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

Robert Garlick

September 9, 2017

A   Balance Tests over Dormitory Status and Period

In table 1 I report mean values of selected student characteristics that are determined prior to university attendance, separately by tracking/random assignment period and by dormitory/non-dormitory status. I report p-values from testing if these means are equal across dormitory students in the tracking and random assignment periods (column 4) and across non-dormitory students in the tracking and random assignment periods (column 7). I also report p-values in column 8 from testing if the changes in means between the tracking and random assignment periods are equal for dormitory and non-dormitory students, i.e. testing if the parallel trends assumption holds. There are differences in three of thirteen characteristics that are significant at the 10% level. However, two of these differences are due to omitting one non-randomly assigned dormitory in the random assignment period, as I discuss in section I.

B   Evaluating Alternative Explanations for the Treatment Effects of Tracking

I consider five alternative explanations that might have generated the mean GPA difference between tracked and randomly assigned dormitory students.

Table 1: Covariate Means by Dormitory Status and Period with Balance Tests

	(1) Entire sample	(2) Tracked dorm	(3) Randomized dorm	(4) p-value (2)=(3)	(5) Tracked non-dorm	(6) Randomized non-dorm	(7) p-value (5)=(6)	(8) p-value (2)-(3)=(5)-(6)
High school GPA	0.090	0.169	0.209	0.136	-0.000	0.000	1.000	0.276
A grade in high school	0.278	0.320	0.327	0.531	0.222	0.253	0.003	0.137
$\leq$ C grade in high school	0.232	0.224	0.196	0.009	0.254	0.250	0.722	0.113
Female	0.515	0.499	0.521	0.056	0.523	0.514	0.451	0.061
Black	0.317	0.503	0.518	0.192	0.116	0.118	0.751	0.359
White	0.425	0.354	0.337	0.127	0.520	0.495	0.033	0.611
Other race	0.258	0.143	0.145	0.831	0.364	0.387	0.047	0.135
English-speaking	0.715	0.593	0.563	0.010	0.851	0.863	0.125	0.003
International	0.142	0.225	0.171	0.000	0.106	0.061	0.000	0.389
Graduated high school in tracking period	0.517	1.000	0.018		1.000	0.036		0.000
Graduated from high school in Cape Town	0.413	0.088	0.083	0.419	0.765	0.754	0.289	0.645
Cape Town high school | A grade	0.414	0.101	0.065	0.005	0.848	0.811	0.058	0.974
Cape Town high school | $\leq$ C grade	0.527	0.146	0.183	0.099	0.798	0.800	0.927	0.254

Notes: Table 1 reports summary statistics of student covariates 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 covariate 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.

The first three explanations are violations of the parallel trends assumption: time-varying student selection regarding whether or not to live in a dormitory, differential time trends in dormitory and non-dormitory students' covariates, and spillover effects of tracking on non-dormitory students. If any of these first three explanations are true, then the strategy laid out in section I will not identify the average treatment effect of tracking on the tracked students and the estimates in section II are not "correct." The fourth explanation is that the treatment effects are an artefact of the grading system and do not reflect any real effect on learning. If this fourth explanation is true, then the results reported in section II are not incorrect but should be interpreted as effects on grades, not learning. The fifth explanation is that dormitory assignment affects GPA through a mechanism other than peer effects. If the fifth explanation is true, then the results reported in section II are not incorrect but should not be interpreted as arising entirely from peer effects.

B.1 Time-varying 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 to live in a dormitory based on the assignment policy. I argue that this selection is unlikely and that my results are robust to accounting for it. First, the change in dormitory assignment policy was not officially announced or widely publicized, limiting students' ability to respond. Second, the results in tables 2 and 1 show that there are similar time changes in dormitory and non-dormitory students' demographic covariates and HSGPA. Third, the results are robust to accounting for differences in covariates 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 2, columns 2 and 3) are very similar to the treatment effects without instruments (table 3).

B.2   Differential Time Trends in Student Covariates

The research design assumes that dormitory and non-dormitory students' GPAs would have the same time trends if the assignment policy had not changed. I present three arguments to support this assumption. 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.[1] The raw data are shown in the first panel of figure 2. 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 2, column 4). This estimate is appropriate if one group of students experiences a transitory shock in 2004/2005, whose influence will be reduced by including 2001-2002 in the sample. A placebo test comparing the difference between Cape Town and non-Cape Town students' GPAs in 2001-2002 and 2004-2005 yields a small positive but insignificant effect of 0.06 (standard error 0.05), showing that the change in GPAs coincides with the change in assignment policy. 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 2, column 5). This estimate is appropriate if the two groups of students have lin-

---

[1]The cluster bootstrap standard errors do not take into account potential clustering within (unobserved) dormitories in 2001–2002 and so are biased downward. 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 2: Treatment Effects with Selection into Dormitory Status, Time Trends, and Other Policy Changes

Outcome	Dorm student (1)	GPA (2)	GPA (3)	GPA (4)	GPA (5)	GPA (6)	# of credits (7)	GPA (8)
Cape Town high school	0.601 (0.015)							
Cape Town high school × tracking period		-0.093 (0.036)				-0.143 (0.090)		-0.135 (0.040)
Dormitory × tracking period			-0.134 (0.053)	-0.092 (0.042)	-0.119 (0.056)		-0.027 (0.037)	
Placebo pre-treatment diff-in-diff					0.058 (0.053)			
Trend-adjusted treatment effect					-0.178 (0.101)			
Student covariates	×	×	×				×	×
Missing data indicators	×	×	×				×	×
Dormitory fixed effects			×				×	×
Instruments								
Linear time trend						×		
Adjusted R²	0.525	0.231	0.232	0.001	0.002	-	0.129	0.228
# dormitory-year clusters	58	58	58	58	58	58	58	54
# dormitory students	6858	6858	6858	6858	6858	6858	7410	6932
# non-dormitory students	6466	6466	6466	6466	6466	6466	7188	7188
# students with missing dormitory status				6161	6161	6161	7188	7188

