but not representative of South Africa.

Approximately half of the 3500-4000 first-year students live in university dormitories.[11] The dormitories provide accommodation, meals, and some organized social activities. Classes and instructors are shared across

[11] The mean dormitory size is 123 students and the 10^{th}, 50^{th}, and 90^{th} percentiles are 50, 112, and 216 students respectively. There are 14 dormitories open for the entire period of study, one that closes in 2006, and another that opens in 2007.

students from different dormitories and students who do not live in dormitories. Dormitory assignment therefore determines the set of residentially proximate peers but not the set of classroom peers. Students are normally allowed to live in dormitories for at most two years. They cannot change dormitories between their first and second year but can move into private accommodation. Dormitory assignment thus determines students' residential peer groups in their first year of university; whether these peer groups persist for a second year is an outcome of students' first year experience. Most students live in two-person rooms and the roommate assignment process varies across dormitories. I do not observe roommate assignments. The other half of the incoming first year students live in private accommodation, typically with family in the Cape Town metropolitan area.

Incoming students were tracked into dormitories up until the 2005 academic year. Tracking was based on a set of national, content-based high school graduation tests taken by all South African grade 12 students.[12] Students with high scores on this examination were assigned to different dormitories than students with low scores. The resultant assignments do not partition the distribution of test scores for three reasons. First, assignment incorporated loose racial quotas, so the threshold score for assignment to the top dormitory was higher for white than black students. Second, most dormitories were single-sex, creating pairs of female and male dormitories at each track. Third, late applicants for admission were waitlisted and assigned to the first available dormitory slot created by an admitted student withdrawing. A small number of high-scoring students thus appear in low-track dormitories and vice versa. These factors generate substantial overlap across dormitories' test scores.[13] However, the mean peer test score for a student in the top quartile of the high school test score distribution was still 0.93 standard deviations higher than for a student in the bottom quartile.

[12] These tests are set, graded, and moderated by a statutory body reporting to the Minister of Education. The tests are nominally criterion-referenced. Students are tested in six subjects. The university converts their letter grades into a single score used for admissions decisions. A time-invariant conversion scale is used to convert international students' A-level or International Baccalaureate scores into a comparable metric.

[13] The overlap is such that it is not feasible to use a regression discontinuity design to study the effect of assignment to higher- or lower-track dormitories. The first stage of such a design does not pass standard instrument strength tests.

From 2006 onward, incoming students were randomly assigned to dormitories. The policy change reflected concern by university administrators that tracking was inegalitarian and contributed to social segregation by income.[14] Assignment used a random number generator with *ex post* changes to ensure racial balance.[15] One small dormitory ($\approx 1.5\%$ of the sample) was excluded from the randomization. This dormitory charged lower fees but did not provide meals. Students could request to live in this dormitory and this resulted in a disproportionate number of low-scoring students under both tracking and randomization. My results are robust to excluding this dormitory.

The change in policy induced a substantial change in students' peer groups. Figure 1 shows how the relationship between students' own high school graduation test scores and their peers' test scores changed. For example, students in the top decile lived with peers who scored approximately 0.5 standard deviations higher under tracking than random assignment; students in the bottom decile lived with peers who scored approximately 0.4 standard deviations lower. This change provides the identifying variation I use to study the effect of tracking.

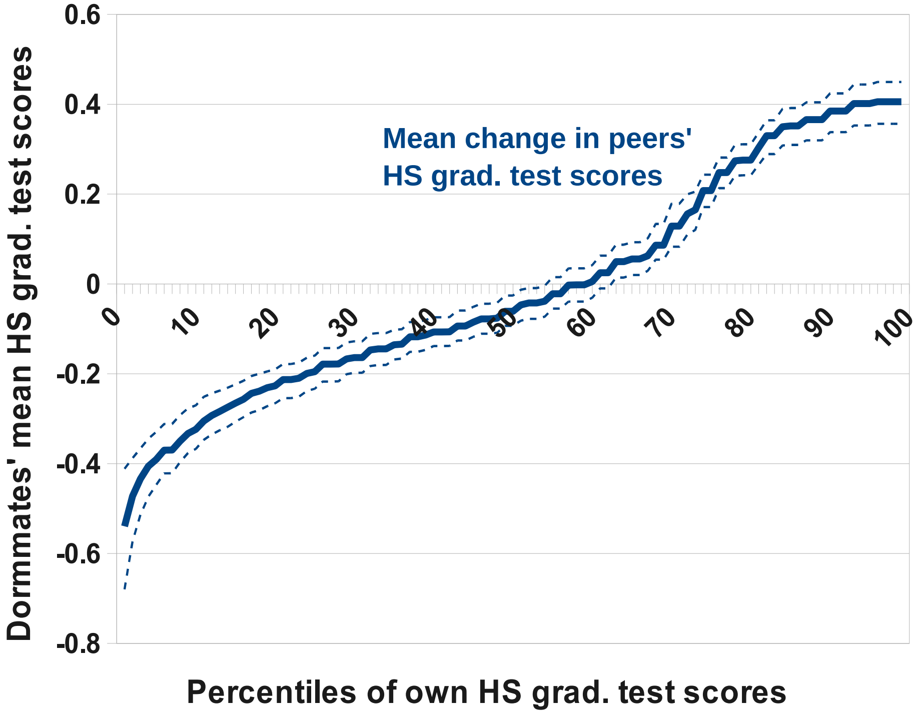
My research design compares the students' first year GPAs between the tracking period (2004 and 2005) and the random assignment period (2007 and 2008). I define tracking as the "treatment" even though it is the earlier policy.[16] I omit 2006 because first year students were randomly assigned to dormitories while second year students continued to live in the dormitories into which they had been tracked. GPA differences between the two periods may reflect cohort effects as well as peer effects. In particular, benchmarking tests show a downward trend in the academic performance of incoming first year students at South African universities over this time period (Higher Education South Africa, 2009). I therefore use a difference-in-differences design that compares the time change in dormitory students' GPAs with

[14] This discussion draws on personal interviews with the university's Director of Admissions and Director of Student Housing.

[15] There is no official record of how often changes were made. In a 2009 interview, the staff member responsible for assignment recalled making only occasional changes.

[16] Defining random assignment as the treatment necessarily yields point estimates with identical magnitude and opposite sign.

Figure 1: Effect of Tracking on Peer Group Composition



Notes: Figure is constructed by estimating a student-level local linear regression of mean dormitory high school test scores against students' own test scores. The regression is estimated separately for tracked and randomly assigned dormitory students and the difference is evaluated at each percentile of the high school grade distribution. The dashed lines show a 95% confidence interval constructed from 1000 replications of a percentile bootstrap.

the time change in non-dormitory students' GPAs over the same period:

$$GPA_{id} = \beta_0 + \beta_1 \text{Dorm}_{id} + \beta_2 \text{Track}_{id} + \beta_3 \text{Dorm}_{id} \times \text{Track}_{id} + \epsilon_{id} \qquad (1)$$

or, with covariates

$$\begin{aligned} GPA_{id} = \beta_0 + \beta_1 \text{Dorm}_{id} + \beta_2 \text{Track}_{id} + \beta_3 \text{Dorm}_{id} \times \text{Track}_{id} \\ + f\left(\vec{X}_{id}\right) + \vec{\mu}_d + \epsilon_{id} \end{aligned} \qquad (2)$$

where i and d index students and dormitories, $Dorm$ and $Track$ are indicator variables equal to 1 for students living in dormitories and for students enrolled in the tracking period, $f(\vec{X}_{id})$ is a function of students' demographic characteristics and high school graduation test scores,[17] and $\vec{\mu}_d$ is a vector of dormitory fixed effects. β_3 equals the average treatment effect of tracking on the tracked students under an "equal trends" assumption: that domitory and non-dormitory students would have experienced the same mean time change in GPAs if the assignment policy had remained constant. The difference-in-differences model identifies only a "treatment on the treated" effect; caution should be exercised in extrapolating this to non-dormitory students. Model 2 relaxes the equal trends assumption to hold conditional on student covariates and dormitory fixed effects. I also estimate models 1 and 2 with inverse probability weights that reweight each group of students to have the same distribution of covariates as the tracked dormitory students.[18] The validity of all of these designs rests on variants on the idea that dormitory students are an appropriate control group for non-dormitory students.

β_3 does not equal the average treatment effect of tracking on the tracked students if dormitory and non-dormitory students have different counterfactual GPA time trends. If the change in assignment policy affects students through channels other than peer effects, then β_3 recovers the correct treatment effect but its interpretation changes. I discuss these concerns in detail in section 8.

The data on students' demographic characteristics and high school test scores (reported in table 1) are broadly consistent with the assumption of equal time trends. Dormitory students have on average slightly higher and more dispersed scores than non-dormitory students on high school gradua-

[17] I report results with a quadratic $f(\cdot)$. The results are very similar when $f(\cdot)$ is linear or cubic.

[18] Unlike the regression-adjusted model 2, the reweighting estimators permit the treatment effect of tracking to vary across student covariates. This is potentially important in this study, where tracking is likely to have heterogeneous effects. However, the regression-adjusted and reweighted estimators reported in section 3 are very similar. DiNardo, Fortin, and Lemiuex (1996) and Hirano, Imbens, and Ridder (2003) discuss reweighting estimators with binary treatments. Reweighted difference-in-differences models are discussed in Abadie (2005) and Cattaneo (2010), who also derive the appropriate weights for estimating treatment-on-the-treated parameters. The reweighted and regression-adjusted model is identified under weaker assumptions than either the rewighted or regression-adjusted model (Robins and Rotnitzky, 1995).

