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

Robert Garlick

Duke University

March 3, 2016

ERID Working Paper Number 220

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

Robert Garlick†

March 3, 2016

Abstract

I study the relative academic performance of students tracked or randomly assigned to South African university dormitories. Tracking reduces low-scoring students' GPAs but has little effect on high-scoring students. This lowers mean GPA and raises GPA dispersion. I also directly estimate peer effects using random variation in peer groups across dormitories. Living with higher-scoring peers raises students' GPAs and this effect is larger for low-scoring students. Peer effects operate largely within race groups but operate both within and across programs of study. This suggests that spatial proximity alone does not generate peer effects. Interaction of some sort is required, but direct academic collaboration is not the relevant form of interaction. I integrate the results from variation in group assignment policies and variation in group composition by drawing on the matching and sorting literatures. Both sets of results imply that own and peer academic performance are substitutes in GPA production and that GPA may be a concave function of peer group performance. The cross-dormitory results correctly predict a negative effect of tracking on low-scoring students but understate the magnitude of the observed effect. I show that this understatement reflects both policy-sensitive parameter estimates and problems with extrapolation outside the support of the data observed under random assignment. This underlines the value of using both cross-policy and cross-group variation to study peer effects.

Keywords: education; inequality; peer effects; tracking
JEL classification: I23; I24; I25; O15

*I am particularly grateful to Jeff Smith, David Lam, John DiNardo, Brian Jacob, and Manuela Angelucci for advice and guidance throughout the preparation of this paper. I also appreciate helpful suggestions from Raj Arunachalam, Emily Beam, John Bound, Tanya Byker, Scott Carrell, Susan Godlonton, Andrew Goodman-Bacon, Italo Gutierrez, Brad Hershbein, Stephen Ross, Rebecca Thornton, Adam Wagstaff, Dean Yang, and seminar participants at ASSA 2014, 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 Jane Hendry, Josiah Mavundla, and Charmaine January at the University of Cape Town. I acknowledge financial support from the University of Michigan's Gerald R. Ford School of Public Policy and Horace H. Rackham School of Graduate Studies. All errors are my own.

†Duke University; robert.garlick@duke.edu

1 Introduction

Group structures are ubiquitous in education and group composition may have important effects on education outcomes. Students in different classrooms, living environments, schools, and social groups are exposed to different peer groups, receive different education inputs, and face different institutional environments. A growing literature shows that students' peer groups influence their education outcomes even without resource and institutional differences across groups.[1] Peer effects play a role in empirical and theoretical research on different ways of organizing students into classrooms and schools.[2] Most studies focus on the effect of assignment or selection into different peer groups for a given group assignment or selection process.[3]

This paper advances the literature by asking a subtly different question: What are the relative effects of two group assignment policies – randomization and tracking or streaming based on academic performance – on the distribution of student outcomes? This contributes to a small but growing empirical literature on optimal group design. Comparison of different group assignment policies corresponds to a clear social planning problem: How should students be assigned to groups to maximize some target outcome, subject to a given distribution of student characteristics? Different group assignment policies leave the marginal distribution of education inputs unchanged. This raises the possibility of improving academic outcomes with few pecuniary costs. Such low cost education interventions are particularly attractive for resource-constrained education systems.

I study peer effects under two different group assignment policies at the University of Cape

[1] Manski (1993) lays out the identification challenge in studying peer effects: do correlated outcomes within peer groups reflect correlated unobserved pre-determined characteristics, common institutional factors, or peer effects – causal relationships between students' outcomes and their peers' characteristics? Many education researchers address this challenge using randomized or controlled variation in peer group composition. Peer effects 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] Examples include 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) on 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.

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 residential peer groups that were respectively homogeneous and heterogeneous in baseline academic performance. I contrast the distribution of first year students' academic outcomes 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. Low-scoring students perform substantially worse under tracking than random assignment, while high-scoring students' GPAs are approximately equal under the two policies. Measures of inequality are substantially higher under tracking than random assignment.

I then use randomly assigned dormitory-level peer groups to estimate directly the effect of living with higher- or lower-scoring peers. I find that all students' GPAs are increasing in the baseline academic performance of their peers. Peer effects operate largely within race groups, suggesting that social proximity mediates peer effects, but do not appear to operate through direct academic collaboration. I estimate a reduced-form but theory-guided model of nonlinear peer effects and find that low-scoring students are more strongly affected by their peer group composition than high-scoring students. I use these peer effects estimated under random assignment to predict the effects of tracking. The predictions are qualitatively consistent with the observed effects of tracking but underestimate the magnitude of the negative effect. This underestimation may be due to policy-sensitive parameters (economic misspecification) or parameters that are only valid within the support of the data generated under random assignment (statistical misspecification). I develop a simple test to separate these two forms of misspecification and conclude that estimating peer effects under random assignment, even when guided by economic theory and model selection criteria, yields estimates that are both economically and statistically misspecified.

This paper contributes to two literatures in economics, on peer effects in education and on academic tracking. I contribute to the peer effects literature by combining variation in peer

group assignment policy with random variation in peer group composition. This is closely related to work by Carrell, Sacerdote, and West (2013). I show that student outcomes are affected both by residential peers' characteristics and by changes in the peer group assignment policy. Both cross-policy and cross-dormitory comparisons show that low-scoring students are more sensitive to changes in peer group composition. I link this finding to the literatures on optimal group composition in the presence of spillovers (Benabou, 1996) and on marriage matching (Becker, 1973). Optimal matching and group composition in these models depend on whether own and peer academic performance are complements or substitutes in GPA production and whether GPA is a convex or concave function of peer academic performance. I find robust evidence of substitutability and mixed evidence of concavity. These parameters have received limited attention in the peer effects literature other than Graham, Imbens, and Ridder (2010) and this is the first finding of substitutability of which I am aware.[4] I also consider the extent to which reduced-form but theory-guided models estimated using random variation in peer group composition can predict the effects of alternative peer group assignment policies. I find these models correctly predict the negative effects of tracking but understate the magnitude. This appears to reflect both economic and statistical model misspecification and reinforces challenges in using peer effects for predicting policy effects raised by Bhattacharya (2009), Carrell, Sacerdote, and West (2013), Graham, Imbens, and Ridder (2010), and Hurder (2012).

I also find that peer effects operate almost entirely within race groups, consistent with results from Hanushek, Kain, and Rivkin (2009) and Hoxby (2000). However, dormitory peer effects in this setting 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. I believe that this finding is novel in the peer effects literature. Taken together, these results suggest that spatial proximity generates peer effects only when students are also socially proximate and likely to interact. But the relevant form of interaction is not

[4] Hoxby and Weingarth (2006) provide a general taxonomy of peer effects other than the linear-in-means model studied by Manski (1993). Several previous studies have found evidence that own and peer characteristics are complements (Booij, Leuven, and Oosterbeek, 2014; Burke and Sass, 2013; Cooley, 2014; Imberman, Kugler, and Sacerdote, 2012; Lavy, Silva, and Weinhardt, 2012).

direct academic collaboration. Peer effects may instead operate through mechanisms such as time use or transfer of soft skills, consistent with Stinebrickner and Stinebrickner (2006).

I contribute to the literature on academic tracking by isolating the role of peer effects in tracking. Most existing papers estimate the effect of school or classroom tracking relative to another assignment policy or of assignment to different tracks.[5] However, tracked and untracked groups may differ on multiple dimensions: peer group composition, instructor behavior, and school resources (Betts, 2011; 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. For example, Duflo, Dupas, and Kremer (2011) show that tracking in Kenyan primary schools has a positive effect on low-track students and argue this is because the benefits of targeted instruction outweigh the cost of weaker peers. 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. Variation in dormitory-level characteristics might in principle affect student outcomes but my results are entirely robust to conditioning on these characteristics.

I argue that my findings are of general interest for three reasons, even though this specific policy decision, tracking or randomly assigning students into dormitories, may not be common. First, the results can be viewed as a mechanism experiment, using the language from Ludwig, Kling, and Mullainathan (2011). I shed light on the nonlinear structure of peer effects and point to an important role for peer effects in classroom tracking studies. Second, prior studies have found substantially stronger residential peer effects on academic outcomes than classroom or study group peer effects (Hoel, Parker, and Rivenburg, 2004; Jain and Kapoor, 2015). This suggests that composition of residential peer groups should be an important policy question. Third, some degree of residential segregation by academic performance may actually be common. Many US universities have "honors colleges" or "honors dormitories," reserved

[5] Betts (2011) reviews an extensive literature comparing tracking to alternative assignment policies. A smaller literature studies the effect of assignment to different tracks in an academic tracking system (Abdulkadiroglu, Angrist, and Pathak, 2011; Ajayi, 2014; Lucas and Mbiti, 2014; Pop-Eleches and Urquiola, 2013).

for students with very strong prior academic performance. Other universities allow students to request specific dormitory allocations after their first year or charge different prices for living in different dormitories, which may generate relatively academically segregated dormitories. I am not aware of any systematic data on the extent of these practices but purely anecdotal evidence suggests that they are common.

My analysis of distributional consequences of peer group composition and assignment rules draws on the treatment effects literature in econometrics. Peer effects and tracking research strongly emphasizes inequality considerations but almost no papers measure inequality (Betts, 2011; Epple and Romano, 2011). I note that an inequality treatment effect of tracking can be obtained using standard methods for quantile treatment effects (Firpo, 2007; Heckman, Smith, and Clements, 1997) and extensions proposed by Firpo (2010) and Rothe (2010). My research design uses difference-in-differences methods to remove time trends in student performance between the tracking and random assignment periods. I therefore use a nonlinear difference-in-differences model (Athey and Imbens, 2006) to obtain quantile and then inequality treatment effects. 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. In section 3 I present the average effects of tracking for the entire sample and for students with different baseline characteristics, including academic performance. I discuss the effects of tracking on the entire GPA distribution in section 4 and show the resultant effects on academic inequality in section 5. I then explore the effects of random assignment to live with higher- or lower-scoring peers in section 6. I present a framework to compare the cross-policy and cross-dormitory results in section 7. In section 8, I test and largely reject alternative explanations for the apparent effects of tracking: time-varying student selection into dormitory or non-dormitory status, differential time trends in student performance between dormitory and non-dormitory students, spillover effects of

tracking on non-dormitory students, limitations of GPA as an outcome measure, and direct effects of dormitory assignment on GPAs that do not reflect peer effects. I conclude in section 9 and outline the conditional nonlinear difference-in-differences model in appendix A.