Notes: Table 2 reports results from the robustness checks discussed in appendices B.1 and B.2. 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 3 shows the IV estimate. Dormitory fixed effects are excluded in columns 1–3 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 "long" difference-in-differences estimate comparing all four observed years of tracking to the two observed years of random assignment. Column 5 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 6 reports the difference between observed GPA under random assignment and predicted GPA from a linear time trend extrapolated from the tracking period. Dormitory fixed effects and student covariates are excluded in columns 4–6 because they are not observed for years before 2004. 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. Standard errors in parentheses are from 1000 bootstrap iterations, 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. The sample size is lower in column 8 than 7 because some dormitories are omitted. The sample size is lower in columns 1–3 than column 7 because high school location is not observed for 9% of students.

ear 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 and 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 2, column 6). 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: figure 2 panel 1 shows that intention-to-treat students (those from high schools outside Cape Town) strongly outperform control students in 2006 and 2007 but not 2008. The year-on-year changes in the GPA difference between students from high schools inside and outside Cape Town *within each policy period* are not trivial: -0.061, 0.095, 0.068, and 0.17 (mean absolute value 0.099). These are generally smaller than the treatment effect of tracking (-0.134) but not statistically significantly smaller. Although the within-policy annual changes in the GPA gap are substantial relative to the average treatment effect, they are small relative to the heterogeneous treatment effect of tracking by high school GPA. To illustrate this point, I estimate a second difference in mean GPA: between students from high schools in and outside Cape Town, for students with above- and below-median HSGPA. The annual changes in this measure are never more than half the size of the difference in treatment effects of tracking for students with above- and below-median HSGPA. Even if there are shocks that differentially affect mean GPA for students by dormitory status, these do not account for the wider GPA observed under tracking.

Second, the time trends in the proportion of graduating high school students who qualify for admission to university are very similar for Cape Town and non-Cape Town high schools between 2001 and 2008 (figure 2 panel 2). 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. The set of students who qualify for university admission is only a 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 0.03 courses, approximately 0.7% of the mean (table 2 column 7). The university also closed one dormitory in 2006 and opened a new dormitory in 2007. The estimated treatment effect is robust to excluding all three dormitories (table 2 column 8).

B.3 Sensitivity to Time-varying Selection and Differential Time Trends

I also conduct a formal sensitivity analysis to understand "how much" time-varying selection on unobserved covariates would be required to fully account for the GPA difference between tracked and untracked dormitory students. I use the relationship between GPA, tracking, and the observed covariates to calibrate the influence of unobserved covariates. This has the same spirit as sensitivity analyses proposed by Altonji, Elder and Taber (2005), Oster (2016), and Rosenbaum and Rubin (1983) but the structure of my analysis is quite different.[2] I assume that the correct GPA model is not equation (1) but is

---

[2]Those analyses use the difference between the estimated treatment effects with and without conditioning on observed covariates to calibrate the potential role of unobserved covariates. I instead use the relationship between the outcome and individual observed covariates to calibrate the possible relationship between the outcome and an unobserved covariate. The Altonji, Elder and Taber (2005) sensitivity analysis implies that my results are

instead

$$\boldsymbol{GPA} = \delta_0 + \boldsymbol{D} \cdot \delta_1 + \boldsymbol{T} \cdot \delta_2 + \boldsymbol{DT} \cdot \delta_3 + \boldsymbol{Z} \cdot \phi + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{D}$ and $\boldsymbol{T}$ are vectors whose elements are equal to one for respectively dormitory students and students in the tracking period and $\boldsymbol{DT}$ is the Hadamard product of $\boldsymbol{D}$ and $\boldsymbol{T}$. $\boldsymbol{Z}$ is an unobserved vector I interpret as "academic orientation," so $\phi > 0$. Omitting $\boldsymbol{Z}$ from the estimating equation will make the OLS estimator of δ_3 inconsistent. To reduce the dimension of the problem, I assume that $Z_i \in \{-1, 1\}$, $\mathbb{E}[Z_i] = \mathbb{E}[Z_i|D_i = 0, T_i = 1] = \mathbb{E}[Z_i|D_i = 0, T_i = 0] = 0$, and $\mathbb{E}[Z_i|D_i = 1, T_i = 1] = \rho = -\mathbb{E}[Z_i|D_i = 1, T_i = 0]$. These assumptions are valid if the mean of $\boldsymbol{Z}$ in the population of South African high school graduates is constant through time, the set of graduates who apply to live in the domitories at this university changes through time, and the set of graduate who apply to attend the university but not live in dormitories does not change through time. These assumptions are not necessary for the sensitivity analysis but they simplify the algebra.

Under these assumptions, the OLS estimator of δ_3 converges in probability to $\delta_3 + \frac{\rho \cdot \phi}{1 - \mu_{D=1,T=1}/\mu_{D=1}}$, where $\mu_{D=1,T=1}$ and $\mu_{D=1}$ are the population proportions of respectively tracked dormitory and all dormitory students. So the direction of the bias is determined by the sign of $\rho \cdot \phi$:

- Students with high academic orientation ($Z_i = 1$) may prefer to live in dormitories under tracking than under random assignment because this exposes them to similar peers who are also focused on academic performance. In this case, $\rho > 0$ and the OLS estimator is upward-biased.