Table 1: Summary Statistics and Balance Tests

	(1)	(2)	(3)	(4)	(5)	(6)
	Entire sample	Track dorm	Random dorm	Track non-dorm	Random non-dorm	Balance test p
Panel A: High school graduation test scores						
Mean score (standardized)	0.088	0.169	0.198	0.000	0.000	0.426
A on graduation test	0.278	0.320	0.325	0.222	0.253	0.108
$\leq$C on graduation test	0.233	0.224	0.201	0.254	0.250	0.198
Panel B: Demographic characteristics						
Female	0.513	0.499	0.517	0.523	0.514	0.103
Black	0.319	0.503	0.524	0.116	0.118	0.181
White	0.423	0.354	0.332	0.520	0.495	0.851
Other race	0.257	0.143	0.144	0.364	0.387	0.124
English-speaking	0.714	0.593	0.560	0.851	0.863	0.001
International	0.144	0.225	0.180	0.106	0.061	0.913
Panel C: Graduated high school in 2004 or earlier, necessary to enroll under tracking						
Eligible for tracking	0.516	1.000	0.027	1.000	0.033	0.124
Eligible | A student	0.475	1.000	0.002	1.000	0.010	0.037
Eligible | $\leq$C student	0.527	1.000	0.039	1.000	0.050	0.330
Panel D: High school located in Cape Town, proxy for dormitory eligibility						
Cape Town high school	0.411	0.088	0.083	0.765	0.754	0.657
Cape Town | A student	0.414	0.101	0.065	0.848	0.811	0.976
Cape Town | $\leq$C student	0.523	0.146	0.186	0.798	0.800	0.224

Notes: Table 1 reports summary statistics of student characteristics at the time of enrollment, for the entire sample (column 1), tracked dormitory students (column 2), randomly assigned dormitory students (column 3), tracked non-dormitory students (column 4), and randomly assigned non-dormitory students (column 5). The p-values reported in column 6 are from testing whether the mean change in each variable between the tracking and random assignment periods is the same for dormitory and non-dormitory students.

tion tests (panel A).[19] They are more likely to be black, less likely to speak English as a home language, and more likely to be international students (panel B). However, the time changes between the tracking and random assignment periods are small and not significantly different between dormitory and non-dormitory students. The notable exception is that the proportion of

[19] I construct students' high school graduation test scores from subject-specific letter grades, following the university's admissions algorithm. I observe grades for all six tested subjects for 85% of the sample, for five subjects for 6% of the sample, and for four or fewer subjects for 9% of the sample. I treat the third group of students as having missing scores. I assign the second group of students the average of their five observed grades but omit them from analyses that sub-divide students by their grades.

English-speaking students moves in different directions. The proportion of students who graduated from high school early enough to enroll in university during the tracking period (2004 or earlier) but did not enroll until random assignment was introduced (2006 or later) is very small and not significantly different between dormitory and non-dormitory students (panel C). I interpret this as evidence that students did not strategically delay their entrance to university in order to avoid the tracking policy. Finally, there is a high and time-invariant correlation between living in a dormitory and graduating from a high school outside Cape Town. This relationship reflects the university's policy of admitting students to live in dormitories if and only if their family lives outside the Cape Town metropolitan region.[20] The fact that this relationship does not change through time provides some reassurance that students are not strategically choosing whether or not to live in dormitories in response to the dormitory assignment policy change. This pattern may in part reflect prospective students' limited information about the dormitory assignment policy: the change was not announced in the university's admissions materials or in internal, local, or national media. On balance, these descriptive statistics support the identifying assumption that dormitory and non-dormitory students' mean GPAs would have experienced similar time changes if the assignment policy had remained constant.[21]

The primary outcome variable is student GPAs in their first year of university. The university did not at this time report students' GPAs or any other measure of average grades. I instead observe students' complete transcripts, which report percentage scores from 0 to 100 for each course. I construct a credit-weighted average score and then transform this to have mean zero and standard deviation one in the control group of non-dormitory

[20] I do not observe students' home addresses, which are used in the university's dormitory admissions decisions. Instead, I match records on students' high school to a publicly available database of high school GIS codes. I then code students as having attended a high school in or outside the Cape Town metropolitan area. This is an imperfect proxy of their home address, as long commutes and boarding schools are not uncommon in South Africa. Furthermore, the university allows students from very low-income neighborhoods on the outskirts of Cape Town to live in dormitories. There are also a small number of students from Cape Town permitted to live in the dormitories for medical reasons or because they have exceptional academic records.

[21] I also test the joint null hypothesis that the mean time changes in all the covariates are equal for dormitory and non-dormitory students. The bootstrap p-value is 0.911.

students, separately by year. The effects of tracking discussed below should therefore be interpreted in standard deviations of GPA. The numerical scores are intended to be time-invariant measures of student performance and are not typically "curved."[22] The nominal ceiling score of 100 does not bind: the highest score any student obtains averaged across her courses is 97 and the 99^{th} percentile of student scores is 84. These features provide some reassurance that my results are not driven by time-varying grading standards or by ceiling effects on the grades of top students. I return to these potential concerns in section 8.

3 Effects of Tracking on Mean Outcomes

Tracked dormitory students obtain GPAs 0.13 standard deviations lower than randomly assigned dormitory students (table 2 column 1). The 95% confidence interval is [-0.27, 0.01]. Including dormitory fixed effects, student demographics, and high school graduation test scores yields a slightly smaller treatment effect of -0.11 standard deviations with a narrowe 95% confidence interval of [-0.17, -0.04] (column 2).[23] The average effect of tracking is thus negative and robust to accounting for dormitory fixed effects and student covariates.[24] This pattern holds for all results reported in the pa-

[22] For example, the mean percentage scores on Economics 1 and Mathematics 1 fluctuate from year to year up to six and nine points respectively, approximately half of a standard deviation.

[23] The bootstrapped standard errors reported in table 2 allow clustering at the dormitory-year level. Non-dormitory students are treated as individual clusters, yielding 60 large clusters and approximately 7000 singleton clusters. I also use a wild cluster bootstrap to approximate the distribution of the test statistic under the null hypothesis of zero average treatment effect (Cameron, Miller, and Gelbach, 2008). The p-values are 0.090 for the regression model with no controls (column 1) and 0.000 for the model with dormitory fixed effects and student covariates (column 3). As a final robustness check, I account for the possibility of persistent dormitory-level shocks with a wild bootstrap clustered at the dormitory level. The p-values are 0.104 and 0.002, respectively with and without fixed effects and student covariates.

[24] The regression-adjusted results in column 2 exclude approximately 9% of students whose high school graduation test scores are missing in my data. I also estimate the treatment effect for the entire sample with missing data indicators and find a very similar result (column 3). Results using both regression adjustment and inverse probability weighting are marginally larger (columns 4 and 5). The reweighted results are robust, though slightly less precisely estimated, to trimming propensity score outliers following Crump, Hotz, Im-

Table 2: Average Treatment Effect of Tracking on Tracked Students

	(1)	(2)	(3)	(4)	(5)
Tracking × Dormitory	-0.129	-0.107	-0.130	-0.144	-0.141
	(0.073)	(0.040)	(0.042)	(0.073)	(0.069)
Tracking	0.000	0.002	-0.013	0.042	-0.009
	(0.023)	(0.021)	(0.020)	(0.057)	(0.049)
Dormitory	0.172	0.138	0.173	0.221	0.245
	(0.035)	(0.071)	(0.072)	(0.061)	(0.064)
Dormitory fixed effects		×	×	×	×
Student covariates		×	×	×	×
Missing data indicators			×		×
Reweighting				×	×
Adjusted R^2	0.006	0.255	0.230	0.260	0.275
# dormitory-year clusters	60	60	60	60	60
# dormitory students	7480	6600	7480	6600	7480
# non-dormitory students	7188	6685	7188	6685	7188

Notes: Table 2 reports results from regressing GPA on indicators for living in a dormitory, the tracking period and their interaction. Columns 2-5 report results controlling for dormitory fixed effects and student covariates: gender, language, nationality, race, a quadratic in high school graduation test scores, and all pairwise interactions. Columns 2 and 4 report results excluding students with missing test scores from the sample. Columns 3 and 5 report results including all students, with missing test scores replaced with zeros and an indicator variable for missing test scores added. Columns 4 and 5 report results from regressions using propensity score weights that reweight all groups to have the same distribution of observed student covariates as tracked dormitory students. Standard errors in parentheses are from 1000 bootstrap replications clustering at the dormitory-year level and re-estimating the weights on each iteration.

per: accounting for student and dormitory characteristics yields narrower confidence intervals and unchanged treatment effect estimates.

How large is a treatment effect of 0.11 to 0.14 standard deviations? This is substantially smaller than the black-white GPA gap at this university (0.46 standard deviations) but larger than the female-male GPA gap (0.09). The effect size is marginally larger than when students are strategically assigned to squadrons at the US Airforce Academy (Carrell, Sacerdote, and West, 2013) and marginally smaller than when Kenyan primary school students are tracked into classrooms (Duflo, Dupas, and Kremer, 2011).[25] These results

bens, and Mitnik (2009). This provides reassuring evidence that the results are not driven by lack of common support on the four groups' observed characteristics. However, the trimming rule is optimal for the average treatment effect with a two-group research design; this robustness check is not conclusive for the average treatment effect on the treated with a difference-in-differences design.

[25] Although Duflo *et al.* find that students perform on average better under tracking, they argue that this reflects the benefits of tailored instruction outweighing peer effects.

provide a consistent picture about the plausible average short-run effects of alternative group assignment policies. These effects are not “game-changers” but they are substantial relative to many other education interventions.[26]

Tracking changes peer groups in different ways: high-scoring students live with higher-scoring peers and low-scoring students live with lower-scoring peers. The effects of tracking are thus likely to vary systematically with students’ demographic and academic characteristics, I explore this heterogeneity in two ways. I first estimate conditional average treatment effects for different subgroups of students. In section 4, I estimate quantile treatment effects of tracking, which show how tracking changes the full distribution of GPAs.

I begin by estimating equation 1 fully interacted with an indicator for students who score above the sample median on their high school graduation test. Above- and below-median students’ GPAs fall respectively 0.24 and 0.01 standard deviations under tracking (cluster bootstrap standard errors 0.06 and 0.07; p-value of difference = 0.014). These very different effects arise despite the fact that above- and below-median students experience “treatments” of similar magnitude. High-scoring students have residential peers who on average score 0.20 standard deviations higher under tracking, while low-scoring students have residential peers who on average score 0.27 standard deviations lower under tracking. This is not consistent with a linear or symmetric response to changes in mean peer quality.[27] Either low-scoring students are more sensitive to changes in their mean peer group composition or outcomes depend on some measure of peer quality other than mean test scores.