2 Research Design and Data

I study a natural experiment at the University of Cape Town in South Africa, where first-year students are allocated to dormitories using either random assignment or academic tracking. This is a selective research university but admits many students from low-performing high 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.[6] The dormitories provide accommodation, meals, and some organized social activities. Classes and instructors are shared across 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 can move out of their dormitory after one year but cannot change to another dormitory. Dormitory assignment thus determines students' residential peer groups in their first year of university; the second year peer group depends on students' location choices. 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 region.

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.[7] Students with high and low scores on this examination were assigned

[6] The mean dormitory size is 123 students and the interdecile range is 50 – 216. There are 16 dormitories in total, one of which closes in 2006 and one of which opens in 2007. I exclude seven very small dormitories that each hold fewer than 10 first-year students.

[7] These tests are developed and moderated by a statutory body reporting to the Minister of Education. The tests are nominally criterion-referenced. Students select six subjects in grade 10 in which they will be

to different dormitories. 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 each dormitory tier 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.[8] 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.[9]

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.[10] Assignment used a random number generator with *ex post* changes to avoid racial imbalance.[11] 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, resulting in a disproportionate number of low-scoring students under both tracking and randomization. Results are robust to excluding this dormitory.

The policy change induced a large change in students' peer groups. Figure 1 shows how the relationship between students' own high school graduation test scores and their peers' test scores

tested in grade 12. The university converts their subject-specific letter grades into a single score for admissions decisions. A time-invariant conversion scale is used to convert international students' A-level or International Baccalaureate scores into a comparable metric.

[8] The overlap too large to use a regression discontinuity design to study the effect of assignment to higher- or lower-track dormitories.

[9] It is not obvious how different the results might be without these departures from tracking. If the treatment effects of tracking depend only on peers' mean high school graduation test scores, then the treatment effects I estimate will be attenuated relative to the effects of pure tracking. However, peer effects may differ within and across gender or race subgroups, as I discuss in section 6. In this case, a tracking system that does not deliberately create racially mixed dormitories might have quite different effects to those that I estimate.

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

[11] 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.

Table 1: Effects of Tracking on Peer Group Composition

	(1) Entire sample	(2) Tracked dorm students	(3) Randomized dorm students	(4) Difference (2)-(3)	(5) p-value (2)=(3)
Mean of HS scores in dormitory	0.198	0.192	0.204	0.012	0.323
Variance of HS scores in dormitory	0.801	0.740	0.857	0.116	0.000
Interquartile range of HS scores in dormitory	1.344	1.244	1.435	0.191	0.000
Fraction of dormmates of own gender	0.917	0.900	0.933	0.033	0.000
Fraction of dormmates of own race	0.424	0.436	0.413	-0.023	0.000
Fraction of dormmates of own language	0.349	0.379	0.320	-0.059	0.000
Fraction of dormmates in own faculty	0.280	0.284	0.275	-0.009	0.013

Notes: Table 2 reports summary statistics on dormitory composition, for all dormitory students entire sample (column 1), tracked dormitory students (column 2) and randomly assigned dormitory students (column 3). The p-values in column 4 are from testing whether the level of each characteristic is equal in the tracking and randomization periods. The statistic "fraction of dormmates of own gender" is defined at the student level and equals, for a female/male student, the proportion of female/male students in her/his dormitory. The other three "fraction of ..." statistics are defined in the same way.

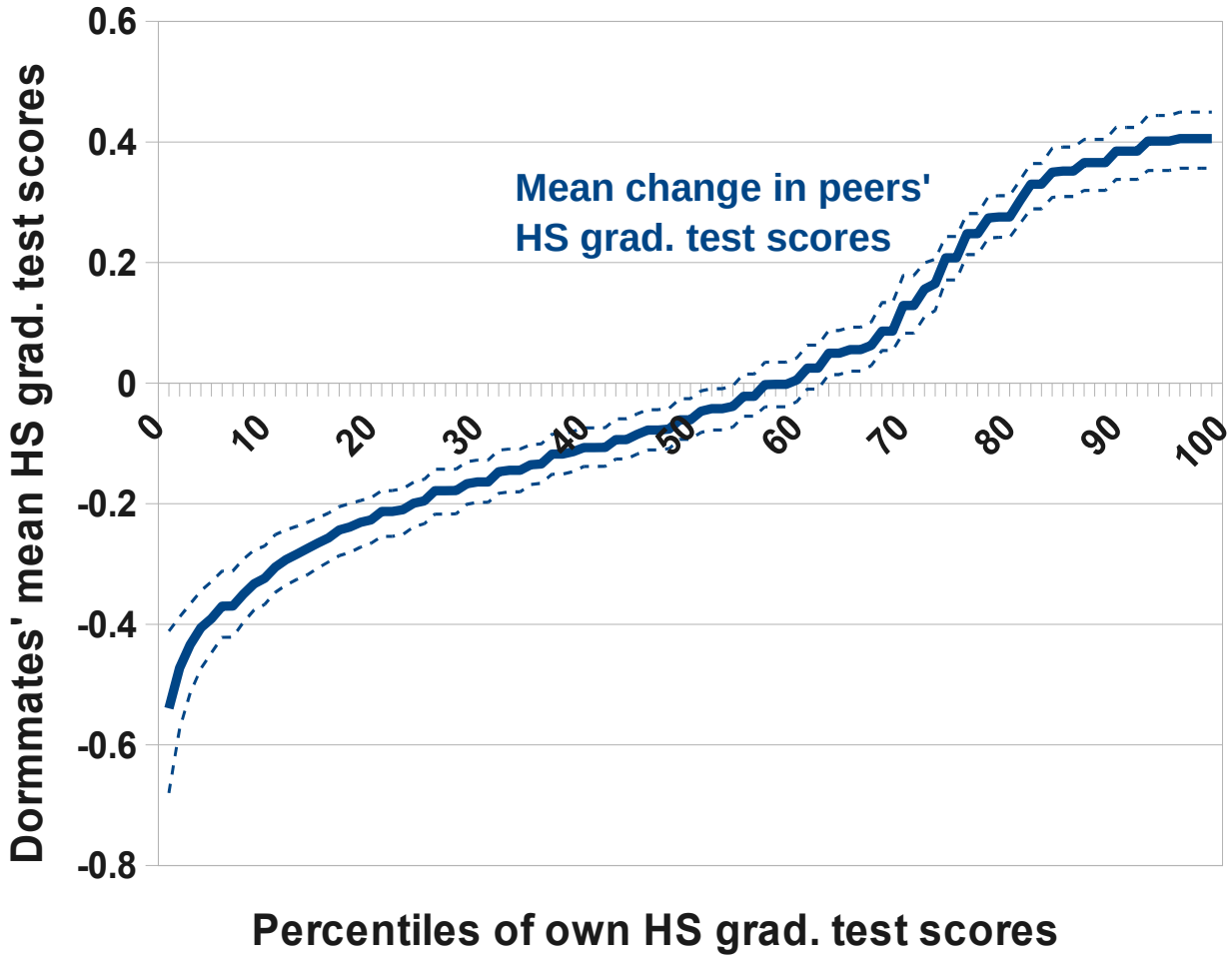
changed. For example, students in the top decile lived with peers who scored approximately 0.4 standard deviations higher under tracking than random assignment; students in the bottom decile lived with peers who scored approximately 0.4 standard deviations lower. This is the identifying variation I use to study the effect of tracking. Table 1 shows that the within-dormitory dispersion of high school graduation test scores is substantially lower under tracking (rows 2-3).[12] Peer groups under tracking are also slightly more homogeneous in terms of race, language, and faculty/college/school of study (rows 5-7). However, cross-dormitory and cross-policy variation in these proportions is not robustly associated with differences in students' outcomes and all results in the paper are robust to controlling for these proportions.

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.[13] 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

[12] The within-dormitory variance and interquartile range remain quite high under tracking because the distribution of high school graduation test scores has a long left tail. The proportion of own-gender peers is also slightly higher under random assignment because a mixed-gender dormitory was closed and a female-only dormitory opened in the random assignment period.

[13] 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: The curve is constructed in three steps. First, I estimate a student-level local linear regression of mean dormitory high school test scores on students' own test scores, separately for tracked and randomly assigned dormitory students. Second, I evaluate the difference between the fitted values at each percentile of the test score distribution. Third, I use a percentile bootstrap with 1000 replications to construct the 95% confidence interval, stratifying by assignment policy.

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

$$GPA_{id} = \beta_0 + \beta_1 \text{Dorm}_{id} + \beta_2 \text{Track}_{id} + \beta_3 \text{Dorm}_{id} \times \text{Track}_{id} + f\left(\vec{X}_{id}\right) + \vec{\mu}_d + \epsilon_{id} \qquad (1)$$

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,[14] 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 a "parallel trends" assumption: that dormitory 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 1 requires only that the parallel trends assumption holds conditional on student covariates and dormitory fixed effects. I also estimate model 1 with inverse probability weights that reweight each group of students to have the same distribution of covariates as the tracked dormitory students.[15]

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 assignment policy change affects students through mechanisms other than peer effects, β_3 recovers the correct treatment effect but its interpretation changes. I discuss these concerns in section 8.

[14] I use a quadratic specification. The results are similar with linear or cubic $f(\cdot)$.

[15] Unlike the regression-adjusted model 1, 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 results in section 3 are very similar in practice. Abadie (2005) and Cattaneo (2010) discuss reweighted difference-in-differences models and derive appropriate weights for treatment-on-the-treated parameters.

The data on students' demographic characteristics and high school test scores (reported in table 2) are broadly consistent with the assumption of parallel time trends. Dormitory students have on average slightly higher and more dispersed scores than non-dormitory students on high school graduation tests (panel A).[16] 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). There are time trends in several characteristics (columns 4 and 7) but these are not significantly different between dormitory and non-dormitory students (column 8). The notable exception is that the proportion of 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. 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 restricting the number of students who live in Cape Town who may be admitted to the dormitory system.[17] 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

[16] 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.

[17] I do not observe students' home addresses, which are used for the university's dormitory admissions. Instead, I match records on students' high schools to a public database of high school GIS codes. I then determine whether students attended high schools in or outside the Cape Town metropolitan area. This is an imperfect proxy of their home address for three reasons: long commutes and boarding schools are fairly common, the university allows students from very low-income neighborhoods on the outskirts of Cape Town to live in dormitories, and a small number of Cape Town students with medical conditions or exceptional academic records are permitted to live in the dormitories.