- Students with low academic orientation ($Z_i = -1$) may prefer to live in dormitories under tracking than under random assignment because

---

very robust to selection on unobserved covariates, because the conditional and unconditional treatment effects estimates in table 3 are very similar to each other.

this exposes them to similar peers who are also more interested in social and leisure activities. In this case, $\rho < 0$ and the OLS estimator is downward-biased.

I can use this formula to construct the "bias-corrected" estimator

$$\hat{\delta}_3^{BC} = \hat{\delta}_3 - \frac{\rho \cdot \phi}{1 - \hat{\mu}_{D=1,T=1}/\hat{\mu}_{D=1}} \tag{2}$$

for any (ρ, ϕ).[3]

I plot $\hat{\delta}_3^{BC}$ against ρ for selected values of ϕ in the first panel of figure 1. I then use the values of (ρ, ϕ) for observed covariates to assess how much selection on $\boldsymbol{Z}$ would be required to account for the apparent treatment effect of tracking. The largest value of $|\rho|$ for any binary covariate in table 2 is 0.091 (for language), shown by the vertical black lines in the figure. If the unobserved covariate $\boldsymbol{Z}$ differs between the tracking and random assignment periods as much as the "most different" observed covariate (language) and is as strongly associated with GPA as race, the bias-corrected treatment effect is approximately -0.10 standard deviations (if academically orientated students select out of tracking) or -0.17 standard deviations (if academically orientated students select into tracking). The bias-adjusted treatment effect of tracking is zero only if unobserved academic orientation is twice as selected as any observed covariate ($\rho \approx 0.2$) and academic orientation predicts students' GPAs as strongly as obtaining mostly As versus mostly Bs on the high school graduation examination ($\phi \approx 0.33$).

The same sensitivity analysis works if the correct GPA model includes both observed covariates $\boldsymbol{X}$ and an unobserved covariate $\boldsymbol{Z}$. I concentrate on the "worst-case" scenario where $\boldsymbol{Z}$ is uncorrelated with each element of $\boldsymbol{X}$ within each group defined by $(\boldsymbol{D}, \boldsymbol{T})$; if the observed and unobserved covariates are

---

[3]This estimator is consistent but still biased because the ratio of the sample mean and sample variance is not an unbiased estimator of the ratio of the population mean and population variance. This bias can be reduced using a higher order Taylor series approximation. This correction is quantitatively unimportant in this application so I omit it for expositional clarity.

Figure 1: "Bias-corrected" estimates of the average treatment effect of tracking

Panel A: Without adjusting for covariates
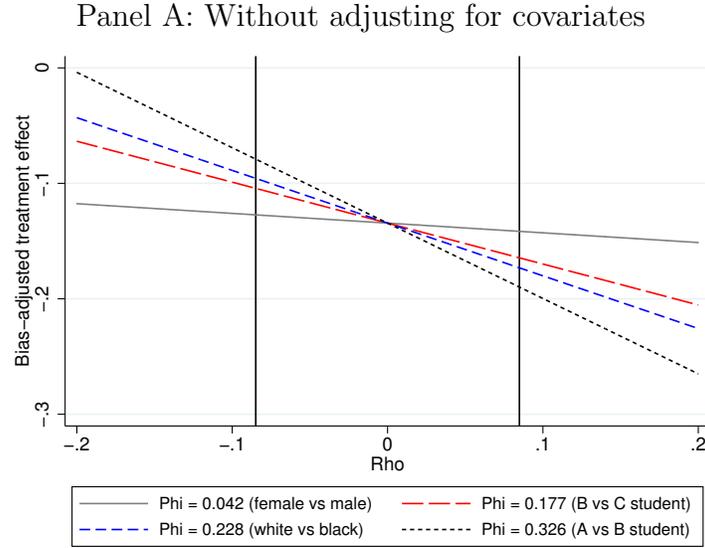
Panel B: Adjusting for individual covariates and dormitory fixed effects
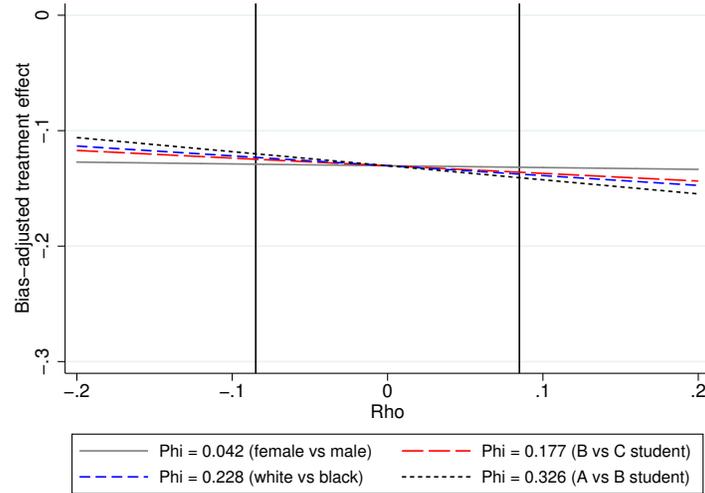
Notes: This figure explores how sensitive the results are to potential differences in unobserved student covariates between the tracking and random assignment periods. These horizontal axis ρ shows "how much" a hypothesized binary unobserved covariate differs between the two periods. The vertical lines provide a benchmark by showing the maximum difference between the two periods for any observed covariate from table 1. The ϕ parameter measures the strength of the relationship between GPA and the unobserved covariate. For example, the solid line indicates the value of ϕ associated with sex. The figure implies that to entirely explain the observed treatment effect, there would need to be an unobserved student covariate with a time trend twice as large as any observed covariate that predicts GPAs more strongly than race.