The near-zero treatment effect of tracking on above-median students is perhaps surprising. Splitting the sample in two may be an insufficiently flexible specification and may fail to discern positive effects on very high-scoring students. I therefore estimate smoothed treatment effects of tracking

[26] For example, McEwan (2013) conducts a meta-study of experimental primary education interventions in developing countries. He finds average effects across studies of 0.12 for class size and composition interventions and 0.06 for school management or supervision interventions. Group re-organization may therefore be an unusually cost-effective option.

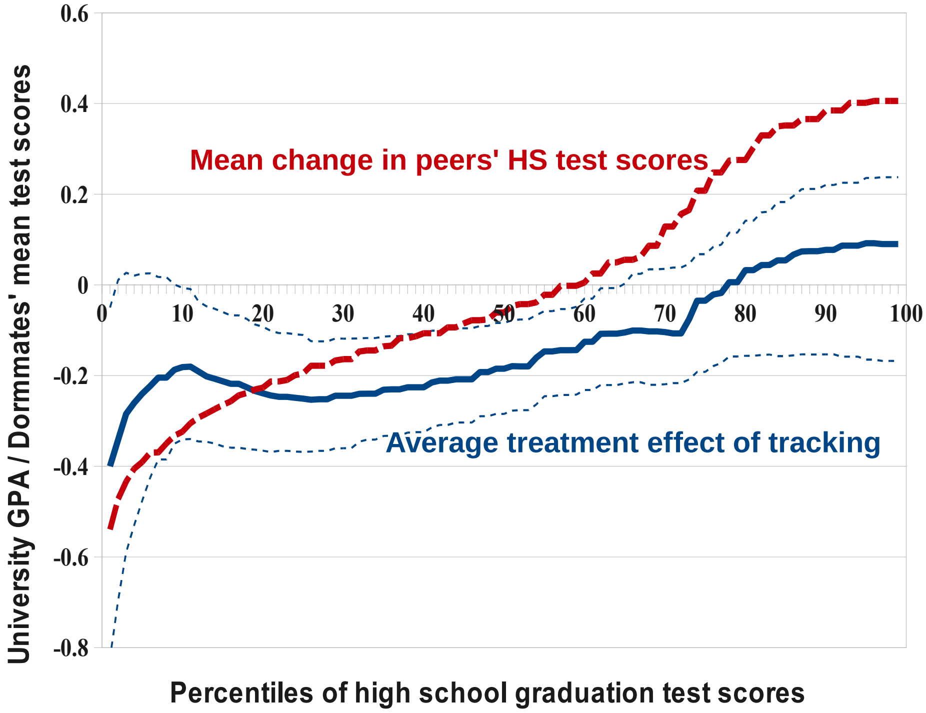
[27] I test whether above- and below-median students have the same ratios of the treatment effect to the change in mean peer high school test scores. The bootstrap p-value is 0.070.

throughout the distribution of high school test scores. Figure 2 shows that treatment effects of tracking are negative for more than half of the distribution. The negative point estimates in the left tail are considerably larger than the positive point estimates in the right tail, though they are not statistically different. I reject equality of the treatment effects and changes in mean peer high school test scores in the right tail but not the left tail. These results reinforce the finding that low-scoring students are substantially more sensitive to changes in peer group composition than high-scoring students. The point estimates suggest that tracking may have a small positive effect on students in the top quartile but this effect is very imprecisely estimated.[28]

There is stronger evidence of heterogeneity across high school test scores than across demographic subgroups. Treatment effects are larger on black than white students: -0.20 versus -0.11 standard deviations. However, this difference is not significant (bootstrap *p*-value 0.488) and shrinks substantially when conditioning on high school graduation test scores. I also estimate a quadruple-differences model allowing the treatment effect of tracking to differ across four race/academic subgroups (black/white × above/below median). The point estimates show that tracking affects below-median students more than above-median students within each race group and affects black students more than white within each test score group. However, neither pattern is significant at any conventional level. While there may be heterogeneity by race conditional on high school tests scores, the sample cannot convincingly detect it. There is no evidence of gender heterogeneity: tracking lowers female and male GPAs by 0.14 and 0.12 standard deviations respectively (bootstrap *p*-value 0.897). I conclude that the primary dimension of treament effect heterogeneity is high school test scores.

[28] A linear difference-in-differences model interacted with quartile or quintile indicators has positive but insignificant point estimates in the top quartile or quintile.

Figure 2: Effects of Tracking on GPA by High School Test Scores



Notes: Figure is constructed by estimating a student-level local linear regression of GPA against high school graduation test scores. The regression is estimated separately for each of the four groups (tracking/randomization period and dormitory/non-dormitory status). The second difference is evaluated at each percentile of the high school test score distribution. The dotted lines show a 95% confidence interval constructed from a nonparametric percentile bootstrap clustering at the dormitory-year level. The dashed line shows the effect of tracking on mean peer group composition, discussed in figure 1.

4 Effects of Tracking on the Distribution of Outcomes

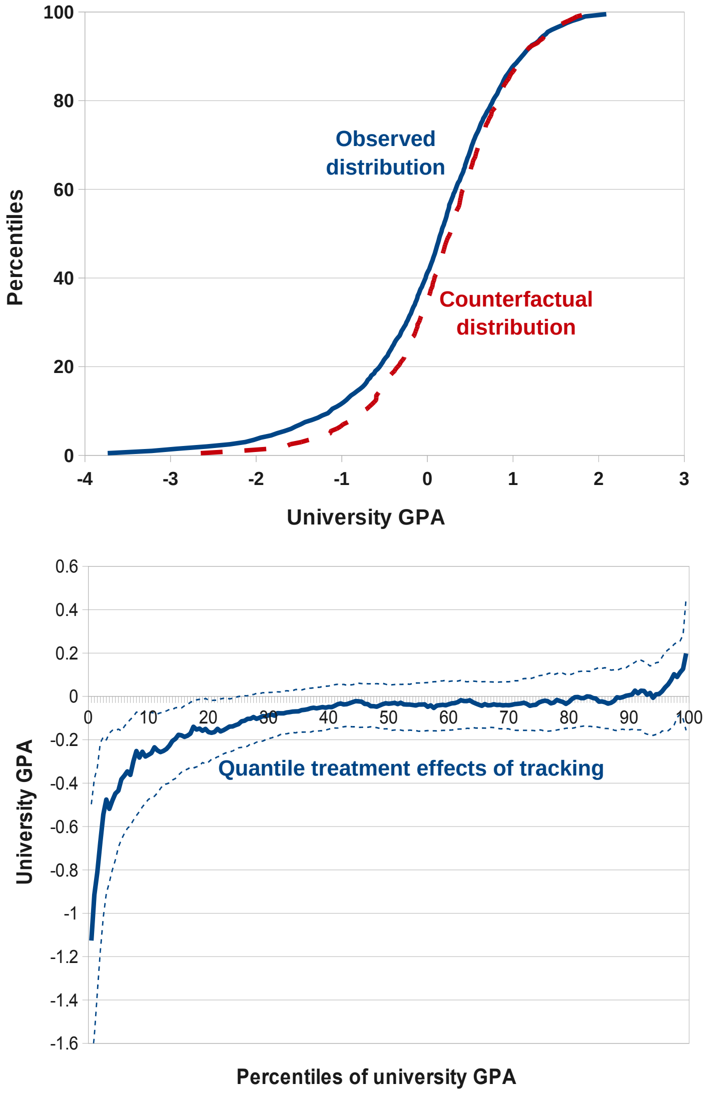
I also estimate quantile treatment effects of tracking on the treated students, which show how tracking changes the full distribution of GPAs. I first construct the counterfactual GPA distribution that the tracked dormitory students would have obtained in the absence of tracking (figure 3, first panel). I then evaluate the horizontal distance between the observed and counterfactual GPA distributions, which equal the quantile treatment effects of tracking on the treated students (figure 3, second panel). This provides substantially more information than the average treatment effect on the tracked students but requires stronger identifying assumptions. Specifically, the average effect is identified under the assumption that any time changes in the mean value of unobserved student-level GPA determinants are common across dormitory and non-dormitory students. The quantile effects are identified under the assumption that there are no time changes in the distribution of unobserved student-level GPA determinants for either dormitory or non-dormitory students. GPA may be subject to secular time trends or cohort-level shocks provided these are common across all students. I discuss the implementation of this model, developed by Athey and Imbens (2006) in appendix A and propose an extension to account flexibly for time trends in observed student-level characteristics.

Figure 3 shows that the negative effects of tracking are concentrated on the left tail. The point estimates are large and negative in the first quintile (0.1 - 1.1 standard deviations), small and negative in the second to fourth quintiles ($\leq$ 0.1 standard deviations), and small and positive in the top quintile ($\leq$ 0.2 standard deviations). The estimates are relatively imprecise, and the 95% confidence interval excludes zero only in the first quintile.[29] This contributes to a consistent impression that the negative average effects of tracking are driven by large negative effects on the left tail of the GPA or high school test score distribution.

There is no necessary relationship between figures 2 and 3. The former figure shows that the average treatment effect of tracking is large and negative for groups of students with low high school graduation test scores. The latter figure shows that the quantile treatment effect of tracking is large and negative on the left tail of the GPA distribution. The quantile results capture treatment effect heterogeneity between and within groups of students with similar high school test scores. However, they provide no information about the effect of tracking on any student or groups of students without ad-

[29] I construct the 95% confidence interval at each half-percentile using a percentile cluster bootstrap. The validity of the bootstrap has not been formally established for the nonlinear difference-in-differences model. However, Athey and Imbens (2006) report that bootstrap confidence intervals have better coverage rates in a simulation study than confidence intervals based on plug-in estimators of the asymptotic covariance matrix.