Table 2: Summary Statistics and Balance Tests

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Entire sample	Track dorm	Random dorm	p-value (2)=(3)	Track non-dorm	Random non-dorm	p-value (5)=(6)	p-value (2)-(3)=(5)-(6)
Panel A: Standardized high school graduation test scores								
Mean score	0.088	0.169	0.198	0.277	0.000	0.000	1.000	0.426
A on graduation test	0.278	0.320	0.325	0.638	0.222	0.253	0.003	0.108
≤C on graduation test	0.233	0.224	0.201	0.029	0.254	0.250	0.723	0.198
Panel B: Demographic characteristics								
Female	0.513	0.499	0.517	0.118	0.523	0.514	0.451	0.103
Black	0.319	0.503	0.524	0.071	0.116	0.118	0.751	0.181
White	0.423	0.354	0.332	0.044	0.520	0.495	0.033	0.851
Other race	0.257	0.143	0.144	0.885	0.364	0.387	0.047	0.124
English-speaking	0.714	0.593	0.560	0.004	0.851	0.863	0.125	0.001
International	0.144	0.225	0.180	0.000	0.106	0.061	0.000	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 | ≤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	0.399	0.765	0.754	0.289	0.657
Cape Town | A student	0.414	0.101	0.065	0.005	0.848	0.811	0.058	0.976
Cape Town | ≤C student	0.523	0.146	0.186	0.079	0.798	0.800	0.927	0.224

Notes: Table 2 reports summary statistics of student characteristics at the time of enrollment, for the entire sample (column 1), dormitory students in the tracking period (column 2), dormitory students in the random assignment period (column 3), non-dormitory students in the tracking period (column 5), and non-dormitory students in the random assignment period (column 6). The p-values in columns 4 and 7 are from testing whether the level of each characteristic is equal in the tracking and randomization periods, respectively for dormitory and non-dormitory students. The p-values reported in column 8 are from testing whether the mean change in each variable between the tracking and random assignment periods is equal for dormitory and non-dormitory students.

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.[18]

The primary outcome variable is first-year students' GPAs. 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 students, separately by year.[19]

[18] 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.

[19] The results are unchanged when I use raw GPA or pool all non-dormitory students rather than standardizing separately by year. See footnote 23 for details.

The effects of tracking 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."[20] 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 3 column 1). The 95% confidence interval is [-0.27, 0.01]. The treatment effect is -0.13 standard deviations with a 95% confidence interval [-0.21, -0.05] after controlling for dormitory fixed effects, student demographics, and high school graduation test scores (column 2).[21] The average effect of tracking is thus negative and robust to accounting for dormitory fixed effects and student covariates.[22] This pattern holds for all results reported in the paper: accounting for student and dormitory characteristics yields narrower confidence intervals and unchanged treatment effect estimates.

How large is a treatment effect of -0.13 standard deviations? This is substantially smaller

[20] For example, mean percentage scores on Economics 1 and Mathematics 1 change by respectively six and nine points from year to year, roughly half of a standard deviation.

[21] The bootstrapped standard errors reported in table 3 allow clustering at the dormitory-year level. Non-dormitory students are treated as individual clusters, yielding 60 large clusters and approximately 7000 singleton clusters. As a robustness check, I also use a wild cluster bootstrap (Cameron, Miller, and Gelbach, 2008). The p-values are 0.090 for the basic regression model (column 1) and < 0.001 for the model with dormitory fixed effects and student covariates (column 3). I also 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 for the models in columns 1 and 3.

[22] The regression-adjusted results in column 3 use missing data indicators for students with missing high school-graduation test scores. I also estimate the treatment effect with these observations excluded and find a similar result (column 2). Effects estimated with both regression adjustment and inverse probability weighting are marginally larger (columns 4 and 5). Trimming propensity score outliers following Crump, Hotz, Imbens, and Mitnik (2009) yields similar but less precise point estimates. This verifies 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 so this robustness check is not conclusive for the average treatment effect on the treated with a difference-in-differences design.

Table 3: 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 3 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 controlling for a missing test score indicator. Columns 4 and 5 report results from propensity score-weighted regressions 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, stratifying by dormitory status and assignment policy, and re-estimating the weights on each iteration.

than the black-white GPA gap at this university (0.46 standard deviations) but larger than the male-female GPA gap (0.09).[23] 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). These results 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.[24]

Tracking changes peer groups in different ways for different students: 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' high school test scores. I explore

[23] This magnitude is robust to using different GPA scales. Using raw GPA, the treatment effect is -2.15 points on a 0-100 scale, while the black-white and male-female gaps in raw GPA are 6.75 and 1.22 respectively. The treatment effect is -0.14 standard deviations when I standardize with respect to the pooled set of all non-dormitory students, rather than standardizing each year separately. The black-white and female-male gaps in this GPA measure are 0.08 and 0.45 respectively.

[24] See McEwan (2013) for a metastudy of effect sizes in education interventions. He finds average effects across studies of 0.12 for class size and composition interventions and 0.06 for school management or supervision interventions, though standard disclaimers about comparability across settings apply.

this heterogeneity in two ways. I first estimate conditional average treatment effects for different subgroups of students. I then estimate quantile treatment effects of tracking in section 4, 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 0.01 and 0.24 standard deviations under tracking (cluster bootstrap standard errors 0.06 and 0.07; p-value of difference 0.014). However, above- and below-median students experience "treatments" of similar magnitude: they have residential peers who score on average 0.20 standard deviations higher and 0.27 standard deviations lower under tracking. This is not consistent with a linear response to changes in mean peer quality.[25] Either low-scoring students are more sensitive to changes in their mean peer group composition or GPA depends on some measure of peer quality other than mean test scores.

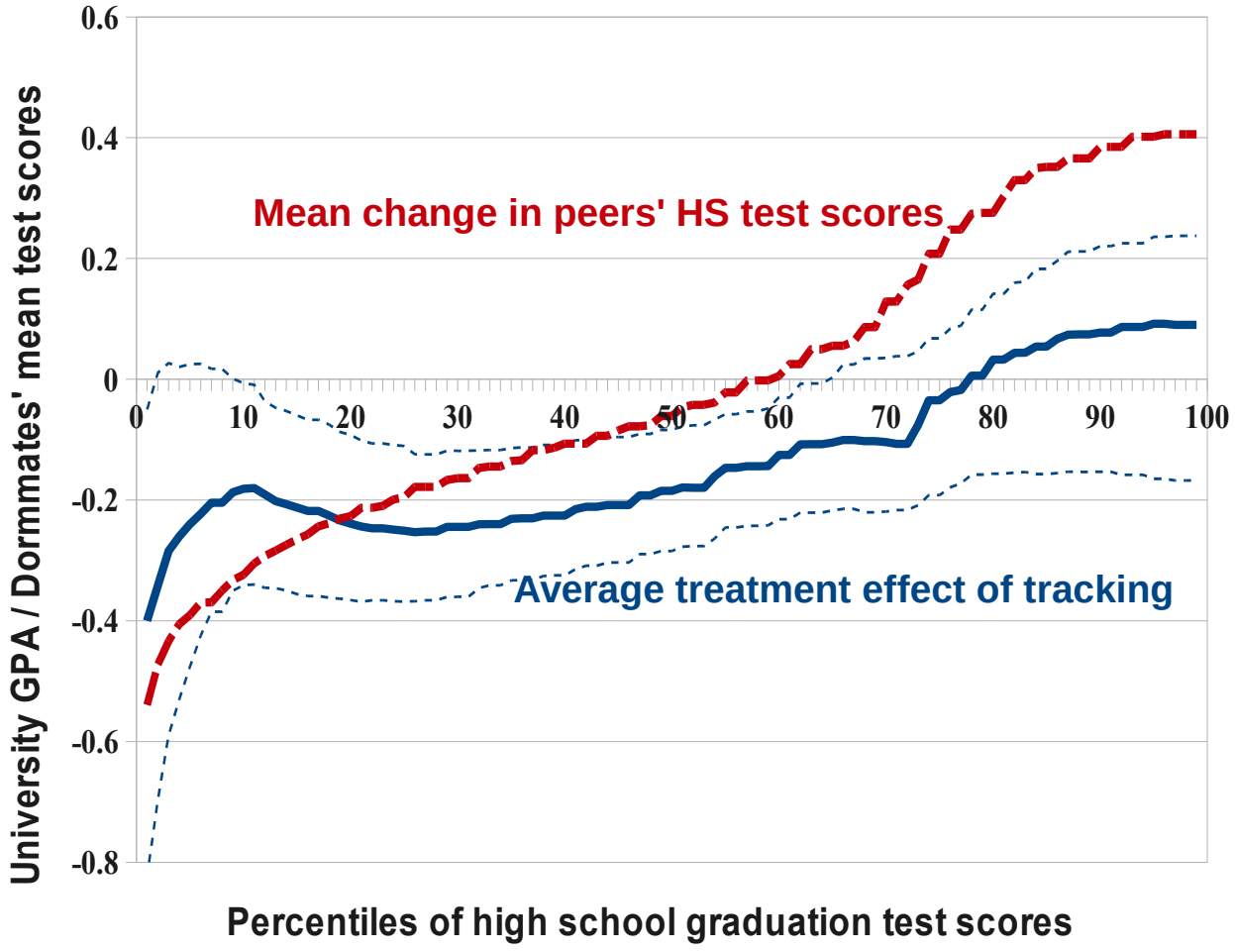
The near-zero treatment effect on above-median students is perhaps surprising. Splitting the sample in two may be too coarse to discern positive effects on very high-scoring students. I therefore estimate treatment effects throughout the distribution of high school test scores. Figure 2 shows that tracking reduces GPA through more than half of the distribution. The negative effects in the left tail are considerably larger than the positive effects in the right tail, though they are not statistically different. Figure 2 also shows the change in mean peer high school test scores (from figure 1). I reject equality of the treatment effects and changes in peer high school test scores in the right 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. Tracking may have a small positive effect on students in the top quartile but this effect is imprecisely estimated.[26]

There is stronger evidence of heterogeneity across high school test scores than demographic

[25] If student GPA is a linear function of peers' mean high school graduation test scores, then the ratio $\Delta GPA/\Delta HS$ should be constant. In particular, this ratio should be equal for above- and below-median students. I reject this hypothesis with a cluster bootstrap p-value of 0.070. This result further motivates my analysis of nonlinear peer effects in section 7.

[26] 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 2 is constructed by estimating a student-level local linear regression of GPA against high school graduation test scores. I estimate the regression separately for each of the four groups (tracking/randomization policy and dormitory/non-dormitory status). I evaluate the second difference between the fitted values 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 and stratifying by assignment policy and dormitory status. The dashed line shows the effect of tracking on mean peer group composition, discussed in figure 1.

subgroups. Treatment effects are larger on black than white students: -0.20 versus -0.11 standard deviations. However, this difference is not significant (cluster bootstrap p-value 0.488) and is almost zero after conditioning on high school test scores. I also estimate a quadruple-differences model allowing the effect of tracking to differ across four race/academic subgroups (black/white $\times$ 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 students within each test score group. However, neither pattern is significant at any conventional level. I thus lack the power to detect any heterogeneity by race conditional on test scores. There is no evidence of gender heterogeneity: tracking lowers female and male GPAs by 0.14 and 0.12 standard deviations respectively (cluster bootstrap p-value 0.897). I conclude that high school test scores are the primary dimension of treatment effect heterogeneity.