10

correlated then accounting for observed covariates will reduce the omitted variable bias. Assume the model is

$$\boldsymbol{GPA} = \boldsymbol{D} \cdot \delta_1 + \boldsymbol{T} \cdot \delta_2 + \boldsymbol{DT} \cdot \delta_3 + \boldsymbol{X} \cdot \boldsymbol{\delta} + \boldsymbol{Z} \cdot \boldsymbol{\phi} + \epsilon. \tag{3}$$

The OLS estimator of δ_3 from (3) is identical to that from the partitioned regression model

$$\boldsymbol{M_X GPA} = \boldsymbol{M_X D} \cdot \delta_1 + \boldsymbol{M_X T} \cdot \delta_2 + \boldsymbol{M_X DT} \cdot \delta_3 + \boldsymbol{Z} \cdot \phi + \epsilon \tag{4}$$

where $\boldsymbol{M_X}$ is the projection matrix $\boldsymbol{I} - \boldsymbol{X}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}$. $\boldsymbol{M_X Z} = \boldsymbol{Z}$ under the assumption that $\boldsymbol{Z}$ is uncorrelated with all elements of $\boldsymbol{X}$. I regress each vector in $(\boldsymbol{GPA}, \boldsymbol{D}, \boldsymbol{T}, \boldsymbol{DT})$ on $\boldsymbol{X}$, use the residuals from these regressions to estimate the parameters of model (3), and construct the biased-corrected estimator $\hat{\delta}_3^{BC}$ from $\hat{\delta}_3$ for selected values of (ρ, ϕ). ϕ is now interpreted as the relationship between GPA and academic orientation conditional on $\boldsymbol{X}$. I specify $\boldsymbol{X}$ to include linear and quadratic terms in HSGPA; an indicator for missing HSGPA; indicators for sex, language, nationality, and race; all pairwise interactions between these variables; and dormitory fixed effects.

I plot $\hat{\delta}_3^{BC}$ against ρ for selected values of ϕ in the second panel of figure 1. The bias-corrected treatment effects are attenuated slightly relative to their values without conditioning on $\boldsymbol{X}$ but the substantive conclusion of the exercise is unchanged.

B.4 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 support this hypothesis. First, the within-race peer effects documented in section IV 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 track-

11

ing 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 propose 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 will 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 random assignment.[4] Prediction (b) can be tested and I find that non-dormitory students with above-median HSGPAs have GPAs that are 0.044 standard deviations higher (standard error 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 (standard error 0.034) higher under tracking than random assignment. The first difference has the wrong sign and both are close to zero. This argument does not rule out the existence of some other social interactions framework that biases the treatment effect. But it does show that one particularly salient framework produces predictions that are not consistent with the data.

Second, the higher raw GPAs for non-dormitory students in the tracking

---

[4]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 $\hat{\gamma}_2 > 0$.

period 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.

B.5 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, truncation, 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 GPA and test score measures but I argue that it is less pressing in this context. Instructors at this university are not encouraged to use grading curves and many examinations are subject to external moderation intended to maintain an approximately time-consistent standard. I observe several patterns in the data that are not consistent with curving. Mean grades in the three largest introductory courses at the university (microeconomics, management, information systems) show year-on-year changes within an assignment policy period of up to 6 points (on a 0 to 100 scale, approximately 1/3 of a standard deviation). Similarly, the 75th and 25th percentiles of the grades within these large first-year courses show year-on-year changes of up to 8 and 7 points respectively. This demonstrates that grades are not strictly curved in at least some large courses. I also examine the treatment effect of tracking on grades in the introductory accounting course, which builds toward an external qualifying examination administered

by South Africa's Independent Regulatory Board for Auditors. This external assessment 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 0.08 standard deviations (cluster bootstrap standard error 0.10, sample size 2099 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. I cannot rule out this concern completely but I argue that it is unlikely to be very important. The nominal grade ceiling of 100 does not bind for any student: the highest grade observed in the dataset is 97/100 and the 99th percentile is 84/100. Some courses may impose ceilings below the maximum grade, which will not be visible in my data. However, the course convenors for Introductory Microeconomics, the largest first-year course at the university, confirmed that they used no such ceilings. The treatment effect of tracking on grades in this course is 0.13 standard deviations (cluster bootstrap standard error 0.05, sample size 4554 students), 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 course-taking behavior: 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 27.6% of her peers are in the same program of study under random assignment. This is 0.8 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 program of study range between -0.19 for engineering and -0.02 for medicine. The average treatment effect

Table 3: Treatment Effects with Alternative Grading Measures

	GPA	% of credits excluded	GPA | non-exclusion
	(1)	(2)	(3)
Dormitory ×	-0.128	0.026	-0.063
tracking period	(0.042)	(0.005)	(0.047)
Student covariates	×	×	×
Missing data indicators	×	×	×
Dormitory fixed effects	×	×	×
Faculty fixed effects	×		
Adjusted R^2	0.244	0.053	0.304
# dormitory-year clusters	58	58	58
# dormitory students	7410	7410	7381
# non-dormitory students	7188	7188	7043

Notes: Table 3 reports results from the robustness checks discussed in appendix B.5. Column 1 reports a difference-in-differences estimate including college/faculty/school fixed effects. Column 2 reports a difference-in-differences estimate with the credit-weighted percentage of courses from which students are academically excluded as the outcome. Column 3 reports a difference-in-differences estimate with GPA calculated using only grades from non-excluded courses as the outcome. The sample size in column 3 is smaller because students are excluded from the regression if they were academically excluded from all courses. Standard errors in parentheses are from 1000 bootstrap iterations, stratifying by assignment policy and dormitory status and clustering by dormitory.

with program of study fixed effects is -0.13 with standard error 0.04 (table 3, column 1). 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 3, columns 2 and 3). Tracking substantially increases the exclusion rate from 3.3 to 6.1% and reduces GPA in non-excluded courses by 0.06 standard deviations, though the latter effect not significantly different to zero. 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 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.