Figure 3: Quantile Treatment Effects of Tracking on the Tracked Students



Notes: The first panel compares the observed distribution of GPAs for tracked dormitory students (solid line) with the counterfactual constructed using the reweighted nonlinear difference-in-differences model discussed in appendix A (dashed line). The propensity score weights are constructed from a model including student gender, language, nationality, race, a quadratic in high school graduation test scores, all pairwise interactions, and dormitory fixed effects. The second panel shows the horizontal distance between the observed and counterfactual GPA distributions evaluated at each half-percentile. The axes are reversed for ease of interpretation. The dotted lines show a 95% confidence interval constructed from a percentile bootstrap clustering at the dormitory-year level and re-estimating the weights on each iteration.

ditional assumptions. See Bitler, Gelbach, and Hoynes (2010) and Heckman, Smith, and Clements (1997) for further discussion on this relationship.[30]

5 Effects of Tracking on Inequality of Outcomes

The counterfactual GPA distribution estimated above also provides information about the relationship between tracking and academic inequality. Specifically, I calculate several standard inequality measures on the observed and counterfactual distributions. The differences between these measures are the inequality treatment effects of tracking on the tracked students.[31] The literature on academic tracking has emphasized inequality concerns Betts (2011). This is the first study of which I am aware to measure explicitly the effect of tracking on inequality. Existing results from the econometric theory literature can be applied directly to this problem (Firpo, 2007, 2010; Rothe, 2010). Identification of these inequality effects requires no additional assumptions beyond those already imposed in the quantile analysis.[32]

Table 3 shows selected inequality measures on the observed and counterfactual GPA distributions. The interquartile range, interdecile range, and standard deviation are all significantly higher under tracking than under the counterfactual.[33] Tracking increases the interquartile range by approximately 12% of its baseline level and the other measures by approximately 20%. This reflects the particularly large negative effect of tracking on the left-most quantiles of the GPA distribution. Tracking thus decreases

[30] Garlick (2012) presents an alternative approach to rank-based distributional analysis. Using this approach, I estimate the effect of tracking on the probability that students will change their rank in the distribution of academic outcomes from high school to the first year of university. I find no effect on several measures of rank changes. Informally, this shows that randomly assigning students to dormitories instead of tracking them helps low-scoring students to "catch-up" to their high-scoring peers but does not facilitate "overtaking."

[31] I apply the same principle to calculate mean GPA for the counterfactual distribution. The observed mean is 0.16 standard deviations lower than the counterfactual mean (bootstrap standard error 0.07). This is consistent with the average effect estimated using a linear difference-in-differences model.

[32] Estimation and inference require additional regularity conditions, which I discuss briefly in the appendix.

[33] I do not calculate other common inequality measures (Gini coefficient, Theil index) because standardized GPA is not a strictly positive variable .

Table 3: Inequality Treatment Effects of Tracking

	(1) Observed distribution	(2) Counterfactual distribution	(3) Treatment effect	(4) Treatment effect in % terms
Interquartile range	1.023	0.907	0.116	12.8
	(0.043)	(0.047)	(0.062)	(6.8)
Interdecile range	2.238	1.857	0.381	20.5
	(0.083)	(0.091)	(0.109)	(5.9)
Standard deviation	0.909	0.766	0.143	18.7
	(0.027)	(0.032)	(0.037)	(4.8)

Notes: Table 3 reports summary measures of academic inequality for the observed distribution of tracked dormitory students' GPA (column 1) and the counterfactual GPA distribution for the same students in the absence of tracking (column 2). The counterfactual GPA is constructed using the reweighted nonlinear difference-in-differences model described in appendix A. Column 3 shows the treatment effect of tracking on academic inequality for the tracked students. Column 4 shows the treatment effect expressed as a percentage of the counterfactual level of inequality. The standard deviation is estimated by $\left(\hat{\mathbb{E}}\left[GPA^2\right] - \left\{\hat{\mathbb{E}}\left[GPA\right]\right\}^2\right)^{0.5}$, with the expectations constructed by integrating the area to the left of the relevant GPA distribution. The distribution is evaluated at half-percentiles to minimize measurement error due to the discrete construction of the counterfactual distribution. Standard errors in parentheses are from 1000 bootstrap replications clustering at the dormitory-year level and stratifying by dormitory status and assignment period.

mean academic outcomes and increases the inequality of academic outcomes. Knowledge of the quantile and inequality treatment effects permits a more comprehensive evaluation of the welfare consequences of tracking. These parameters might inform an inequality-averse social planner's optimal trade-off between efficiency and equity if the mean effect of tracking were positive, as has been found in some other contexts.

6 Effects of Random Variation in Dormitory Composition

The principal research design uses cross-policy variation by comparing tracked and randomly assigned dormitory students. My second research design uses cross-dormitory variation in dormitory composition induced by random assignment. I first use a standard test to confirm that residential peer effects are present in this setting, providing additional evidence that the main results are not driven by confounding factors. I document differences in peer effects within and between demographic and academic subgroups within dormitories, providing some information about mechanisms. In section 7, I explore whether peer effects estimated using random dormitory assignment can predict the distributional effects of tracking. I find that low-scoring students are more sensitive to changes in peer group composition than high-scoring students, which is qualitatively consistent with the effect of tracking. Quantitative predictions are, however, sensitive to model specification choices.

I first estimate workhorse linear-in-means peer effects model (Manski, 1993; Sacerdote, 2001):

$$GPA_{id} = \alpha_0 + \alpha_1 HS_{id} + \alpha_2 \overline{HS}_d + \vec{\alpha}\vec{X}_{id} + \vec{\mu}_d + \epsilon_{id}, \tag{3}$$

where HS_{id} and $\overline{HS}_d$ are individual and mean dormitory high school graduation test scores, $\vec{X}_{id}$ is a vector of student demographic characteristics, and $\vec{\mu}$ is a vector of dormitory fixed effects. α_2 measures the average gain in GPA from a one standard deviation increase in the mean high school graduation test scores of one's residential peers.[34] Random dormitory assignment ensures that $\overline{HS}_d$ is uncorrelated with individual students' unobserved characteristics so α_2 can be consistently estimated by least squares.[35] Random dormitory assignment means that average high school graduation test scores will be equal in expectation. α_2 is identified using sample variation in mean high school test scores across dormitories due to finite numbers of students in each dormitory. This variation is relatively low: the range and variance of dormitory means are approximately 10% of the range and variance of individual test scores. Given this limited variation, the results should be interpreted with caution.

I report estimates of equation 3 in table 4, using the sample of all dor-

[34] α_2 captures both "endogeneous" effects of peers' GPA and "exogeneous" effects of peers' high school graduation test scores, using Manski's terminology. Following the bulk of the peer effects literature, I do not attempt to separate these effects.

[35] To test whether assignment was random, I explore the balance of high school graduation test scores and demographics across dormitories. I regress each variable on a vector of dormitory fixed effects, calculate the Wald test statistic for the test of joint equality and sum these test statistic across all characteristics. The bootstrap p-value is 0.885. However, the dormitories are marginally unbalanced on high school graduation test scores in one of the two years of random assignment.

Table 4: Peer Effects from Random Assignment to Dormitories

	(1)	(2)	(3)	(4)	(5)	(6)
Own HS graduation	0.362	0.332	0.331	0.400	0.373	0.373
test score	(0.014)	(0.014)	(0.014)	(0.024)	(0.023)	(0.023)
Own HS graduation				0.137	0.144	0.142
test score squared				(0.017)	(0.017)	(0.017)
Mean dorm HS graduation	0.241	0.222	0.220	0.221	0.208	0.316
test score	(0.093)	(0.098)	(0.121)	(0.095)	(0.103)	(0.161)
Mean dorm HS graduation				0.306	0.311	-0.159
test score squared				(0.189)	(0.207)	(0.316)
Own × mean dorm HS				-0.129	-0.132	-0.132
graduation test score				(0.073)	(0.069)	(0.069)
p-value of test against				0.000	0.000	0.000
equivalent linear model						
Adjusted R^2	0.213	0.236	0.248	0.244	0.270	0.278
# students	3068	3068	3068	3068	3068	3068
# dormitory-year clusters	30	30	30	30	30	30

Notes: Table 4 reports results from estimating equations 3 (columns 1-3) and 5 (columns 4-6). Columns 2, 3, 5, and 6 control for students' gender, language, nationality and race. Columns 3 and 6 include dormitory fixed effects. The sample is all dormitory students in the random assignment period with non-missing high school graduation test scores. Standard errors in parentheses are from 1000 bootstrap replications clustering at the dormitory-year level.

mitory students in the random assignment period. I find that $\hat{\alpha}_2 \approx 0.22$, a result that is robust to conditioning on student demographics and dormitory fixed effects. This implies that moving a student from the dormitory with the lowest observed mean high school graduation test score to the highest would increase her GPA by 0.18 standard deviations. This implies large peer effects relative to existing estimates (Sacerdote, 2011). Stinebrickner and Stinebrickner (2006) suggest a possible reason for this pattern. They document that peers' study time is an important driver of peer effects and that peer effects are larger using a measure that attaches more weight to prior study behavior: high school GPA instead of SAT scores. I measure peer characteristics using scores on a content-based high graduation examination, while SAT scores are the most common measure in the existing literature. However, the coefficient from the dormitory fixed effects regression is fairly imprecisely estimated (90% confidence interval from 0.02 to 0.42) so the magnitude should be interpreted with caution. This may reflect

Table 5: Subgroup Peer Effects from Random Assignment to Dormitories

	(1)	(2)	(3)	(4)
Own HS graduation	0.327	0.327	0.369	0.322
test score	(0.016)	(0.016)	(0.017)	(0.017)
Mean dorm HS graduation test	0.203	0.162		
score for own race	(0.059)	(0.083)		
Mean dorm HS graduation test	-0.007	-0.035		
score for other races	(0.055)	(0.091)		
Mean dorm HS graduation test			0.050	0.099
score for own faculty			(0.045)	(0.048)
Mean dorm HS graduation test			0.198	0.190
score for other faculties			(0.062)	(0.083)
Adjusted R^2	0.219	0.243	0.214	0.249
# students	3068	3068	3068	3068
# dormitory-year clusters	30	30	30	30