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 GPA distribution, following Athey and Imbens (2006). I first construct the counterfactual GPA distribution that the tracked dormitory students would have obtained in the absence of tracking (figure 3, first panel). The horizontal distance between the observed and counterfactual GPA distributions at each quantile equals the quantile treatment effect of tracking on the treated students (figure 3, second panel). 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 ($\leq$ 0.2 standard deviations), and small and positive in the top quintile ($\leq$ 0.2 standard deviations). The estimates are relatively imprecise; the 95% confidence interval excludes zero only in the first quintile.[27] This reinforces the pattern that the negative

[27] 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.

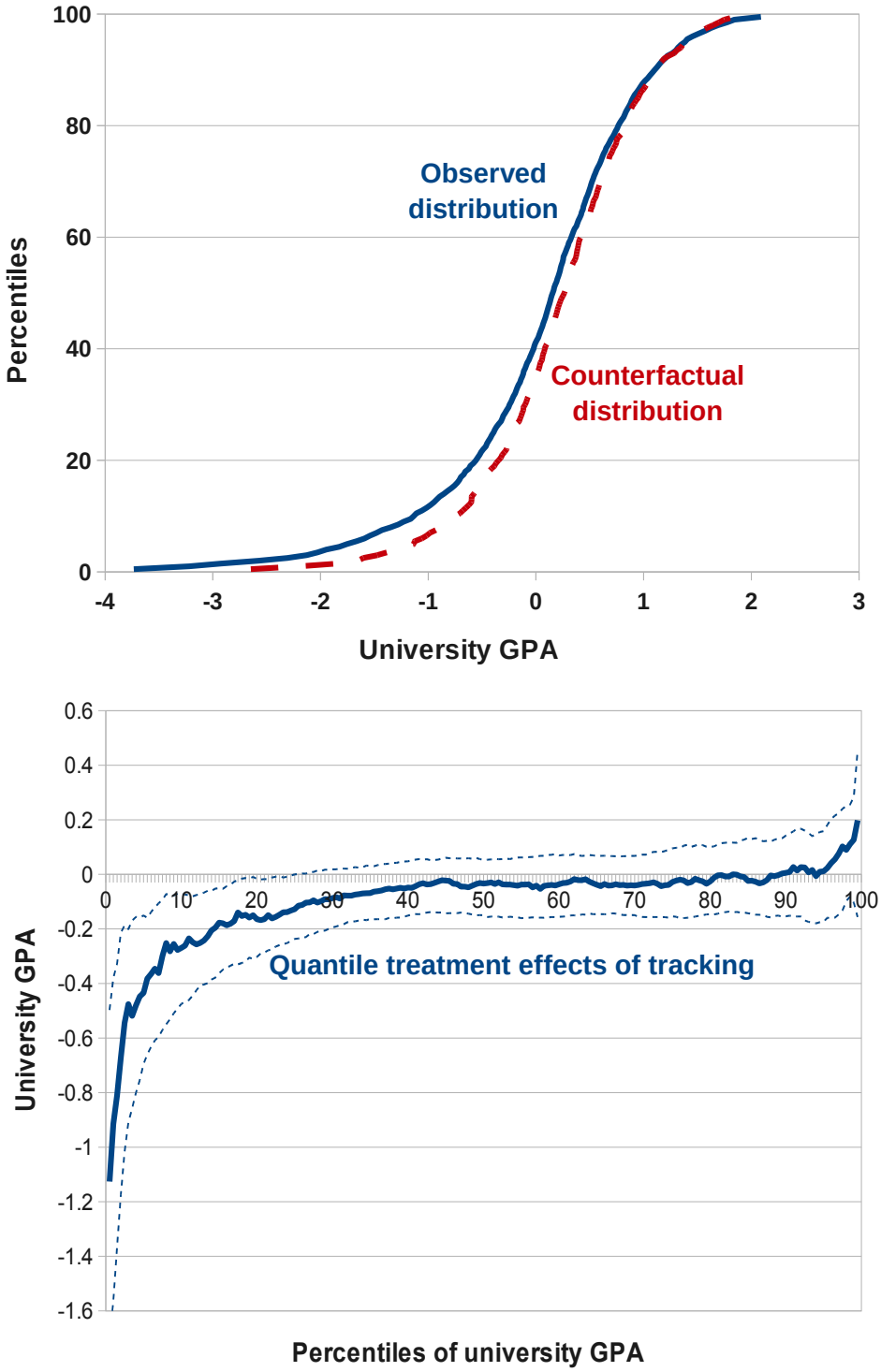
average effect of tracking is driven by large negative effects on the left tail of the GPA or high school test score distribution.

This provides substantially more information than the average treatment effect but requires stronger identifying assumptions. Specifically, the average effect is identified under the assumption that any time changes in the mean value of unobserved GPA determinants are common across dormitory and non-dormitory students. The quantile effects in a difference-in-differences model 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 (Athey and Imbens, 2006). These results should be interpreted with caution because there is some evidence of time trends in the characteristics listed in table 2, casting doubt on the identifying assumptions of the nonlinear difference-in-differences model. I propose an extension of the Athey-Imbens model to account flexibly for time trends in observed student characteristics using reweighting and discuss the implementation of this model in appendix A. The estimated quantile treatment effects are very similar with and without reweighting the data to adjust for these time trends.

There is no necessary relationship between figures 2 and 3. Figure 2 shows that the average treatment effect of tracking is large and negative for students with low high school graduation test scores. Figure 3 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 do not recover treatment effects on specific students or groups of students without additional assumptions. See Bitler, Gelbach, and Hoynes (2010) for further discussion on this relationship.[28]

[28] 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 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 random dormitory assignment, relative to tracking, helps low-scoring students to catch-up to their high-scoring peers but does not facilitate overtaking.

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



Notes: The first panel shows the observed GPA distribution for tracked dormitory students (solid line) and 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, stratifying by assignment policy and dormitory status and re-estimating the weights on each iteration.

Table 4: 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	12.79
	(0.043)	(0.047)	(0.062)	(6.8)
Interdecile range	2.238	1.857	0.381	20.52
	(0.083)	(0.091)	(0.109)	(5.9)
Standard deviation	0.909	0.766	0.143	18.67
	(0.027)	(0.032)	(0.037)	(4.8)

Notes: Table 4 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 the tracked students. Column 4 shows the treatment effect expressed as a percentage of the counterfactual level. Standard errors in parentheses are from 1000 bootstrap replications clustering at the dormitory-year level and stratifying by assignment policy and dormitory status.

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.[29] The literature on academic tracking emphasizes 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.

Table 4 shows inequality measures for the observed and counterfactual GPA distributions. The standard deviation and interquartile and interdecile ranges are all significantly higher under tracking than under the counterfactual.[30] 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 lowest quantiles of the GPA distribution.

[29] 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 (cluster bootstrap standard error 0.07). This is consistent with the average effect from the linear difference-in-differences models in section 3.

[30] I do not calculate other common inequality measures such as the Gini coefficient and Theil index because standardized GPA is not a strictly positive variable.

Tracking thus decreases mean academic outcomes and increases academic inequality. 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 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 peer group composition induced by random assignment. I first use a standard test to confirm the presence of residential peer effects, providing additional evidence that the main results are not driven by confounding factors. I document differences in dormitory-level peer effects within and between demographic and academic subgroups, 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, however, understate the effects of tracking.

I first estimate the standard linear-in-means model (Manski, 1993):

$$GPA_{id} = \alpha_0 + \alpha_1 HS_{id} + \alpha_2 \overline{HS}_d + \vec{\alpha}\vec{X}_{id} + \vec{\mu}_d + \epsilon_{id}, \tag{2}$$

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

Table 5: 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						
Student covariates		×	×		×	×
Dormitory fixed effects			×			×
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 5 reports results from estimating the linear peer effects model in 2 (columns 1-3) and the quadratic peer effect model in equation 4 (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.

high school graduation test scores of one's residential peers.[31] 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.[32] However, random assignment also means that average high school graduation test scores are equal in expectation. α_2 is identified using sample variation in 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 scores. As Angrist (2013) shows, results should be interpreted with caution in this case.

I report estimates of equation 2 in table 5, using the sample of all dormitory students in the

[31] α_2 captures both "endogenous" effects of peers' GPA and "exogenous" 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.

[32] The observed dormitory assignments are consistent with randomization. I regress each baseline characteristic on a vector of dormitory indicators and test the hypothesis that the coefficients on the dormitory indicators are all equal. The bootstrap p-values for this test are 0.762 for mean high school graduation test scores, 0.857 for proportion black, 0.917 for proportion white, 0.963 for proportion other races, 0.895 for proportion English-speaking, and 0.812 for proportion international. I also run a seemingly unrelated regression model to test for equality of all six characteristics across all dormitories and fail to reject equality (bootstrap p-value 0.886).

Table 6: 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)
Student covariates		×		×
Dormitory fixed effects		×		×
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 6 reports results from estimating equation 3 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.

random assignment period. I find that $\hat{\alpha}_2 = 0.22$, which is robust to conditioning on student demographics and dormitory fixed effects. Hence, 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. These effects are large 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 peer effect mechanism 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 school graduation test, while SAT scores are a common measure in existing research. 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.[33] This may reflect the limited variation in $\overline{HS}_d$.

The linear-in-means model can be augmented to allow the effect of residential peers to vary

[33] 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.

within and across sub-dormitory groups. Specifically, I explore within- and across-race peer effects by estimating:

$$GPA_{ird} = \alpha_0 + \beta_1 HS_{ird} + \beta_2 \overline{HS}_{rd} + \beta_3 \overline{HS}_{-rd} + \vec{\beta}\vec{X}_{ird} + \vec{\mu}_d + \epsilon_{ird}. \quad (3)$$

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 6, column 2). The difference shows that peer effects operate primarily within race groups but it is quite imprecisely estimated (bootstrap p-value 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 along other dimensions such as race, which remains highly salient in South Africa.[34] This indicates that interaction 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 3 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). These results imply that within-faculty peer effects are not systematically stronger than cross-faculty peer effects.[35] 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 mechanisms through

[34] 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.