B.6 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 attributes 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 relation-

ship 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 attributes. 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 HSGPA and the attributes of low-track dormitories. I test this hypothesis by estimating equation (2) with an interaction between $HSGPA_{id}$ and the rank of dormitory d during the tracking period. The interaction term has a small and insignificant coefficient: 0.004, with standard error 0.005. 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 attributes.

This does not rule out the possibility of time-varying effects of dormitory attributes or of effects of time-varying attributes. 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 high school graduation test results are published in newspapers and the university publishes the minimum HSGPA required for admission to specific programs of study. 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 instruc-

tors observed students' dormitory assignments and assigned lower scores to students in low-track dormitories, 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 (standard error 0.05): almost identical to the average effect across all courses. I conclude that the headline results cannot be entirely explained by discriminatory grading.

C   Reweighted Nonlinear Difference-in-Differences Model

Athey and Imbens (2006) propose a model for recovering quantile treatment on the treated effects in a difference-in-differences setting. The standard linear difference-in-differences model recovers only the average treatment effect on the treated. Tracking is an inherently heterogeneous treatment, where students' treatment depends on their pre-treatment covariates, so it is likely to have heterogeneous treatment effects. In this appendix I describe Athey and Imbens's model, highlight departures from the identifying assumptions in my application, and explain how I implement the model to generate the results reported briefly in section III and discussed in detail in appendix D. I also explain how I condition on pre-treatment covariates.

The original model is identified under four assumptions. Athey and Imbens (2006) propose two models, with different identifying assumptions. I use only the "changes-in-changes" model. This is identified under arguably weaker assumptions than the alternative "quantile difference-in-differences" model. Results are not sensitive to the choice of model. I list each assumption and then discuss whether it is likely to hold in my application. Throughout this discussion, I use $T_i = 1$ to denote students in tracking period, $D_i = 1$ to denote dormitory students, and $\boldsymbol{X}_i$ to denote a vector of observed pre-treatment covariates.

A1: GPA in the absence of tracking is generated by the production function $GPA = h(U, \boldsymbol{X}, T)$. U is an unobserved scalar random variable and h is monotonically increasing in U.[5] This assumption implies that GPA does not depend directly on D. Neither the monotonicity assumption, nor the conditional independence of GPA and D are testable. However, I can test the related condition that the relationship between GPA and each element of $\boldsymbol{X}$ does not differ between non-dormitory students and randomly assigned dormitory students. To implement this test, I test if $\beta_3 = 0$ in the model $GPA_{id} = \beta_0 + \text{Dorm}_{id} \cdot \beta_1 + X_{id} \cdot \beta_2 + \text{Dorm}_{id} \cdot X_{id} \cdot \beta_3 + \epsilon_{id}$ for each pre-treatment covariate in $\boldsymbol{X}$. I reject this condition at the 5% level for three of the pre-treatment covariates listed in table 1.

A2: The distribution of U is constant through time for each group: $U \perp T | D, \boldsymbol{X}$. This assumption is not directly testable. However, I can test the related condition that the distribution of each element of $\boldsymbol{X}$ is constant through time. I reject this condition for some binary demographic measures and for some summary statistics of the HSGPA distribution shown in table 1.[6]

A3: The support of dormitory students' GPA is contained in that of non-dormitory students' GPA: $supp(GPA | D = 1, \boldsymbol{X}) \subseteq supp(GPA | D = 0, \boldsymbol{X})$. This assumption is testable and holds for my full dataset and for each subsample defined by a discrete element of $\boldsymbol{X}$, as listed in table 1.

A4: The distribution of GPA is strictly continuous. This assumption is testable and holds approximately in my data. There are 5490 unique GPA values for 14598 observations. No value accounts for more than 0.3% of the observations.

I conclude that assumptions $A3$ and $A4$ are plausible in this application, but that assumptions $A1$ and $A2$ may not hold. In particular, the time

---

[5]As is standard for identification of treatment-on-the-treated parameters, no assumption is required about the model that generates treated outcomes. So tracked students may experience a completely different GPA production function with complex interactions between own and peer characteristics.

[6]I also reject the equality of the distribution HSGPA across the tracking and random assignment periods using Kolmogorov-Smirnov tests, implemented separately for dormitory and non-dormitory students.

trends in several elements of $\boldsymbol{X}$ mean that it is important to condition on pre-treatment covariates. Under these four assumptions, the quantile treatment effect of tracking on the tracked students at each quantile q is given by the horizontal distance between the observed GPA distribution for tracked students and the counterfactual distribution:

$$
\begin{aligned}
\Delta(q) = {} & \mathbb{E}_{\mathcal{X}} \left[F^{-1}_{GPA|D=1,T=1,\boldsymbol{X}}(q) \right] \\
& - \mathbb{E}_{\mathcal{X}} \left[F_{GPA|D=1,T=0,\boldsymbol{X}} \left(F^{-1}_{GPA|D=1,T=0,\boldsymbol{X}} \left(F_{GPA|D=1,T=0,\boldsymbol{X}}(q) \right) \right) \right],
\end{aligned}
\tag{5}
$$