Notes: Table 5 reports results from estimating equation 4 using race subgroups (columns 1-2) and faculty subgroups (columns 3-4). "Faculty" refers to colleges/schools within the university such as commerce and science. Columns 2 and 4 include dormitory fixed effects and control for students' gender, language, nationality and race. The sample is all dormitory students in the random assignment period with non-missing high school graduation test scores. Standard errors in parentheses are from 1000 bootstrap replications clustering at the dormitory-year level.

the limited variation in $\overline{HS}_d$.[36]

The linear-in-means model can be augmented to allow the effect of residential peers to vary within and across sub-dormitory groups. Specifically, I explore within- and across-race peer effects by estimating:

$$GPA_{igd} = \alpha_0 + \beta_1 HS_{ird} + \beta_2 \overline{HS}_{rd} + \beta_3 \overline{HS}_{-rd} + \vec{\beta}\vec{X}_{ird} + \vec{\mu}_d + \epsilon_{ird}. \quad (4)$$

For student i of race r in dormitory d, $\overline{HS}_{rd}$ and $\overline{HS}_{-rd}$ denote the mean high school graduation test scores for other students in dormitory d of, respectively, race r and all other race groups. $\hat{\beta}_2$ and $\hat{\beta}_3$ equal 0.16 and -0.04 respectively (table 5, column 2). The difference strongly suggests that peer effects operate primarily within race groups but it is quite imprecisely estimated (bootstrap p-value equality 0.110). I interpret this as evidence that spatial proximity does not automatically generate peer effects. Instead, peer groups are formed through a combination of spatial proximity and proximity

[36] As a robustness check, I use a wild cluster bootstrap to approximate the distribution of the test statistic under the null hypothesis of zero peer effect. This yields p-values of 0.088 using dormitory-year clusters and 0.186 using dormitory clusters.

along other dimensions such as race, which remains highly salient in South Africa.[37] This indicates that interation patterns by students may mediate residential peer effects, meaning that estimates are not policy-invariant.

I also explore the content of the interaction patterns that generate residential peer effects by estimating equation 4 using faculty/school/college groups instead of race groups. The estimated within- and across-faculty peer effects are respectively 0.10 and 0.19 (cluster bootstrap standard errors 0.05 and 0.08). Despite their relative imprecision, these results suggests that within-faculty peer effects are not systematically stronger than cross-faculty peer effects.[38] This result is not consistent with peer effects being driven by direct academic collaboration such as joint work on problem sets or joint studying for examinations. Interviews with students at the university suggest two channels through which peer effects operate: time allocation over study and leisure activities, and transfers of tacit knowledge such as study skills and academic norms. This is consistent with prior findings of strong peer effects on study time (Stinebrickner and Stinebrickner, 2006) and social activities (Duncan, Boisjoly, Kremer, Levy, and Eccles, 2005). Research in the United States has also noted that non-traditional and first generation students often lack information about how to navigate academic bureaucracy and interact with faculty. Low-scoring students at this university are likely to come from schools and families with limited prior exposure to university study. Exposure to more prepared peers may be important in acquiring this

[37] I find a similar result using language instead of race to define subgroups. This pattern could also arise if students sort into racially homogeneous geographic units by choosing rooms within their assigned dormitories. As I do not observe roommate assignments, I cannot test this mechanism.

[38] Each at the University of Cape Town is registered in one of six faculties: commerce, engineering, humanities and social sciences, health sciences, law and science. Some students take courses exclusively within their faculty (engineering, health sciences) while some courses overlap across multiple faculties (introductory statistics is offered in commerce and science, for example). I obtain similar results using course-specific grades as the outcome and allowing residential peer effects to differ at the course level. For example, I estimate equations 3 and 4 with Introductory Microeconomics grades as an outcome. I find that there are strong peer effects on grades in this course ($\hat{\alpha}_2 = 0.34$ with s.e. 0.15) but they are not driven primarily by other students in the same course ($\hat{\beta}_2 = 0.06$ with s.e. 0.17 and $\hat{\beta}_3 = 0.17$ with s.e. 0.15). This, and other course-level regressions, are consistent with the main results but the smaller sample sizes yield relatively imprecise estimates that are somewhat sensitive to the inclusion of covariates.

information.

Combining the race- and faculty-level peer effects results indicates that spatial proximity alone does not generate peer effects. Some direct interaction is also necessary and may be more likely when students are socially as well as spatially proximate. However, the relevant form of the interaction is not direct academic collaboration. The research design and data cannot conclusively determine what interactions do generate the estimated peer effects.

7 Reconciling Results from the Different Designs

The linear-in-means model restricts average GPA to be invariant to any group reassignment of students across groups: reassigning a strong student to a new group has equal but oppositely signed effects on her old and new peers' average GPA. If the true GPA production function is linear, then the average treatment effect of tracking relative to random assignment must be zero. I therefore estimate a more general production function that permits nonlinear peer effects:

$$\begin{aligned} GPA_{id} = \gamma_0 + \gamma_1 HS_{id} + \gamma_2 \overline{HS}_d + \gamma_{11} HS_{id}^2 + \gamma_{22} \overline{HS}_d^2 \\ + \gamma_{12} HS_{id} \times \overline{HS}_d + \vec{\gamma} \vec{X}_{id} + \vec{\mu}_d + \epsilon_{id} \end{aligned} \tag{5}$$

This is a parsimonious specification that permits average outcomes to vary over assignment processes but may not be a perfect description of the GPA production process. In particular, I limit attention to the mean as a summary measure of dormitory characteristics.[39] γ_{12} and γ_{22} are the key parameters of the model. γ_{12} indicates whether own and peer high school graduation test scores are complements or substitutes in GPA production. Equivalently, γ_{12} indicates whether GPA is super- or submodular in own and peer scores. If $\gamma_{12} < 0$, the GPA gain from high-scoring peers is larger for low-scoring students. In classic binary matching models, this parameter gov-

[39] See Carrell, Sacerdote, and West (2013) for an alternative parameterization and Graham (2011) for background discussion. Equation 5 has the attractive feature of aligning with theoretical literatures on binary matching and on neighborhood segregation. The results are qualitatively similar if dormitory-year means are replaced with medians.

erns whether positive or negative assortative matching is output-maximizing (Becker, 1973). In matching models with more than two agents, this parameter is not sufficient to characterize the output-maximizing set of matches. γ_{22} indicates whether GPA is a concave or convex function of peers' mean high school graduation test scores. If $\gamma_{22} < 0$, total output is highest when mean test scores are identical in all groups. If $\gamma_{22} > 0$, total output is highest when some groups have very high means and some groups have very low means. This parameter has received relatively little attention in the peer effects literature but features prominently in some models of neighborhood effects (Benabou, 1996; Graham, Imbens, and Ridder, 2013). Tracking will deliver higher total GPA than random assignment if both parameters are positive and vice versa. If the parameters have different signs, the average effect of tracking is ambiguous.[40]

Estimates from equation 5 are shown in table 4 columns 4, 5 (adding controls for student demographics) and 6 (adding dormitory fixed effects). $\hat{\gamma}_{12}$ is negative and marginally statistically significant across all specifications. The point estimate of -0.13 (s.e. 0.07) in column 6 implies the GPA gain from an increase in peers' mean test scores will be 0.2 standard deviations larger for students at the 25th percentile of the high school test score distribution than students at the 75th percentile. This is consistent with the section 4 result that low-scoring students are hurt more by tracking than high-scoring students are helped. However, the sign and magnitude of $\hat{\gamma}_{22}$ flips from positve to negative with the inclusion of dormitory fixed effects. This provides mixed evidence regarding the concavity or convexity of the GPA production function.

I draw three conclusions from these results. First, there is clear evidence of nonlinear peer effects from the cross-group variation generated under random assignment. Likelihood ratio tests prefer the nonlinear models in

[40] To derive this result, note that $\mathbb{E}[\overline{HS}_d|HS_{id}] = HS_{id}$ under tracking and $\mathbb{E}[HS_{id}]$ under random assignment. Hence, $\mathbb{E}[HS_{id}\overline{HS}_d] = \mathbb{E}[\overline{HS}_d^2] = \mathbb{E}[HS_{id}^2]$ under tracking and $\mathbb{E}[HS_{id}]^2$ under random assignment. Plugging these results into equation 5 for each assignment policy yields $\mathbb{E}[Y_{id}|\text{Tracking}] - \mathbb{E}[Y_{id}|\text{Randomization}] = \sigma_{HS}^2 \left(\gamma_{22} + \gamma_{12}\right)$. This simple demonstration assumes an infinite number of students and dormitories. This assumption is not necessary but simplifies the exposition.

Table 6: Observed and Predicted GPA Using Different Production Function Specifications

	(1) Quartile 4	(2) Quartile 1
Panel A: Mean GPA		
Observed	0.761	-0.486
Predicted, without dorm fixed effects	0.889	-0.345
Predicted, least squares with dorm dummies	0.698	-0.433
Predicted, within-group transformation	0.689	-0.503
Panel B: Mean treatment effect of tracking		
Estimated from DD design	0.032	-0.225
Predicted, without dorm fixed effects	0.223	-0.050
Predicted, least squares with dorm dummies	0.041	-0.139
Predicted, within-group transformation	0.032	-0.195

Notes: Table 6 panel A reports observed GPA (row 1) and predicted GPA from three different models. All predictions use observed regressor values from tracked dormitory students and estimated coefficients from randomly assigned dormitory students. The first prediction uses coefficients generated by estimating equation 5 without dormitory fixed effects (shown in column 5 of table 4). The second prediction uses coefficients generated by estimating equation 5 without dormitory indicator variables (shown in column 6 of table 4). The third prediction uses coefficients generated by estimating equation 5 with data from a within-dormitory transformation (shown in column 6 of table 4). The second and third predictions differ because the values of the dormitory fixed effects respectively are and are not used in the prediction.

columns 4-6 to the corresponding linear models in columns 1-3. Second, peer effects estimates relying on randomly induced cross-group variation may be sensitive to the support of the data. Using dormitory fixed effects reduces the variance of $\overline{HS}_d$ from 0.19 to 0.11. This results in different conclusions about the curvature of the GPA production function in columns 5 and 6. Third, the results from the fixed effects specification (column 6) are qualitatively consistent with the fact that tracking lowered mean GPA relative to random assignment.