[35] Each student at the University of Cape Town is registered in one of six faculties: commerce, engineering, health sciences, humanities and social 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 2 and 3 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 cluster bootstrap standard error 0.15) but they are not driven primarily by other students in the same course ($\hat{\beta}_2 = 0.06$ and $\hat{\beta}_3 = 0.17$ with cluster bootstrap standard errors 0.17 and 0.15). This, and other course-level regressions, are consistent with the main results but the smaller sample sizes yield less precise estimates that are somewhat sensitive to the inclusion of covariates.

which peer effects operate: time allocation over study and leisure activities, and transfers of tacit knowledge such as study skills, norms about how to interact with instructors, and strategies for navigating academic bureaucracy. 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).

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 is more likely when students are also socially 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 Cross-Policy and Cross-Dormitory Results

The linear-in-means model restricts average GPA to be invariant to any group reassignment: moving 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:

$$\begin{aligned} GPA_{id} = \gamma_0 + \gamma_1 HS_{id} + \gamma_2 \overline{HS}_d + \gamma_{11} HS_{id}^2 + \gamma_{22} \overline{HS}_d^2 \\ + \gamma_{12} HS_{id} \times \overline{HS}_d + \vec{\gamma} \vec{X}_{id} + \vec{\mu}_d + \epsilon_{id} \end{aligned} \tag{4}$$

This is a parsimonious specification that permits average outcomes to vary over assignment processes and aligns with theoretical models of matching markets and neighborhood segregation.[36] γ_{12} and γ_{22} are the key parameters of the model. γ_{12} indicates whether own and peer high

[36] This specification assumes that GPA depends on peer characteristics only through the dormitory mean. The results are very similar if dormitory-year means are replaced with medians, though the estimates are slightly noisier. Some previous work finds a relationship between individual outcomes and the variance of peer characteristics (Sacerdote, 2011). My results are very similar when I control for the dormitory-year standard deviation of high school gradution test scores. The coefficient on the standard deviation itself is negative (-0.12 to -0.16) but not statistically significant. See Carrell, Sacerdote, and West (2013) for an alternative

school graduation test scores are complements or substitutes in GPA production. If $\gamma_{12} < 0$, the GPA gain from high-scoring peers is larger for low-scoring students. In classic binary matching models, this parameter governs whether positive or negative assortative matching is output-maximizing (Becker, 1973). In matching models with more than two agents, γ_{12} 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 higher when mean test scores are identical in all groups. If $\gamma_{22} > 0$, total output is higher 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, 2010). Tracking will deliver lower total GPA than random assignment if both parameters are negative and vice versa. If the parameters have different signs, the average effect of tracking is ambiguous.[37]

Estimates from equation 4 are shown in table 5 columns 4, 5 (controlling for student demographics) and 6 (with dormitory fixed effects). $\hat{\gamma}_{12}$ is negative and marginally statistically significant across all specifications. The point estimate of -0.13 (cluster bootstrap standard error 0.07) implies the GPA gain from an increase in peers' mean test scores is 0.2 standard deviations larger for students at the 25^{th} percentile of the high school test score distribution than students at the 75^{th} 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 of $\hat{\gamma}_{22}$ flips from positive to negative with the inclusion of dormitory fixed effects. It is thus unclear whether GPA is concave or convex in mean peer group test scores.

I draw three conclusions from these results. First, there is clear evidence of nonlinear peer effects from the cross-dormitory variation generated under random assignment. Likelihood ratio tests prefer the nonlinear models in columns 4-6 to the corresponding linear models in columns

parameterization and Graham (2011) for background discussion.

[37] 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, $\mathbb{E}[HS_{id}\overline{HS}_d]$ and $\mathbb{E}[\overline{HS}_d^2]$ both equal $\mathbb{E}[HS_{id}^2]$ under tracking and $\mathbb{E}[HS_{id}]^2$ under random assignment. Plugging these results into equation 4 for each assignment policy yields $\mathbb{E}[Y_{id}|\text{Tracking}] - \mathbb{E}[Y_{id}|\text{Randomization}] = \sigma_{HS}^2(\gamma_{22} + \gamma_{12})$. This simple demonstration assumes an infinite number of students and dormitories but a similar result holds with a finite population.

1-3 of table 5. Second, peer effects estimates using randomly induced cross-dormitory 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 leads to 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 negative average treatment effect of tracking.

Are the coefficient estimates from equation 4 also quantitatively consistent with the observed treatment effects of tracking? I answer this question by comparing observed GPA^{Track} to predicted GPA using the coefficient estimates reported in table 5 column 6: $\hat{GPA}^{\text{Track}} = X^{\text{Track}} \cdot \hat{\gamma}^{\text{Random}}$. The mean difference between the observed and predicted values is a consistent estimator of the average treatment effect of tracking on the tracked students (ATET) if the regression model is correctly specified (Fortin, Lemiuex, and Firpo, 2011).[38] Comparing this estimate of the ATET to the difference-in-differences estimate tests whether within-policy variation can predict the effect changing the assignment policy to tracking.

The ATET from this prediction exercise is -0.08 standard deviations (cluster boostrap standard error 0.11), compared to the difference-in-differences estimate of -0.13. The treatment effects for students with below- and above-median high school graduation test scores are -0.14 and -0.01 standard deviations respectively (standard errors 0.12 and 0.13). The prediction exercise is fairly accurate for above-median students but substantially understates the negative effect on below-median students, which is -0.24 in the difference-in-differences framework.[39] The coefficient estimates from the linear peer effects model (column 3 of table 5) come closer to matching the estimates from the difference-in-differences framework for below–median students. This highlights a tension between the within-sample model selection tests, which strongly prefer the quadratic model, and the out-of-sample predictive accuracy, which is similar for the linear

[38] The observed and predicted values can also differ if the mean values of any unobserved determinants of *standardized* GPA differ between the two periods. Any mean time trends that are common to the dormitory and non-dormitory students are already accounted for because GPA in each period is standardized with reference to the non-dormitory students. So the two estimators of the ATET should differ only due to functional form issues.

[39] The results are similar using the coefficient estimates from the quadratic peer effects model without demographic controls or dormitory fixed effects (columns 4 and 5 of table 5).

and nonlinear models.

The difference between the observed and predicted effects of tracking may reflect economic or statistical misspecification of equation 4. I call the model *statistically* misspecified if it correctly represents a policy-invariant data generating process within but not outside the support of the data observed under random assignment. I call the model *economically* misspecified if it represents a data generating process that is not policy invariant, even within the support of the data observed under random assignment. The tracking policy generates a wider support of $\bar{HS}_d$ than the random assignment policy and peer effects may operate differently in very low- and high-scoring dormitories than mid-scoring dormitories. The tracking policy may also generate systematically different patterns of peer interaction, such as more common within-dormitory friendships if students have a preference for academically homogenous social groups. The former and latter concerns correspond respectively to statistical and economic misspecification.

I explore the relative importance of economic and statistical misspecification by comparing prediction error within and outside the support of the dormitory means observed under random assignment. First, I estimate the quadratic peer effects model (equation 4) on the sample of randomly assigned dormitory students. Second, I use the coefficient estimates to predict GPA for tracked dormitory students and calculate the squared prediction error for each student. Third, I calculate the mean squared prediction error for three subgroups of tracked students: those in dormitories with mean high school graduation test scores below, within, and above the range observed under random assignment. I refer to these as low-, mid-, and high-track dormitories respectively. Fourth, I calculate squared prediction error for each randomly assigned dormitory student using leave-one-out estimation.[40] The mean squared prediction error for randomly assigned dormitory students (0.464, with standard error 0.026) provides a benchmark for the prediction error without either economic or statistical misspecification.[41] The mean

[40] I estimate the model using observations $1, \ldots, i-1, i+1, \ldots N$ to obtain $\hat{\gamma}_{-i}$ and calculate the squared prediction error for student i as $(GPA_i - X_i\hat{\gamma}_{-i})^2$. This yields an out-of-sample prediction error for randomly assigned dormitory students that is directly comparable to the prediction error for tracked dormitory students.

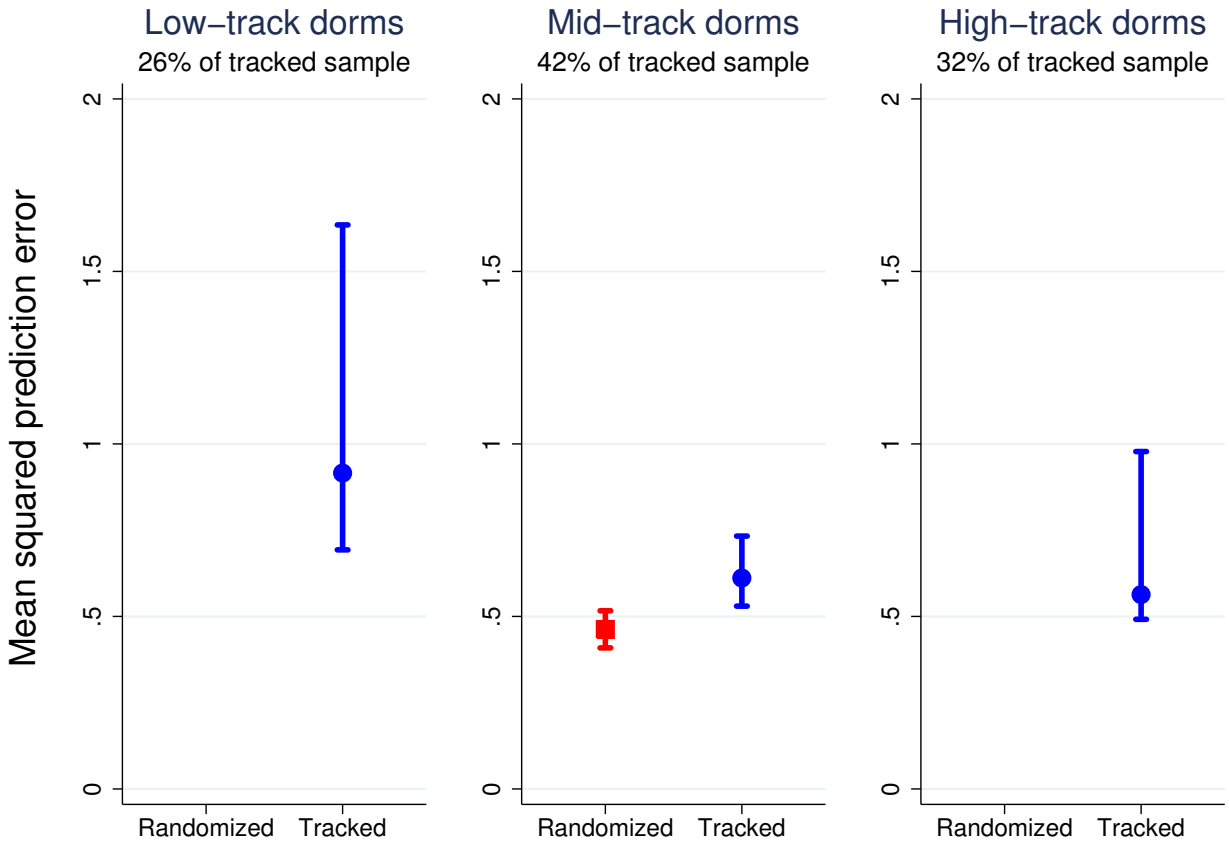
[41] Standard errors are estimated by bootstrapping the entire prediction process: estimating the quadratic peer effects model using randomly assigned students and calculating squared prediction error for each randomly assigned and tracked student. The bootstrap algorithm resamples dormitory-year clusters over 1000 replications.

squared prediction error for tracked students in mid-track dormitories (0.611 with standard error 0.054) also reflects economic misspecification. The mean squared prediction error for tracked students in low- and high-track dormitories (0.720 with standard error 0.706) also reflects both economic and statistical misspecification.