where the expectation is taken over the joint distribution of $\boldsymbol{X}$. Intuitively, we construct the counterfactual GPA distribution in three steps. Consider student A, who lives in a dormitory during the random assignment period and has GPA g and pre-treatment covariates $\boldsymbol{x}$. Under assumptions $A1$ and $A4$, A must have unobserved scalar $u = h^{-1}(g; \boldsymbol{x}, 1)$. Student B, who also has GPA g, has pre-treatment covariates $\boldsymbol{x}$, and attends the university during the random assignment period but does not live in a dormitory, will have the same unobserved scalar u. We can compare *GPA levels* across dormitory and non-dormitory students because GPA does not depend directly on whether students live in dormitories. Under assumptions $A1$ and $A2$, B will have the same rank in the GPA distribution as student C, who has unobserved scalar u, pre-treatment covariates $\boldsymbol{x}$, does not live in a dormitory and attends the university during the tracking period. We can compare *GPA ranks* across time periods because there are no time trends in the distribution of $F_U(\cdot)$ and because h is a monotonic function of u. We can now compare C to student D, who has unobserved scalar u, pre-treatment covariates $\boldsymbol{x}$, and lives in a dormitory during the tracking period in the counterfactual world in which tracking was not implemented. Under assumptions $A1$ and $A4$, students C and D will have the same GPA. We have therefore identified one value in the counterfactual GPA distribution for tracked students in the absence of tracking. We repeat this exercise for all values of GPA observed amongst dormitory students to construct the entire counterfactual distribution. $A3$

ensures that for all GPA values observed amongst dormitory students, there are non-dormitory students with the same GPAs to allow comparison.

I condition on $\boldsymbol{X}$ using a weighting procedure that reweights the sample of students in the random assignment period to have the same distribution of pre-treatment covariates as the sample of students in the tracking period. Specifically, I estimate a logit regression of dormitory status on pre-treatment covariates, separately for dormitory and non-dormitory students, construct the probability that any student will appear in the random assignment period $Pr(T = 1|D, \boldsymbol{X})$, and construct $\omega(\boldsymbol{X}, D, T) = T + \frac{(1-T)\cdot Pr(T=1|D,\boldsymbol{X})}{Pr(T=0|D,\boldsymbol{X})}$. The ω term equals one for all students in the tracking period and is large (respectively small) for students in the random assignment period whose pre-treatment covariates are similar to (respectively different from) those in the tracking period. I can then define the reweighted GPA distribution $F_{GPA_\omega^{D0}}(\cdot)$ as the distribution of $GPA \cdot Pr(T = 1|D, \boldsymbol{X})/Pr(T = 0|D, \boldsymbol{X})$ and rewrite equation (5) as:

$$F_{GPA|D=1,T=1}^{RW,CF}(g) = F_{GPA_\omega^{10}}\left(F_{GPA_\omega^{00}}^{-1}\left(F_{GPA^{01}}(g)\right)\right) \tag{6}$$

This is a direct adaptation of the reweighting techniques used in wage decompositions and program evaluation (DiNardo, Fortin and Lemiuex, 1996; Firpo, 2007; Hirano, Imbens and Ridder, 2003). Athey and Imbens (2006) recommend two alternative ways to account for pre-treatment covariates, neither of which is appropriate in this application. First, a fully nonparametric method that applies the model separately to each value of the pre-treatment covariates. This is feasible only if the dimension of $\boldsymbol{X}$ is low. Second, a parametric method that applies the model to the residuals from a regression of GPA on $\boldsymbol{X}$. This is valid only under the strong assumption that the pre-treatment covariates $\boldsymbol{X}$ 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 any element of $\boldsymbol{X}$. For example, different treatment effects on students with high and low HSGPAs would violate this restriction.

I implement the model in four steps:

1. For $D \in \{0,1\}$, regress T on student sex, language, nationality, race, linear and quadratic terms in HSGPA, all twoway interactions, and an indicator variable for observations with missing HSGPA using a logit model.

2. Construct the predicted probability $\hat{Pr}(T_i = 1 | D_i, \boldsymbol{X}_i)$ for each student i in the random assignment period.

3. Evaluate equation 6 at each half-percentile of the GPA distribution (i.e. quantiles 0.5 to 99.5). The first panel of figure 2 displays this counterfactual GPA distribution for tracked students, along with the observed GPA distribution.

4. Calculate the difference between observed and counterfactual distributions at each half-percentile. The second panel of 2 shows this difference.

5. 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:
$\mathbb{E}\left[GPA^{RW,CF}\right] \approx \frac{1}{198} \cdot \sum_{p=2}^{199} \left[\frac{1}{2}F_{GPA^{11}}^{RW,CF}(p) + \frac{1}{2}F_{GPA^{11}}^{RW,CF}(p-1)\right]$. 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 estimating both quantile and summary treatment effects using this model is available at www.robgarlick.com/code.

D   Quantile and Inequality Effects of Tracking

In this appendix I report results from using the nonlinear difference-in-differences model to estimate quantile treatment effects of tracking on the tracked students (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 2, 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 2, second panel). The point estimates are large and negative in the bottom quintile (0.2 - 1.1 standard deviations), small and negative for most of the distribution ($\leq$ 0.2 standard deviations), and small and positive in the top decile ($\leq$ 0.2 standard deviations). The estimates are relatively imprecise; the 95% confidence interval excludes zero only in the bottom tercile.[7] This reinforces the conclusion that the negative average effect of tracking is driven by large negative effects on students with low academic performance, whether that performance is measured in terms of university GPA or high school graduation test scores.

There is no necessary relationship between figures 3 and 2. Figure 3 shows that the average treatment effect of tracking is large and negative for students with low HSGPAs. Figure 2 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 HSGPAs. However, they do not recover treatment effects

---

[7]I construct pointwise 95% confidence intervals 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.

on specific students or groups of students without additional assumptions. See Bitler, Gelbach and Hoynes (2016) for further discussion on this relationship.[8]

I also report results from two alternative approaches to estimating this model. First, I show results without using reweighting to account for differences in pre-treatment covariates (first panel of figure 3). Second, I show results after dropping students with missing HSGPA instead of using a missing data indicator (second panel of figure 3). The point estimates differ slightly but the general pattern of results is consistent across all three implementations: large negative effects in the left tail, small negative effects in the middle of the distribution, and small positive effects in the extreme right tail.[9]

The nonlinear model provides substantially more information than the average treatment effect but requires stronger identifying assumptions. In particular, 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 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. The time trends in some covariates shown in table 1 cast doubt on this identifying assumption but the similarity of the quantile treatment effects with and without adjusting for covariates shows that violations of this assumption are not necessarily quantitatively important.