Are the coefficient estimates from equation 5 quantitatively, as well as qualitatively, consistent with the observed treatment effects of tracking? I combine coefficients from estimating equation 5 in the sample of randomly assigned dormitory student with observed values of individual- and dormitory-level regressors in the sample of tracked dormitory students. I then predict the level of GPA and the treatment effect of tracking for students in the first and fourth quartiles of the high school graduation test score distribution. I compare these predictions to observed GPA for tracked dormitory students and to the difference-in-differences treatment effect of tracking estimated.

The results in table 6 show that the predictions are sensitive to specification of equation 5. Excluding dormitory fixed effects (row 2) yields very inaccurate predictions, with GPA and treatment effects too high for students in the top and bottom quartiles. This reflects the estimated convexity of the GPA production function without dormitory fixed effects ($\hat{\gamma}_{22} = 0.31$ but insignificant). After including dormitory fixed effects, the production function is not convex ($\hat{\gamma}_{22} = -0.16$ but insignificant) and own and peer test scores are substitutes ($\hat{\gamma}_{12} = 0.13$). The fixed effects estimates therefore predict negative and zero treatment effects on the first and fourth quartiles respectively, matching the difference-in-differences estimates. However, the first quartile estimates are quite sensitive to specifying the fixed effects with dormitory dummies (row 3) or using a within-group data transformation (row 4).

This exercise illustrates that a simple reduced form model of the GPA production function can come close to predicting the treatment effects of tracking. However, the predictions are extremely sensitive to specification choices regarding covariates and group fixed effects, which in turn influence the support of the data. These are precisely the choices for which economic theory is likely to provide little guidance. Statistical model selection criteria are also inconclusive in this setting.[41] This sensitivity may be due to out-of-sample extrapolation, potential dependence of GPA on dormitory-level statistics other than the mean, or behavioral responses by students that make peer effects policy-sensitive.

[41] For example, the Akaike and Bayesian information criteria are lower for the models respectively with and without dormitory fixed effects, while a likelihood ratio test for equality of the models has p-value 0.083. Hurder (2012) also attempts to use peer effects estimates to predict the effects of changing the peer group assignment rule and reaches a similar conclusion.

8 Alternative Explanations for the Apparent Effects of Tracking

I consider four alternative explanations that might have generated the observed GPA difference between tracked and randomly assigned dormitory students. The first two explanations are violations of the "parallel time changes" assumption: time-varying student selection regarding whether or not to live in a dormitory and differential time trends in dormitory and non-dormitory students' characteristics. The third explanation is that the treatment effects are an artefact of the grading system and do not reflect any real effect on learning. The fourth explanation is that dormitory assignment affects GPA through a mechanism other than peer effects; this would not invalidate the results but would change their interpretation.

8.1 Selection into Dormitory Status

The research design assumes that non-dormitory students are an appropriate control group for any time trends or cohort effects on dormitory students' outcomes. This assumption may fail if students select whether or not to live in a dormitory based on the assignment policy. I argue that such behavior is unlikely and that my results are robust to accounting for selection. First, the change in dormitory assignment policy was not officially announced or widely publicised, limiting students' ability to respond. Second, table 1 shows that there are approximately equal time changes in dormitory and non-dormitory students' demographic characteristics and high school graduation test scores. Third, the results are robust to accounting for small differences in these time changes using regression or reweighting.

Fourth, the admission rules cap the number of students whose families live in Cape Town who may be admitted to the dormitory system. Given this rule, I use an indicator for whether each student attended a high school outside Cape Town as an instrument for whether the student lives in a dormitory. High school location is an imperfect proxy for home location, which I do not observe. Nonetheless, the instrument strongly predicts dormitory status: 76% of non-Cape Town students live in dormitories compared to 8%

of Cape Town students. The intention-to-treat and instrumented treatment effects (table 7, columns 2 and 3) are very similar to the treatment effects without instruments (table 2).

8.2 Differential Time Trends in Student Characteristics

The research design assumes that dormitory and non-dormitory students' GPAs do not have different time trends for reasons unrelated to the change in assignment policy. I present two arguments against this concern. First, I extend the analysis to include data from the 2001–2002 academic years ("early tracking"), in addition to 2004–2005 ("late tracking") and 2007–2008 (randon assignment). I do not observe dormitory assignments in 2001–2002 so I report only intention-to-treat effects.[42] The raw data are shown in figure 4 panel A. I estimate the effect of tracking under several possible violations of the parallel trends assumption. The average effect of tracking comparing 2001-2005 to 2007-2008 is -0.090 with standard error 0.044 (table 7, column 4). This estimate is appropriate if one group of students experiences a transitory shock in 2004/2005. A placebo test comparing the difference between Cape Town and non-Cape Town students' GPAs in 2001-2002 and 2004-2005 yields a small positive but insignificant effect of 0.058 (standard error 0.052). I use the placebo test result to construct a "trend-adjusted" treatment effect of -0.175 with standard error 0.100 (table 7, column 6). This estimate is appropriate if the two groups of students have linear but non-parallel time trends and are subject to common transitory shocks (Heckman and Hotz, 1989). Finally, I adjust for any GPA trend by estimating a linear time trend in the GPA gap between Cape Town and non-Cape Town students from 2001 to 2005. I then project that trend into 2007–2008 and estimate the deviation of the GPA gap from its predicted level. This method yields a treatment effect of random assignment relative to tracking of 0.141 with standard error 0.093 (table 7, column 5). This estimate is appropriate if the two groups of

[42] The cluster bootstrap standard errors do not take into account potential clustering within (unobserved) dormitories in 2001–2002 and so may be downward-biased. I omit the 2003 academic year because the data extract I received from the university had missing identifiers for approximately 80% of students in that year. I omit 2006 because first year students were randomly assigned to dormitories that still contained tracked second year students. The results are robust to including 2006, which is shown in figure 4.

Table 7: Robustness Checks

Outcome	Dorm						No. of			% credits	GPA | non-
(default is GPA)	student						credits			excluded	exclusion
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Cape Town high school	0.601										
	(0.019)										
Cape Town high school		-0.093		-0.090		-0.115					
× tracking period		(0.034)		(0.044)		(0.055)					
Dormitory ×			-0.133				-0.013	-0.139	-0.165	0.027	-0.077
tracking period			(0.050)				(0.038)	(0.043)	(0.044)	(0.005)	(0.050)
Cape Town high school					0.141						
× randomization period					(0.093)						
Placebo pre-treament						0.058					
diff-in-diff						(0.052)					
Trend-corrected						-0.175					
treatment effect						(0.100)					
Sample period (default				2001-	2001-	2001-					
is 2004-2008)				-2008	-2008	-2008					
Dormitory fixed effects							×	×	×	×	×
Student covariates	×	×	×				×	×	×	×	×
Missing data indicators	×	×	×				×	×	×	×	×
Instruments			×								
Faculty fixed effects									×		
Pre-treatment trend					×						
Adjusted R^2	0.525	0.231	0.231	0.002	0.000	0.000	0.127	0.242	0.229	0.052	0.302
# dormy-year clusters	60	60	60	60	60	60	60	52	60	60	60
# dormitory students	6915	6915	6915	8509	8509	8509	7480	6795	7480	7480	7449
# non-dorm students	6466	6466	6466	14203	14203	14203	7188	7188	7188	7188	7043

Notes: Table 7 reports results from the robustness checks discussed in subsections 8.1 - 8.3. Columns 1–3 show the relationship between students' GPA (outcome), whether they live in dormitories (treatment) and whether they graduated from high schools located outside Cape Town (instrument). The coefficient of interest is on the treatment or instrument interacted with an indicator for whether students attended the university during the tracking period. Column 1 shows the first stage estimate, column 2 shows the reduced form estimate, and column shows the IV estimate. Dormitory fixed effects are excluded because they are colinear with the treatment indicator. Columns 4–6 use data from 2001, 2002, 2004, 2005, 2007, and 2008 to test the parallel time trends assumption. Column 4 estimates a standard difference-in-differences model comparing all four observed years of tracking to the two observed years of random assignment. Column 5 estimates the difference between the observed GPA under random assignment and the predicted GPA from a linear time trend extrapolated from the tracking period. Column 6 shows the placebo difference-in-differences test comparing the first two years of tracking to the last two years of tracking and the difference between the actual and placebo effects following Heckman and Hotz (1989). Column 7 estimates equation 2 with the credit-weighted number of courses as the outcome. Column 8 estimates equation 2 excluding dormitories that are either observed in only one period or use a different admission rule. Column 9 estimates equation 2 including college/faculty/school fixed effects. Column 10 estimates equation 2 with the credit-weighted percentage of courses from which students are academically excluded as the outcome. Column 11 estimates equation 2 with GPA calculated using only grades from non-excluded courses as the outcome. Standard errors in parentheses are from 1000 bootstrap replications. The bootstrap resamples dormitory-year clusters except for the 2001-2002 data in columns 4-6, for which dormitory assignments are not observed.

students have non-parallel time trends whose difference is linear. The effect of tracking is relatively robust across the standard difference-in-differences model and all three models estimated under weaker assumptions. However, there is some within-policy GPA variation through time: intention-to-treat students (those from high schools outside Cape Town) strongly outperform control students in 2006 and 2007 but not 2008. The reason for this divergence is unclear.