These results show that there are important roles for both economic and statistical misspecification. Mean squared prediction error is 32% higher for tracked than randomly assigned dormitory students over the original support of dormitory means, verifying that model estimated under random assignment is economically misspecified for tracking. This is consistent with the finding from section 6 that student interaction patterns can mediate spatial peer effects. Mean squared prediction error is 55% higher outside the original support for dormitory means, verifying that the model estimated under random assignment is also statistically misspecified. However, outside-support mean squared prediction error is very imprecisely estimated so the latter result should be interpreted with caution.

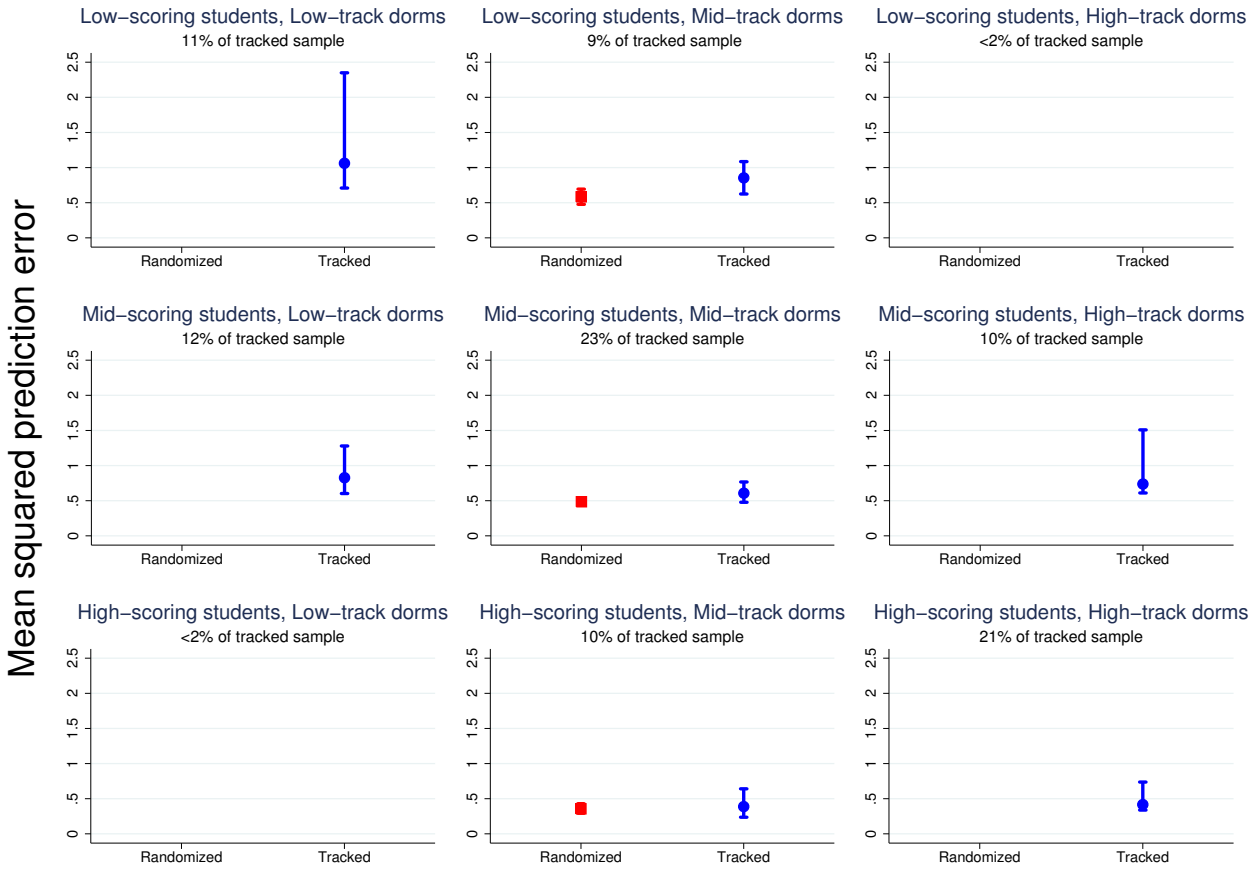
I show the mean squared prediction for randomly assigned students and tracked students in different dormitory types in figure 4. The mean squared prediction error is similar for students in mid-track dormitories, which are within the observed support under random assignment, and high-track dormitories, which are not. But the mean squared prediction error is substantially higher and less precisely estimated for students in low-track dormitories. Figure 5 shows mean squared prediction error disaggregated by dormitory track and by student type: students in the first, second/third, and fourth quartiles of the high school graduation test score distribution are defined as low-, mid-, and high-scoring respectively. This allows me to compare tracked mid-scoring students in mid-track dormitories to randomly assigned mid-scoring students. Even in this case, where the tracked and randomly assigned dormitory students have similar academic backgrounds and live in similar (on average) dormitories, mean squared prediction error is 25% higher for tracked than randomly assigned dormitory students. This difference is much more likely to reflect economic than statistical misspecification. These disaggregated results reinforce three findings from the aggregate analysis: the model estimated using random variation in

Figure 4: Mean Squared Prediction Error by Dormitory Track, for Tracked and Randomly Assigned Dormitory Students



Notes: Blue circles show mean squared prediction error for tracked students. Red squares show mean squared prediction error for randomly assigned students. Squared prediction error is the difference between observed GPA and predicted GPA using the coefficient estimates from estimating the quadratic peer effects model (equation 4) on the sample of randomly assigned dormitory students. The prediction error for randomly assigned dormitory students is derived from leave-one-out models so all errors are out-of-sample predictions. The 95% confidence intervals are from a percentile bootstrap algorithm that resamples dormitory-year observations, estimates the quadratic peer effects model, and estimates individual prediction errors 1000 times. Low-, mid-, and high-track dormitories are defined as those with mean high school graduation test scores below, within, and above the observed range under random assignment.

Figure 5: Mean Squared Prediction Error by Dormitory Track and Student Type, for Tracked and Randomly Assigned Dormitory Students



Notes: Blue circles show mean squared prediction error for tracked students. Red squares show mean squared prediction error for randomly assigned students. Squared prediction error is the difference between observed GPA and predicted GPA using the coefficient estimates from estimating the quadratic peer effects model (equation 4) on the sample of randomly assigned dormitory students. The prediction error for randomly assigned dormitory students is derived from leave-one-out models so all errors are out-of-sample predictions. The 95% confidence intervals are from a percentile bootstrap algorithm that resamples dormitory-year observations, estimates the quadratic peer effects model, and estimates individual prediction errors 1000 times. Low-, mid-, and high-track dormitories are defined as those with mean high school graduation test scores below, within, and above the observed range under random assignment. Low, mid-, and high-scoring students are defined as those in the first, second/third, and fourth quartiles of the distribution of high school graduation test scores.

dormitory composition is both economically and statistically misspecified, and prediction is least accurate for low-scoring students in low-scoring dormitories.[42] I conclude that using peer effects estimated under random assignment to predict outcomes under tracking is both economically and statistically problematic. This reinforces the negative effects of tracking on low-scoring students in this setting. Tracking reduces their GPA even more than predicted, which may complicate the design of interventions to offset these negative effects.

8 Alternative Explanations for the Effects of Tracking

I consider five alternative explanations that might have generated the observed GPA difference between tracked and randomly assigned dormitory students. The first three explanations are violations of the parallel time changes assumption: time-varying student selection regarding whether or not to live in a dormitory, differential time trends in dormitory and non-dormitory students' characteristics, and spillover effects of tracking on non-dormitory students. The fourth explanation is that the treatment effects are an artefact of the grading system and do not reflect any real effect on learning. The fifth 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 publicized,

[42] The last result could in principle be explained by a left-skewed GPA scale. This scale would inflate prediction errors in the left tail relative to the right tail. I explore this concern by implementing the entire analysis using rank in the GPA distribution as the outcome, rather than GPA level. Results are similar with this alternative outcome measure, which reduces concerns about the role of the outcome scale.

limiting students' ability to respond. Second, table 2 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, admission rules cap the number of students from 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 address, which I do not observe. Nonetheless, the instrument strongly predicts dormitory status: 76% of non-Cape Town students and 8% of Cape Town students live in dormitories. The intention-to-treat and instrumented treatment effects (table 7, columns 2 and 3) are very similar to the treatment effects without instruments (table 3).

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 three 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 (random assignment). I do not observe dormitory assignments in 2001–2002 so I report only intention-to-treat effects.[43] The raw data are shown in the first panel of figure 6. 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 -0.09 with standard error 0.04 (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

[43] The cluster bootstrap standard errors do not take into account potential clustering within (unobserved) dormitories in 2001–2002 and so are 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.