The counterfactual GPA distribution estimated above also provides information about the relationship between tracking and the dispersion of academic outcomes. Specifically, I calculate several standard measures of dispersion or

---

[8]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.

[9]I also estimate the model using two alternative specifications of $\boldsymbol{X}$: omitting the quadratic term and twoway interactions, and including a cubic term and three-way interactions. Results are robust across these specifications as well.

Figure 2: Quantile Treatment Effects of Tracking on the Tracked Students
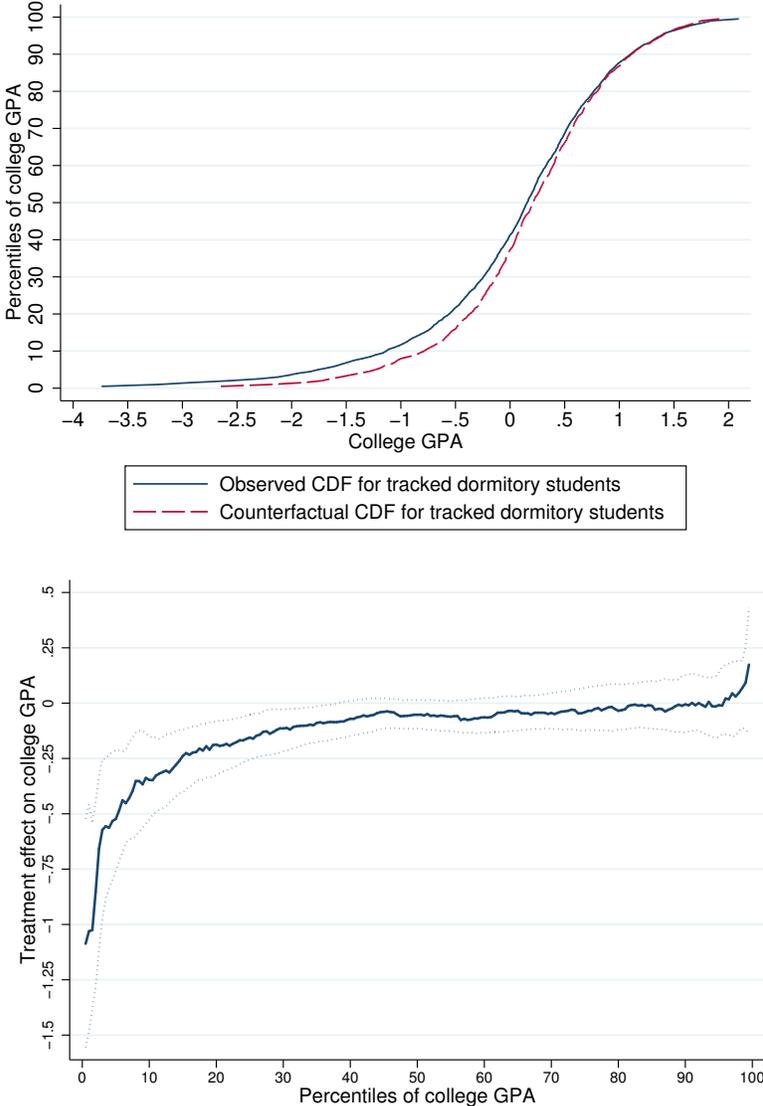
Notes: The first panel shows the observed GPA distribution for tracked dormitory students (solid line) and the counterfactual distribution constructed using the reweighted nonlinear difference-in-differences model (dashed line). The propensity score weights are constructed from a model including student sex, language, nationality, race, a quadratic in HSGPA, 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.

Figure 3: Quantile Treatment Effects of Tracking on the Tracked Students
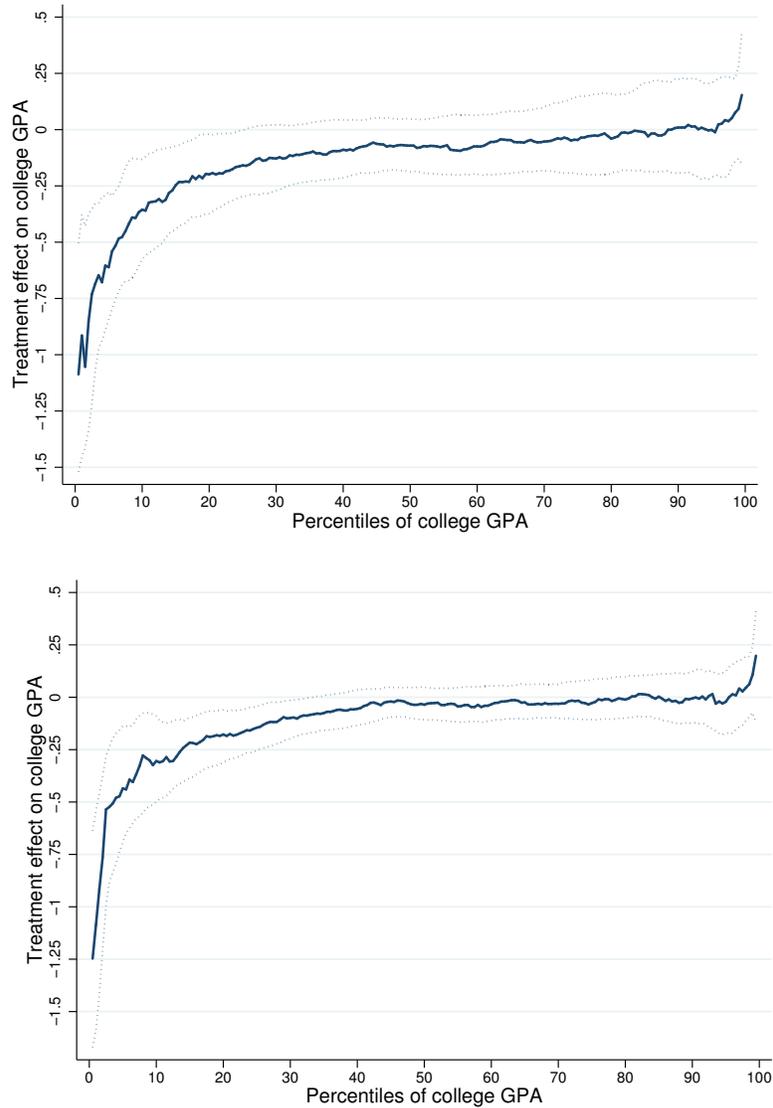
Notes: Both panels show quantile treatment effects of tracking on the tracked students, with different approaches to adjusting for covariate differences between the tracking and random assignment periods. The first panel shows treatment effect without any adjustment for covariates using the original Athey-Imbens model. The second panel shows treatment effects with reweighting used to adjust for differences: student sex, language, nationality, race, a quadratic in high school graduation test scores, all pairwise interactions, and dormitory fixed effects. Students with missing high school graduation test scores are dropped from the sample used to construct the second panel. The dotted lines show 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.