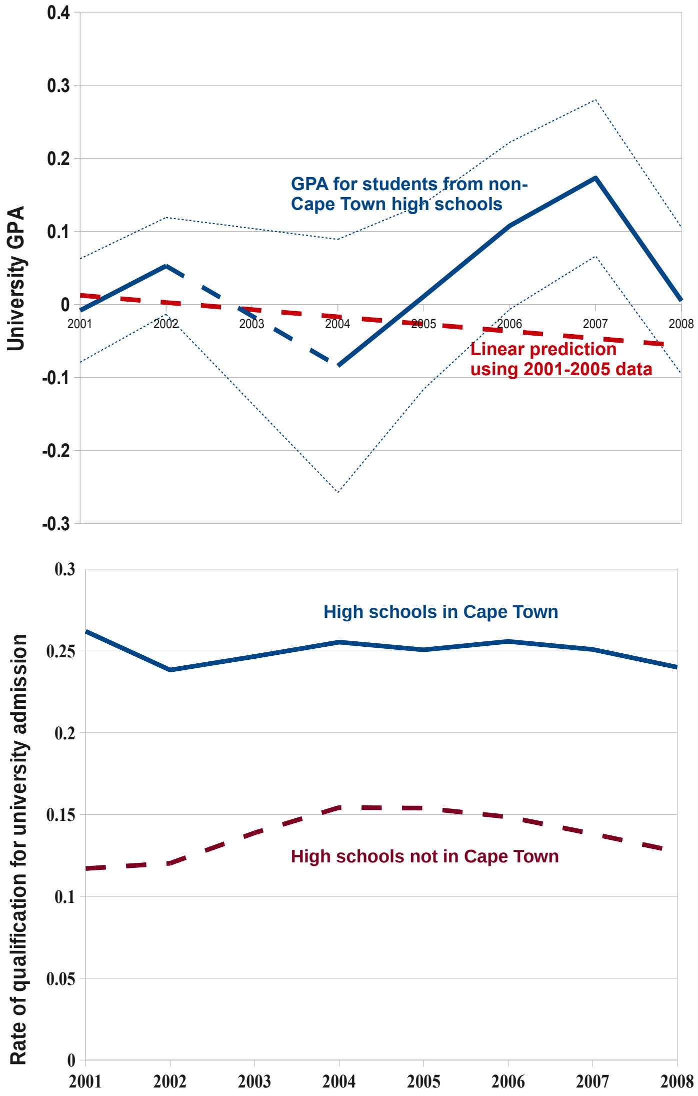
Second, the time trends in the proportion of graduating high school students who qualify for admission to university are very similar for Cape Town and non-Cape Town high schools between 2001 and 2008 (figure 4 panel B). Hence, the pools of potential dormitory and non-dormitory students do not have different time trends. This helps to address any concern that students make different decisions about whether to attend the University of Cape Town due to the change in the dormitory assignment policy. However, the set of students who qualify for university admission is an imperfect proxy for the set of potential students at this university. Many students whose high school graduation test scores qualify them for admission to a university may not qualify for admission to this relatively selective university.

Third, the results are not driven by two approximately simultaneous policy changes that occurred at the university. The university charged flat tuition fees up until 2005 and charged per-credit tuition fees from 2006. This may have changed the number of courses for which students registered. However, the number of credit-weighted courses remains constant through for dormitory and non-dormitory students, with a difference-in-differences estimate of 0.013, less than 0.4% of the mean (table 7 column 7). The university also closed one dormitory in 2006 and opened a new dormitory in 2007, as well as reserving one cheaper dormitory for low-income students under both policies. The estimated treatment effect is robust to excluding all three dormitories (table 7 column 8).

8.3 Limitations of GPA as an Outcome Measure

I explore four ways in which the results might be driven by the grading system, rather than peer effects: curving, ceiling effects, course choices, and

Figure 4: Long-term Trends in Student Academic Performance



Notes: The first panel shows mean GPA for first year university students from high schools outside Cape Town. The time series covers the tracking period (2001-2005) and the random assignment period (2006-2008). Mean GPA for students from Cape Town high schools is, by construction, zero in each year. Data for 2003 is missing and replaced by a linear imputation. The second panel shows the proportion of grade 12 students whose score on the high school graduation examination qualified them for admission to university. The mean qualification rate for high schools in Cape Town is 0.138 in the tracking period (2001 - 2005) and 0.133 in the random assignment period (2007 - 2008). The mean qualification rate for high schools outside Cape Town is 0.250 in the tracking period (2001 - 2005) and 0.245 in the random assignment period (2007 - 2008). The second difference is 0.001 (bootstrap standard error 0.009) or, after weighting by the number of grade 12 students enrolled in each school, 0.007 (standard error 0.009).

course exclusions. First, instructors may use "curves" that keep features of the grade distribution constant through time within each course. Under this hypothesis, the effects of tracking may be negative effects on dormitory students relative to non-dormitory students, rather than negative effects on absolute performance. This would not invalidate the main result but would certainly change its interpretation. This is a concern for most test score measures but I argue that it is less pressing in this context. Instructors at this university are not encouraged to use grading curves and many examinations are subject to external moderation intended to maintain an approximately time-consistent standard. I observe several patterns in the data that are not consistent with curving. Mean grades in the three largest introductory courses at the university (microeconomics, management, information systems) show year-on-year changes within an assignment policy period of up to 6 points (on a 0 to 100 scale, approximately 1/3 of a standard deviation). Similarly, the 75th and 25th percentiles of the grades within these large first-year courses show year-on-year changes of up to 8 and 7 points respectively. This demonstrates that grades are not strictly curved in at least some large courses. I also examine the treatment effect of tracking on grades in the introductory accounting course, which builds toward an external qualifying examination administered by South Africa's Independent Regulatory Board for Auditors. This external assessment standard for accounting students, although it is only administered only after they graduate, reduces the scope for assessment to respond to events within the university. Tracking reduces mean grades in the introductory accounting course by 0.11 standard deviations (cluster bootstrap standard error 0.12, 2107 students). This provides some reassurance that tracking does indeed reduce the academic competence of low-scoring students.

Second, tracking may have no effect on high-scoring students if they already obtain near the maximum GPA and are constrained by ceiling effects. I cannot rule out this concern completely but I argue that it is unlikely to be central. The nominal grade ceiling of 100 does not bind for any student: the highest grade observed in the dataset is 97/100 and the 99th percentile is 84/100. Some courses may impose ceilings below the maximum grade, which will not be visible in my data. However, the course convenors for

Introductory Microeconomics, the largest first-year course at the university, confirmed that they used no such ceilings. The treatment effect of tracking on grades in this course is 0.130 standard deviations (cluster bootstrap standard error 0.056), so the average effect across all courses is at least similar to the average effect in a course without grade ceilings.

Third, dormitory students may take different classes in the tracking and random assignment periods and grading standards may differ across these classes. I find some evidence of changes in the type of courses students take: dormitory students take slightly fewer commerce and science classes in the tracking than random assignment period, relative to non-dormitory students. However, the effect of tracking is consistently negative within each type of class. The treatment effects for each faculty/school/college range between -0.23 for engineering and -0.04 for medicine. The average treatment effect with faculty fixed effects is -0.165 with standard error 0.044 (table 7, column 9). I conclude that the main results are not driven by time-varying course-taking behavior.

Fourth, the university employs an unusual two-stage grading system which does explain part of the treatment effect of tracking. Students are graded on final exams, class tests, homework assignments, essays, and class participation and attendance, with the relative weights varying across classes. Students whose weighted scores before the exam are below a course-specific threshold are excluded from the course and do not write the final exam. These students receive a grade of zero in the main data, on a 0-100 scale. I also estimate the treatment effect of tracking on the credit-weighted percentage of courses from which students are excluded and on GPA calculated using only non-excluded courses (table 7, columns 10 and 11). Tracking substantially increases the exclusion rate from 0.037 to 0.064 and reduces GPA in non-excluded courses by 0.077 standard deviations, though the latter effect is imprecisely estimated. I cannot calculate the hypothetical effect of tracking if all students were permitted to write exams but these results show that tracking matters in qualitatively similar ways at the intensive and extensive grading margins. This finding is consistent with the negative effect of tracking being concentrated on low-scoring students, who are most at risk of course exclusion. The importance of course exclusions also suggests that peer effects operate from early in the semester, rather than being concentrated during final exams.

8.4 Other Channels Linking Dormitory Assignment to GPA

I ascribe the effect of tracking on dormitory students' GPAs to changes in the distribution of peer groups. However, some other aspect of the dormitories or assignment policy may account for this difference. Dormitories differ in some of their time-invariant characteristics such as proximity to the main university campus and within-dormitory study space. The negative treatment effect of tracking is robust to dormitory fixed effects, which account for any relationship between dormitory features and GPA that is common across all types of students. Dormitory fixed effects do not account for potential interaction effects between student and dormitory characteristics. In particular, tracking would have a negative effect on low-scoring students' GPAs even without peer effects if there is a negative interaction effect between high school graduation test scores and the characteristics of low-track dormitories. I test this hypothesis by estimating equation 3 with an interaction term between HS_{id} and the rank of dormitory d during the tracking period. The interaction term has a small and insignificant coefficient (0.003 with cluster bootstrap standard error 0.006), showing that low-scoring students do not have systematically different GPAs when they are randomly assigned to previously low-track dormitories. This result is robust to replacing the continuous rank measure with an indicator for below-median rank dormitories. I conclude that the results are not explained by time-invariant dormitory characteristics.

This does not rule out the possibility of time-varying effects of dormitory characteristics or of effects of time-varying characteristics. I conducted informal interviews with staff in the university's Office of Student Housing and Residence Life to explore this possibility. There were no substantial changes to dormitories' physical facilities but there was some routine staff turnover, which I do not observe in my data. It is also possible that assignment to a low-track dormitory may directly harm low-scoring students through stereotype threat. Their dormitory assignment might continuously remind them of their low high school graduation test score and undermine their confidence or work ethic (Steele and Aronson, 1995). I cannot directly test this explanation and so cannot rule it out. However, the consistent results from the cross-policy and cross-dormitory analyses suggest that peer effects explain the bulk of the observed treatment effect of tracking, Wei (2009) also notes that evidence of stereotype threat outside laboratory conditions is rare.

9 Conclusion

This paper describes the effect on student GPAs of tracked dormitory assignment relative to random assignment at the University of Cape Town in South Africa. I show that under tracking the mean GPA was lower and the level of GPA inequality higher. This result arises because students' GPAs are higher when living with high-scoring peers but low-scoring students are more sensitive to residential peer group composition than high-scoring students. These peer effects arise largely through interaction with own-race peers and the relevant form of interaction does not appear to be direct academic collaboration. I present an extensive set of robustness checks supporting a causal interpretation for these results.