Table 7: Robustness Checks

Outcome	**1**{Dorm student}	GPA	GPA	GPA	GPA	GPA	No. of credits	GPA	GPA	% credits excluded	GPA | non-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-treatment						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 time 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
# dorm-year clusters	60	60	60	60	60	60	60	52	60	60	60
# dormitory students	6915	6915	6915	6915	6915	6915	7480	6795	7480	7480	7449
# non-dorm students	6466	6466	6466	6466	6466	6466	7188	7188	7188	7188	7043
# students with missing dorm status				9331	9331	9331					

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 first stage outcome. Columns 4–6 use data from 2001-2002, 2004-2005, and 2007-2008 to test the parallel time trends assumption. Column 4 reports a difference-in-differences estimate comparing all four observed years of tracking to the two observed years of random assignment. Column 5 reports the difference between observed GPA under random assignment and predicted GPA from a linear time trend extrapolated from the tracking period. Column 6 reports 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 main and placebo effects following Heckman and Hotz (1989). Column 7 reports a difference-in-differences estimate with the credit-weighted number of courses as the outcome. Column 8 reports a difference-in-differences estimate excluding dormitories that are either observed in only one period or use a different admission rule. Column 9 reports a difference-in-differences estimate including college/faculty/school fixed effects. Column 10 reports a difference-in-differences estimate with the credit-weighted percentage of courses from which students are academically excluded as the outcome. Column 11 reports a difference-in-differences estimate with GPA calculated using only grades from non-excluded courses as the outcome. Standard errors in parentheses are from 1000 bootstrap replications, stratifying by assignment policy and dormitory status. The bootstrap resamples dormitory-year clusters except for the 2001-2002 data in columns 4-6, for which dormitory assignments are not observed.

students' GPAs in 2001-2002 and 2004-2005 yields a small positive but insignificant effect of 0.06 (standard error 0.05). I subtract the placebo test result from the original treatment effect estimate to obtain a "trend-adjusted" treatment effect of -0.18 with standard error 0.10 (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 estimate 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 0.14 with standard error 0.09 (table 7, column 5). This estimate is appropriate if the two groups of 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 (shown in the second panel of figure 6). 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 at the university. The university charged a flat tuition fee up to 2005 and per-credit fees from 2006. This may have changed the number of courses for which students registered. However,

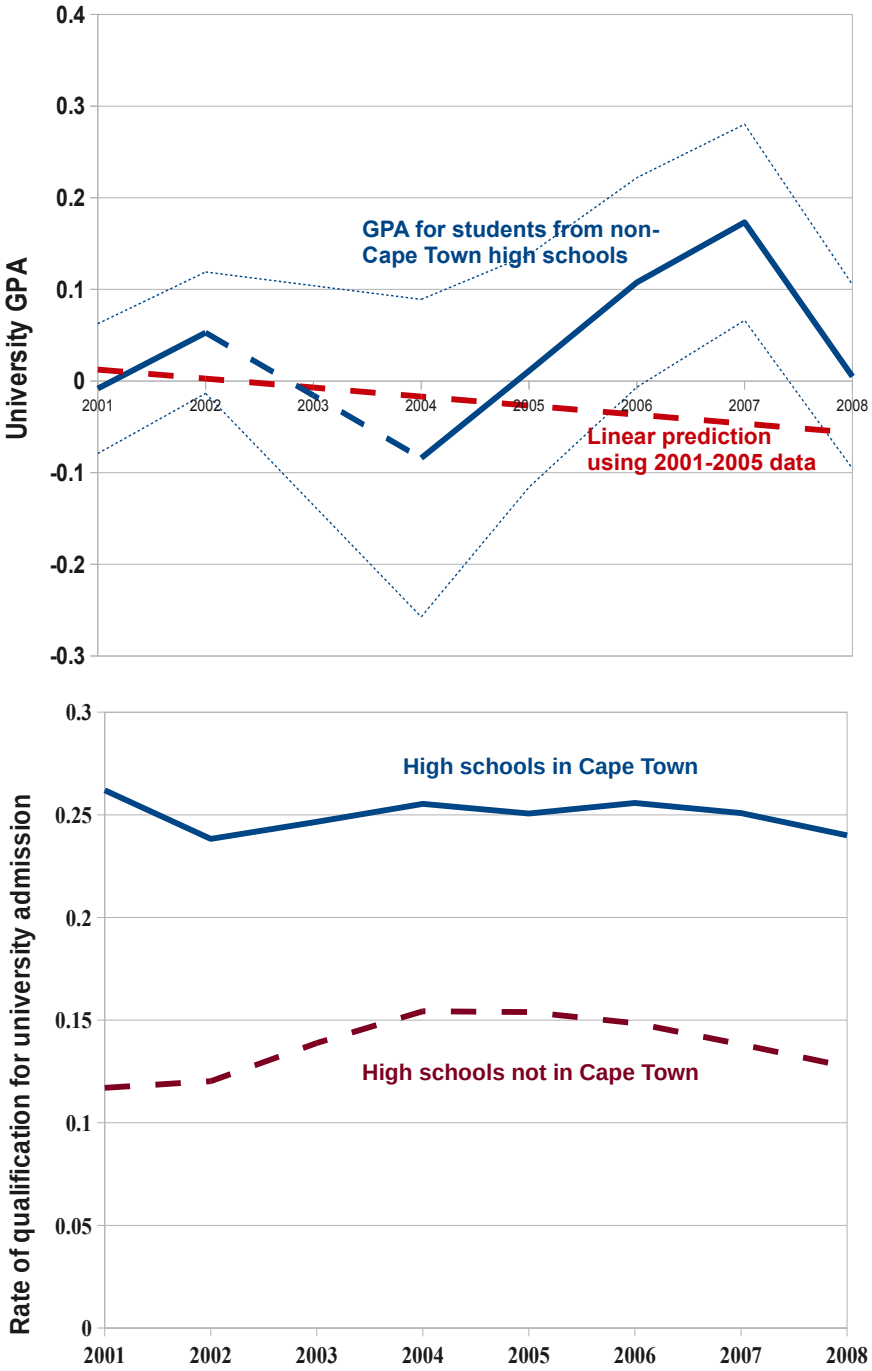
the credit-weighted number of courses remained constant for dormitory and non-dormitory students, with a difference-in-differences estimate of 0.013, less than 0.4% of the mean (table 7 column 7). The university also closed one dormitory in 2006 and opened a new dormitory in 2007, as well as reserving 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 Spillover effects of tracking on non-dormitory students

Could all or part of the estimated treatment effects of tracking have been generated by spillover effects on non-domitory students? Two pieces of evidence suggest spillovers are possible. First, the within-race peer effects documented in section 6 suggest that peer effects are mediated by patterns of social interaction, and interactions may occur between dormitory and non-dormitory students. Second, raw GPAs are slightly higher for non-dormitory students in the tracking than the random assignment period. If spillovers occur, and raise the raw GPAs for non-dormitory students in the tracking period, then the treatment effects of tracking estimated using difference-in-differences will be overstated. I present two arguments against this explanation.

First, I develop a specific framework of spillovers that generates a positive effect of tracking on non-dormitory students' GPAs. I show that this framework generates additional testable predictions that are not consistent with the data. Assume that students have a preference for academically homogeneous social groups. Then tracked dormitory students will interact mainly with their dormmates and randomly assigned dormitory students interact more often with non-dormitory students. High-scoring non-dormitory students will thus interact with fewer high-scoring dormitory students under tracking and low-scoring non-dormitory students will interact with fewer low-scoring dormitory students under tracking. If the parameter estimates from equation 4 also apply to non-residential peer groups, then (a) mean GPA for non-dormitory students will be higher under tracking and (b) high-scoring and low-scoring non-dormitory students will have respectively lower and higher GPAs under tracking than ran-

Figure 6: 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 dotted lines show a 95% confidence interval constructed from 1000 replications of a percentile bootstrap stratifying by assignment policy and dormitory status. The bootstrap resamples dormitory-year clusters for 2004-2008, the only years in which dormitory assignments are observed. 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).

dom assignment.[44] Prediction (b) can be tested and I find that non-dormitory students with above-median high school graduation test scores have GPAs that are 0.044 standard deviations higher (s.e. 0.028) under tracking than random assignment. Non-dormitory students with below-median high school graduation test scores have GPAs that are 0.003 standard deviations (s.e. 0.034) higher under tracking than random assignment. The first difference has the wrong sign and both are very close to zero. This argument does not rule out the existence of some other social interactions framework that biases the treatment effect. I can simply show that one particularly salient framework produces predictions that are not consistent with the data.

Second, the change in non-dormitory students' mean GPAs is consistent with the results of "benchmarking" (i.e. aptitude) tests that 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 conclude that spillovers from dormitory to non-dormitory students are unlikely to generate the observed pattern of treatment effects, though I cannot directly test spillover mechanisms without data on social networks or time use.

8.4 Limitations of GPA as an Outcome Measure

I explore four ways in which the grading system might pose a problem for validity or interpretation of the results: curving, ceiling effects, course choices, and 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 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 stan-

[44] Prediction (a) follows because $\hat{\gamma}_{12} < 0$ in equation 4, so the negative effect of tracking on high-scoring non-dormitory students will be smaller than the positive effect of tracking on low-scoring non-dormitory students. Prediction (b) follows because $\hat{\gamma}_2 > 0$.

dard. 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 75^{th} and 25^{th} 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 for accounting students, although it is only administered only after they graduate, reduces the scope for internal assessment to change through time. Tracking reduces mean grades in the introductory accounting course by 0.11 standard deviations (cluster bootstrap standard error 0.12, sample size 2107 students). This provides some reassurance that tracking reduces real academic performance.

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 99^{th} 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 0.13 standard deviations (cluster bootstrap standard error 0.06), 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, with different grading standards, in the tracking and random assignment periods. There are some changes in the type of courses students take: dormitory students take slightly fewer commerce and science classes and slightly

more engineering and social science classes in the tracking than random assignment period, relative to non-dormitory students. Courses are also marginally more concentrated by dormitory in the tracking period. The average student lives in a dormitory where 27.5% of her peers are in the same faculty/school/college under random assignment. This is 0.9 percentage points higher under tracking (standard error 0.3). 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 -0.17 with standard error 0.04 (table 7, column 9). I conclude that the main results are not driven by time-varying course-taking behavior.

Fourth, the university employs a 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 3.7 to 6.4% and reduces GPA in non-excluded courses by 0.08 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 reduces grades at both the intensive and extensive 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.

8.5 Other Mechanisms 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 feature 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 interactions 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 2 with an interaction between HS_{id} and the rank of dormitory d during the tracking period. The interaction term has a small and insignificant coefficient: 0.003, cluster bootstrap standard error 0.006. Hence, low-scoring students do not have systematically lower GPAs when 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 or discrimination by instructors. Stereotype threat would occur if students' dormitory assignment informed or continuously reminded them of their high school graduation test score and undermined low-scoring students' confidence or motivation (Steele and Aronson, 1995). I cannot directly test this hypothesis and so cannot rule it out. However, dormitory assignment probably provided students with

limited information about their academic rank because course-specific admissions thresholds are publicly available. The consistent results from the cross-policy and cross-dormitory analyses also suggest that peer effects explain much of the observed treatment effect of tracking. Discriminatory grading would occur if instructors observed students' dormitory assignments and assigned lower scores to students in low-track, conditional on the quality of their work. I test this hypothesis by estimating the treatment effect of tracking in the largest first-year course at the university, Introductory Microeconomics. Most assessment in this course uses electronically-graded multiple choice tests, leaving no scope for instructor discrimination. The effect of tracking in this course is -0.13, with cluster bootstrap standard error 0.06: large, negative and almost identical to the average effect across all courses. I conclude that the headline results cannot be entirely explained by discriminatory grading.