inequality on the observed and counterfactual distributions. The differences between these measures are the treatment effects of tracking on the dispersion of tracked students' GPAs.[10] The literature on academic tracking emphasizes inequality concerns (Betts, 2011) and the effect of tracking on dispersion or inequality is sometimes obvious from quantile treatment effects. However, this is the first study of which I am aware to quantify the effect of tracking on the dispersion of outcomes. Existing results from the econometric theory literature can be applied directly to this problem (Firpo and Pinto, 2016; Rothe, 2010). Identification of these effects requires no additional assumptions beyond those already imposed in the quantile analysis.

Table 4 shows dispersion measures for the observed and counterfactual GPA distributions. The standard deviation, interquartile range, and interdecile range are all significantly higher under tracking than under the counterfactual. Tracking increases the interquartile range by approximately 12% of its baseline level and the other measures by approximately 18%. This reflects the particularly large negative effect of tracking on the lowest quantiles of the GPA distribution. Tracking thus decreases mean GPA and increases GPA dispersion. Knowledge of the quantile and dispersion 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.

---

[10]I apply the same principle to calculate mean GPA for the counterfactual distribution. The observed mean is 0.125 standard deviations lower than the counterfactual mean (cluster bootstrap standard error 0.041). This is very similar to the average effect from the linear difference-in-differences models reported in section II.

Table 4: Treatment Effects of Tracking on GPA Dispersion

	(1) Observed distribution	(2) Counterfactual distribution	(3) Treatment effect	(4) Treatment effect in % terms
Mean	0.052	0.177	-0.125	-70.62
	(0.062)	(0.045)	(0.041)	
Standard deviation	0.909	0.769	0.140	18.21
	(0.028)	(0.027)	(0.033)	
Interquartile range	1.023	0.909	0.114	12.54
	(0.046)	(0.039)	(0.052)	
Interdecile range	2.238	1.896	0.342	18.04
	(0.087)	(0.085)	(0.108)	

Notes: Table 4 reports summary measures of academic dispersion 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 C. 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 iterations, where I estimate the full distributions, the summary statistics, and the treatment effect in each iteration.

References

Altonji, Joseph, Todd Elder, and Christopher Taber. 2005. "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools." *Journal of Political Economy*, 113(1): 151–184.

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

Betts, Julian. 2011. "The Economics of Tracking in Education." In *Handbook of the Economics of Education Volume 3.* , ed. Eric Hanushek, Stephen Machin and Ludger Woessmann, 341–381. Elsevier.

Bitler, Marianne, Jonah Gelbach, and Hilary Hoynes. 2016. "Can Variation in Subgroups' Average Treatment Effects Explain Treatment Effect Heterogeneity? Evidence from a Social Experiment." *Review of Economics and Statistics*, forthcoming.

DiNardo, John, Nicole Fortin, and Thomas Lemiuex. 1996. "Labor Market Institutions and the Distribution of Wages, 1973 - 1992: A Semiparametric Approach." *Econometrica*, 64(5): 1001–1044.

Firpo, Sergio. 2007. "Efficient Semiparametric Estimation of Quantile Treatment Effects." *Econometrica*, 75(1): 259–276.

Firpo, Sergio, and Cristine Pinto. 2016. "Identification and Estimation of Distributional Impacts of Interventions Using Changes in Inequality Measures." *Journal of Applied Econometrics*, 31(3): 457–486.

Garlick, Robert. 2012. "Mobility Treatment Effects: Identification, Estimation and Application." Mimeo.

Heckman, James, and Joseph 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.

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, Keisuke, Guido Imbens, and Geert Ridder. 2003. "Efficient Estimation of Average Treatment Effects Using the Propensity Score." *Econometrica*, 71(4): 1161–1189.

Oster, Emily. 2016. "Unobservable Selection and Coefficient Stability: Theory and Validation." *Journal of Business and Economic Statistics*, forthcoming.

Rosenbaum, Paul, and Don Rubin. 1983. "Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome." *Journal of the Royal Statistical Society Series B*, 45(2): 212–218.

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

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