My findings demonstrate that different policies for assigning students to peer groups can have large effects on their academic performance. Academic tracking into residential groups, and perhaps other noninstructional groups, may generate a substantially worse distribution of academic performance than random assignment. However, caution should be exercised in using my results to judge holistically the relative merits of the two policies. Tracking clearly harms low-scoring students but some (imprecise) results suggest a positive effect on high-scoring students. Changing the assignment policy would then entail a transfer from one group of students to another and, as academic outputs are typically non-tradeable, it may not be possible to Pareto rank different policies. Many non-measured student outcomes may also be affected by different group assignment policies. For example, high-scoring students' performance may be unaffected by tracking because the rise in their peers' academic proficiency induces them to substitute time away from studying toward leisure. In future work I plan to study the long-term effects of tracking versus random assignment on graduation rates, time-to-degree, and labor market outcomes. These results will permit a more comprehensive evaluation of the relative merits of the two group assignment policies.

Despite these provisos, my findings provide important evidence regarding the importance of peer group assignment policies. I provide what appears to be the first well-identified 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 total effect of instructional tracking is positive, this may combine a negative direct peer effect of tracking with a positive effect due to changes in instructor behavior. My findings also suggest that policymakers can change the distribution of students' academic performance by rearranging the groups in which these students interact while leaving the marginal distribution of inputs into the education production function unchanged. This is attractive in any setting but particularly in resource-constrained developing countries. While the external validity of any result is always questionable, my findings may be particularly relevant to universities serving a diverse student body that includes both high performing and academically underprepared students. This is particularly relevant to selective universities with active affirmative action programs, such as those studied in Bertrand, Hanna, and Mullainathan (2010).

The examination of peer effects under random assignment in sections 6 and 7 also points to fruitful avenues for future research. Peer effects estimated under random assignment had limited ability to predict the effects of a change in assignment policy and residential peer effects appear to be mediated by students' patterns of interaction. This highlights the risk of relying too heavily on reduced form estimates that do not accurately capture the behavioral content of peer effects. Research that combines peer effects estimated under different peer group assignment policies with detailed data on social interactions and explicit models of network formation may provide additional insights.

References

Abadie, A. (2005): "Semiparametric Difference-in-Differences Estimators," *Review of Economic Studies*, 72, 1–19.

Abdulkadiroglu, A., J. Angrist, and P. Pathak (2011): "The Elite Illusion: Achievement Effects at Boston and New York Exam Schools," Working Paper 17264, National Bureau of Economic Research.

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.

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

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.

Becker, G. (1973): "A theory of marriage: Part I," *Journal of Political Economy*, 81, 813–846.

Benabou, R. (1996): "Equity and Efficiency in Human Capital Investment: The Local Connection," *Review of Economic Studies*, 63(2), 237–264.

Bertrand, M., R. Hanna, and S. Mullainathan (2010): "Affirmative Action in Education: Evidence from Engineering College Admissions in India," *Journal of Public Economics*, 94(1/2), 16–29.

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.

Bitler, M., J. Gelbach, and H. Hoynes (2010): "Can Variation in Subgroups' Average Treatment Effects Explain Treatment Effect Heterogeneity? Evidence from a Social Experiment," mimeo.

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.

Brock, W., and S. Durlauf (2001): "Interactions-based Models," in *Handbook of Econometrics Volume 5*, ed. by J. Heckman, and E. Leamer, pp. 3297–3380. Elsevier.

Brunello, G., and D. Checchi (2007): "Does School Tracking Affect Equality of Opportunity? New International Evidence," *Economic Policy*, 22(52), 781–861.

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

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., F. Malmstrom, and J. West (2008): "Peer Effects in Academic Cheating," *Journal of Human Resources*, XLIII(1), 173–207.

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

Cattaneo, M. (2010): "Efficient Semiparametric Estimation of Multi-Valued Treatment Effects under Ignorability," *Journal of Econometrics*, 155, 138–154.

Cooley, J. (2013): "Can Achievement Peer Effect Estimates Inform Policy? A View from Inside the Black Box," *Review of Economics and Statistics*, Forthcoming, University of Cambridge.

Crump, R., J. Hotz, G. Imbens, and O. Mitnik (2009): "Dealing with Limited Overlap in Estimation of Average Treatment Effects," *Biometrika*, 96(1), 187–199.

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.

Duncan, G., J. Boisjoly, M. Kremer, D. Levy, and J. Eccles (2005): "Peer Effects in Drug Use and Sex among College Students," *Journal of Abnormal Child Psychology*, 33(3), 375–385.

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.

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

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," Discussion Paper 4841, IZA.

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): "Mobility Treatment Effects: Identification, Estimation and Application," Working paper.

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 (2013): "Measuring the Average Outcome and Inequality Effects of Segregation in the Presence of Social Spillovers," Working Paper 16499, National Bureau of Economic Research.

Hanushek, E., J. Kain, and S. Rivkin (2009): "New Evidence about Brown v. Board of Education: The Complex Effects of School Racial Composition on Achievement," *Journal of Labor Economics*, 27(3), 349–383.

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

Heckman, J., and J. Hotz (1989): "Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training," *Journal of the American Statistical Association*, 84(408), 862–880.

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.

Higher Education South Africa (2009): "Report to the National Assembly Portfolio Committee on Basic Education," Available online at www.pmg.org.za/report/20090819-national-benchmark-tests-project-standards-national-examination-asses.

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," Mimeo.

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.

Hurder, S. (2012): "Evaluating Econometric Models of Peer Effects with Experimental Data," Mimeo.

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

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

Lavy, V., O. Silva, and F. Weinhardt (2012): "The Good, the Bad and the Average: Evidence on the Scale and Nature of Ability Peer Effects in Schools," *Journal of Labor Economics*, 30(2), 367–414.

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

McEwan, P. (2013): "Improving Learning in Primary Schools of Developing Countries: A Meta-Analysis of Randomized Experiments," Mimeo.

Meghir, C., and M. Palme (2005): "Educational Reform, Ability, and Family Background," *American Economic Review*, 95(1), 414–424.

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

Pischke, S., and A. Manning (2006): "Comprehensive versus Selective Schooling in England in Wales: What do we Know?," Working Paper 12176, National Bureau of Economic Research.

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

Robins, J., and A. Rotnitzky (1995): "Semiparametric Efficiency in Multivariate Regression Models with Missing Data," *Journal of the American Statistical Association*, 90(429), 122–129.

Rothe, C. (2010): "Nonparametric Estimation of Distributional Policy Effects," *Journal of Econometrics*, 155(1), 56–70.

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.

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

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

Steele, C., and J. Aronson (1995): "Stereotype Threat and the Intellectual Test Performance of African Americans," *Journal of Personality and Social Psychology*, 69(5), 797–811.

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.

Wei, T. (2009): "Stereotype Threat, Gender, and Math Performance: Evidence from the National Assessment of Educational Progress," Mimeo.

A Reweighted Nonlinear Difference-in-Differences Model (For Online Publication)

Athey and Imbens' (2006) nonlinear difference-in-differences model constructs quantile treatment on the treated effects in a difference-in-differences setting. This provides substantially more information than the standard linear difference-in-differences model, which constructs only the average treatment effect on the treated. However, the model requires stronger identifying assumptions and the original model provides a limited way to deal with covariates.

The original model is identified under four assumptions. Define T as an indicator variable equal to one in the tracking period and zero in the random assignment period and D as an indicator variable equal to one for dormitory students and zero for non-dormitory students. The identifying assumptions are:

(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 that of non-dormitory students' GPAs: $supp(GPA|D = 1) \subseteq supp(GPA|D = 0)$.

(A4) The distribution of GPA is strictly continuous.[43]

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. The q^{th} quantile treatment effects of tracking on the treated students is defined as the horizontal difference between the observed and counterfactual distributions at quantile q: $F^{-1}_{GPA|D=1,T=1}(q) - F^{CF,-1}_{GPA|D=1,T=1}(q)$.

These identifying assumptions may hold conditional on some covariates $X_1, \ldots, X_k$ but not unconditionally. In my application, some of the demographic characteristics shown in table 1 are not stable through time. If these

[43] The GPA measure I use is approximately continuous. There are 5215 unique values of GPA, so each value accounts for an average of 0.02% of the total mass. The most common value accounts for only 0.26% of the total mass.

covariates are determinants of GPA, then the stationarity assumption *A2* is unlikely to hold when the covariates are included in U. The assumption may, however, still be valid after conditioning on the covariates.

Athey and Imbens briefly propose two ways to include observed covariates in the model. First, a nonparametric method that applies the model separately to each value of the covariates. This is feasible only if there are a small number of discrete covariates. Second, a parametric method that residualizes GPA by regressing it on the covariates and applies the model to the residuals. This is valid only under the strong assumption that the observed covariates X and unobserved scalar U are additively separable in the GPA production function. It is also valid only if the functional form of the covariates used for residualization is correctly specified. If either assumption fails, the residualization scheme will not recover consistent estimates of the quantile treatment effects.

I instead use a reweighting scheme that avoids the assumption of additive separability and may be more robust to specification errors. Specifically, I define the reweighted counterfactual distribution as

$$F^{RW,CF}_{GPA^{11}}(g) = F_{GPA^{10}_{\omega}}\left(F^{-1}_{GPA^{00}_{\omega}}\left(F_{GPA^{01}}\left(g\right)\right)\right) \tag{6}$$

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

$$\hat{\tau}^{QTT}(q) = \hat{F}^{-1}_{GPA^{11}}(q) - \hat{F}^{-1,RW,CF}_{GPA^{11}}(q) \tag{7}$$

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) suggests 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 3, 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 (6) 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.

The Stata code for implementing this estimator is available on my website. I do not attempt to estimate the counterfactual minimum and maximum, as inference on these parameters is known to be highly problematic (Horowitz, 2001).