9 Conclusion

This paper describes the effect of tracking relative to random dormitory assignment on student GPAs at the University of Cape Town in South Africa. I show that tracking lowered mean GPA and increased GPA inequality. This result occurs because living with high-scoring peers has a larger positive effect on low-scoring students' GPAs than on high-scoring students' GPAs. 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 show that different peer group assignment policies can have substantial effects on students' academic outcomes. Academic tracking into residential groups, and perhaps other noninstructional groups, may generate a substantially worse distribution of academic performance than random assignment. However, my results do not permit a comprehensive evaluation of 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 may thus entail a transfer from one group of students to another and, as academic outputs are not directly tradeable, Pareto-ranking the two policies may not be possible. Non-measured student outcomes may also be affected by different group assignment policies. For example, high-scoring students' GPAs 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 on graduation rates, time-to-degree, and labor market outcomes. This will permit a more comprehensive evaluation of the two group assignment policies. One simple revealed preference measure of student welfare under the two policies is the proportion of dormitory students who stay in their dormitory for a second year. Tracking reduces this rate for students with above- and below-median high school test scores by 0.4 percentage points and 6.7 percentage points respectively (cluster bootstrap standard errors 3.5 and 3.8 respectively). Low-scoring dormitory students may thus be aware of the negative effect of tracking and respond by leaving the dormitory system early.

Despite these provisos, my findings shed light on the importance of peer group assignment policies. I provide what appears to be the first cleanly 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) suggest that a positive total effect of instructional tracking may combine a negative peer effect of tracking on low-scoring students with a positive effect due to changes in instructor behavior. My findings suggest that policymakers can change the distribution of students' academic performance by rearranging the groups in which these students interact without changing the marginal distribution of inputs into the education production function. 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 (Bertrand, Hanna, and Mullainathan, 2010).

The examination of peer effects under random assignment also points to fruitful avenues for future research. As in Carrell, Sacerdote, and West (2013), peer effects estimated under random assignment do not robustly predict the effects of a new assignment policy and residential peer effects appear to be mediated by students' patterns of interaction. This highlights the risk of relying on reduced form estimates that do not capture the behavioral content of peer effects. Combining peer effects estimated under different 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 Economics and Statistics*, 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.

Ajayi, K. (2014): "Does School Quality Improve Student Performance? New Evidence from Ghana," Working paper, Boston University.

Angrist, J. (2013): "The Perils of Peer Effects ," *Labour Economics*, 30(C), 98–108.

Angrist, J., and K. Lang (2004): "Does School Integration Generate Peer Effects? Evidence from Boston's Metco Program," *American Economic Review*, 94(5), 1613–1634.

Arnott, R. (1987): "Peer Group Effects and Educational Attainment," *Journal of Public Economics*, 32, 287–305.

Athey, S., and G. Imbens (2006): "Identification and Inference in Nonlinear Difference-in-differences Models," *Econometrica*, 74(2), 431–497.

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 Experi-

ment," Mimeo.

Booij, A., A. Leuven, and H. Oosterbeek (2014): "Ability Peer Effects in University: Evidence from a Randomized Experiment," Mimeo.

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. (2014): "Can Achievement Peer Effect Estimates Inform Policy? A View from Inside the Black Box," *Review of Economics and Statistics*, 96(3), 514–523.

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.

Fortin, N., T. Lemiuex, and S. Firpo (2011): "Decomposition Methods in Economics,"

in *Handbook of Labor Economics Volume 4A*, ed. by O. Ashenfelter, and D. Card. North-Holland.

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(8/9), 1455–1475.

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

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 (2010): "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.

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 Economics and Statistics*, 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.

Hoel, J., J. Parker, and J. Rivenburg (2004): "Peer Effects: Do First-Year Classmates, Roommates, and Dormmates Affect Students Academic Success," Working paper, Reed College.

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," Working paper, Harvard University.

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.

Jain, T., and M. Kapoor (2015): "The Impact of Study Groups and Roommates on Academic Performance," *Review of Economics and Statistics*, 1(97), 44–54.

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.
Lucas, A., and I. Mbiti (2014): "Effects of School Quality on Student Achievement: Discontinuity Evidence from Kenya," *American Economic Journal: Applied Economics*, 3(6), 234–263.
Ludwig, J., J. Kling, and S. Mullainathan (2011): "Mechanism Experiments and Policy Evaluations," *Journal of Economic Perspectives*, 3(25), 17–38.
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.
Pop-Eleches, C., and M. Urquiola (2013): "Going to a Better School: Effects and Behavioral Responses," *American Economic Review*, 103(4), 1289–1324.
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.
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.

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

Athey and Imbens (2006) establish a model for recovering 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 recovers only the average treatment effect on the treated. However, the model requires stronger identifying assumptions.

The original model is identified under five 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 assumptions are:

A1 GPA in the absence of tracking is generated by the unknown production function $GPA = h(U, T)$, where U is an unobserved scalar random variable. GPA does not depend directly on D.

A2 The production function $h(u, t)$ is strictly increasing in u for $t \in \{0, 1\}$.

A3 The distribution of U is constant through time for each group, in this case dormitory and non-dormitory students: $U \perp T|D$.

A4 The support of dormitory students' GPA is contained in that of non-dormitory students' GPA: $supp(GPA|D = 1) \subseteq supp(GPA|D = 0)$.[45]

A5 The distribution of GPA is strictly continuous.[46]

These assumptions are sufficient to identify the counterfactual distribution of tracked dormitory students' GPAs in the absence of tracking, $F^{CF}_{GPA|D=1,T=1}(\cdot) = F_{GPA^{10}}\left(F^{-1}_{GPA^{00}}\left(F_{GPA^{01}}(\cdot)\right)\right)$. These are the outcomes that tracked students would have obtained if they had been randomly assigned. The q^{th} quantile treatment effect 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 covariate vector X but not unconditionally. In my application, some of the demographic characteristics show time trends (table 2). If these characteristics are subsumed in U and in turn influence GPA, then the stationarity assumption *A3* will fail. The assumption may, however, hold after conditioning on X.

[45] This assumption is testable and holds in my data.

[46] This assumption is testable and holds approximately in my data. There are 5505 unique GPA values for 14668 observations. No value accounts for more than 0.3% of the observations.

Athey and Imbens discuss two ways to include observed covariates in the model. First, a fully nonparametric method that applies the model separately to each value of the covariates. This is feasible only if the dimension of X is low. Second, a parametric method that applies the model to the residuals from a regression of GPA on X. This is valid only under the strong assumption that the observed covariates X and unobserved scalar U are independent (conditional on D) and additively separable in the GPA production function. Substantively, the additively separable model is misspecified if the treatment effect of tracking at any quantile varies with X. For example, different treatment effects on students with high and low high school graduation test scores would violate this restriction.

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 at each value g as

$$F^{RW,CF}_{GPA^{11}}(g) = F_{GPA^{10}_{\omega}}\left(F^{-1}_{GPA^{00}_{\omega}}\left(F_{GPA^{01}}\left(g\right)\right)\right) \tag{5}$$

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.[47] This is a straightforward adaptation of the reweighting techniques used in wage decompositions and program evaluation (DiNardo, Fortin, and Lemiuex, 1996; Hirano, Imbens, and Ridder, 2003). The counterfactual distribution is identified under conditional analogues of assumptions $A1 - A5$.[48]

[47] I could instead use $F^{RW,CF}_{GPA^{11}}(g) = F_{GPA^{10}_{\omega}}\left(F^{-1}_{GPA^{00}_{\omega}}\left(F_{GPA^{01}_{\omega}}\left(g\right)\right)\right)$ as the counterfactual distribution, with weights $Pr(T = 1, D = 1|X)/Pr(T = t, D = d|X)$ for $(d, t) \in \{0, 1\}^2$. This reweights all four groups of students to have the same distribution of observed characteristics. Balancing the distributions may increase the plausibility of the assumption that dormitory and non-dormitory students share the same production function $h(x, u, t)$. Results from this model are similar, but with larger negative effects in the left tail.

[48] Note that the conditional version of assumption $A4$ is more restrictive than the unconditional version. The common support assumption may hold for the marginal distribution $F_{GPA}(\cdot)$ but not for the conditional distribution $F_{GPA|X}(\cdot|x)$ for some value of x.

Hence, the q^{th} quantile treatment effect on the tracked students is

$$\tau^{QTT}(q) = F^{-1}_{GPA^{11}}(q) - F^{-1,RW,CF}_{GPA^{11}}(q). \tag{6}$$

I do not formally establish conditions for consistent estimation. Firpo (2007) recommends a series logit model for the propensity score $Pr(T = 1|D = d, X)$ with the polynomial order chosen using cross-validation. I report results with a quadratic function; the treatment effects are similar with linear, cubic, and quartic specifications. I implement the estimator in four steps:

1. For $D \in \{0, 1\}$, I regress T on student gender, language, nationality, race, a quadratic in high school graduation test scores and all pairwise interactions. I construct the predicted probability $\hat{Pr}(T = 1|D, X)$.

2. I evaluate equation 5 at each half-percentile of the GPA distribution (i.e. quantiles 0.5 to 99.5). I plot this counterfactual GPA distribution for tracked students in the first panel of figure 3, along with the observed GPA distribution.

3. I plot the difference between observed and counterfactual distributions at each half-percentile in the second panel of figure 3.

4. I construct a 95% bootstrap confidence interval at each half-percentile, clustering at the dormitory-year level and stratifying by (D, T).

I also use the estimated counterfactual distribution to construct summary statistics such as the mean and variance of counterfactual GPA. I approximate the mean by Riemann integrating the area of the left of the counterfactual distribution:
$\mathbb{E}\left[GPA^{RW,CF}\right] \approx \frac{1}{198} \cdot \sum_{p=2}^{199} \left[\frac{1}{2} F^{RW,CF}_{GPA^{11}}(p) + \frac{1}{2} F^{RW,CF}_{GPA^{11}}(p-1)\right]$. The second uncentered moment of the counterfactual distribution can be constructed in the same way using the square of the counterfactual distribution function. I then construct the variance using $\mathbb{E}\left[\left(GPA^{RW,CF}\right)^2\right] - \left(\mathbb{E}\left[GPA^{RW,CF}\right]\right)^2$. These statistics are measured with error due to the linear approximation

used in the Riemann integration. The measurement error decreases as the number of evaluation points increases. The measurement error is zero if all students obtain the same counterfactual GPA, in which case the distribution function is linear.

Stata code for implementing this estimator is available at www.robgarlick.com/code.