Peer effects under tracking and mixing: Evidence from South African college dormitories[*]

Robert Garlick[†]

August 8, 2011

Abstract

I study a natural experiment in the South African higher education system that provides unique insights into the role of peer effects under different peer assignment regimes. A large research university historically tracked incoming students into dormitories, so that they lived with peers with similar levels of baseline academic performance. In 2006, the university instituted a regime of random assignment to dormitories for incoming students. I compare the distribution of outcomes under the two regimes, using students living outside the dormitory system to control for time trends in achievement.

This research design estimates peer effects using variation across the type of assignment regime used to construct peer groups. I argue that this is more informative for a policy-maker interested in the optimal peer assignment regime than existing research designs, almost all of which use variation across peer groups within a given regime. In this design, high and low track students differ only in the composition of their residential peer groups but otherwise attend the same classes and experience the same institutional environment. This allows me to isolate the pure peer effect of tracking, avoiding the conflation of peer effects with differences in school resources or teaching strategies that face studies of school- or classroom-level tracking.

I estimate that tracking reduced students' grade point averages (GPA) by approximately 0.1 standard deviations on average. I then integrate existing reweighting and nonlinear difference-in-difference models to estimate the full counterfactual distribution of outcomes in the absence of tracking. This demonstrates that treatment effects are particularly large and negative in the lower tail (up to 0.5 standard deviations) but are non-positive up to the highest quantiles, suggesting that tracking did not help even the strongest students who might *a priori* be expected to benefit the most. This distribution of treatment effects increased standard measures of inequality by up to 15%.

I interpret these results as evidence that in this setting own and peer ability are substitutes in the production of academic achievement. Under this assumption, students seek to engage in negative assortative matching, with strong students pairing with weak students. However, tracking raises the cost of matching with dissimilar peers and so pushes students toward welfare-reducing positive assortative matching.

JEL codes: I24, I25, O15

[*] I thank John DiNardo, Brian Jacob, and Jeffrey Smith for comments on earlier drafts of this paper and Raj Arunachalam, Emily Beam, John Bound, Tanya Byker, Susan Godlonton, Andrew Goodman-Bacon, Italo Gutierrez, Brad Hershbein, David Lam, Rebecca Thornton, Dean Yang and seminar participants at the University of Michigan, PACDEV conference and MEA conference for helpful suggestions. All errors are my own.

[†] PhD candidate in economics and public policy, University of Michigan; rgarlick@umich.edu

1 Introduction

Peer effects have grown into an important topic of economic inquiry over the past few decades, with a growing literature examining the role of peer groups in determining education (Angrist and Lang, 2004), health (Fowler and Christakis, 2008), labor market outcomes (Mas and Moretti, 2009) and participation in social programs (Bertrand, Luttmer, and Mullainathan, 2000). Manski (1993) lays down the fundamental challenge for identifying peer effects: peer groups may form due to similarities in unobserved characteristics and group members typically face common environmental factors. These unobserved individual and group characteristics may influence outcomes of interest and complicate separate identification of the causal effect of peer interaction.

A large education literature responds to this challenge by exploiting natural experiments in which institutional rules generate random variation in peer (group) characteristics (Sacerdote, 2001; Foster, 2006; Carrell, Fullerton, and West, 2009). Most of these research designs compare the outcomes of individuals who are randomly assigned to peers with different baseline characteristics and so produce well-identified estimates of the across-peer or across-group effects for the given random assignment regime. For example, Stinebrickner and Stinebrickner (2006) find that students obtain higher grades in college when randomly assigned to live with roommates with high grades from high school. A related literature estimates peer effects in academic environments employing tracking regimes (Jackson, 2010; Pop-Eleches and Urquiola, 2011). These papers exploit the fact that when students are tracked into schools or classrooms on a baseline measure of academic performance, there are cutoff levels of performance that sharply separate otherwise similar students into high and low tracks with peers of very different ability. This generates (locally) well-identified estimates of the across-group effects for the given tracked assignment regime. For example, Ding and Lehrer (2007) find that students whose entrance test scores qualify them for admission to “high track” high schools in China obtain higher scores on college entrance tests.

Both research designs estimate peer effects using *across-group within-regime* variation in peer characteristics, where the assignment regime is either random or tracked. This sheds light on the effect of assignment to different peers or peer groups under each regime. However, these designs cannot speak to the effect of *different assignment regimes*. This is arguably the more relevant parameter for a policy-maker who wishes to find the optimal rule for grouping a given set of students in order to maximize some measure of aggregate academic achievement. This distinction has been explored in a number of theoretical papers (Arnott, 1987; Glewwe, 1997) but has received little attention in the empirical literature. Duflo, Dupas, and Kremer (2011) appear to provide the only well-identified estimates of peer effects using across-regime variation, rather than simply across-group variation for a given assignment regime. Carrell, Sacerdote, and West (2011) provide a cautionary example of the risks of using peer effects estimated using within-regime variation to predict the effect of changing regimes. They assign students to groups to maximize achievement by the lowest ability students using estimates derived from a previous random assignment regime and find that these students are actually hurt by their intervention.

I address this gap in the literature by analyzing the causal effect of tracking into peer groups on students' academic outcomes, relative to random assignment. I study a natural experiment at a South African university: incoming students were historically tracked into dormitories to live with students whose academic performances in standardized high school graduation tests was similar to their own. This tracking regime was replaced in 2006 with a policy of randomly assigning incoming students to dormitories. I use this policy change to identify the effect of tracking on the distribution of students' grade point averages (GPAs), relative to random assignment. I find a robust and precisely estimated 0.1 standard deviation fall in GPA under tracking. This effect is concentrated amongst students with low socio-economic status and low high school graduation test scores. I also construct the full counterfactual distribution of GPAs that would have occurred in the absence of tracking. This shows that the treatment effect of tracking is particularly large for students at the bottom of the distribution, with effects of up to 0.5 standard deviations at the 5^{th} and 10^{th} percentiles. Perhaps surprisingly, effects are positive (though very small and imprecisely estimated) at even the highest quantiles, suggesting that tracking had little effect on high achieving students. To illustrate the extent to which tracking is a regressive policy intervention, I also show that tracking significantly increases GPA inequality by up to 15%.

I claim that I can assign a causal interpretation to these effects for three reasons. First, I use students living in private accommodation as a control group to remove any common trends in unobserved student characteristics or grading policies. Second, I argue that the nature of the policy change, particularly the fact that it was not widely communicated, limited the scope for students to respond to the policy change by choosing when to attend the university and whether to live in a dormitory. Third, I control for a rich set of baseline student characteristics and show that my estimates are highly robust to a range of flexible controls for students and dormitory characteristics.

This paper contributes to the literature on three levels. First, I provide what appears to be only the second study of the effect of different peer assignment regimes. Second, I am able to separate the pure peer effects of tracking from other conflating factors. The research design employed by Duflo, Dupas, and Kremer (2011) cannot separate direct peer effects from indirect effects operating through teachers' behavioral responses to facing relatively homogeneous or heterogeneous classrooms (under tracking and random assignment respectively). I argue that part of the reason I find negative effects of tracking while they find positive effects is that peer effects in their research design are offset by teachers' behavioral responses. Research designs using across-group variation under a tracking regime typically face even greater problems, as high and low track schools and classrooms differ in many respects other than student characteristics. For example, the distribution of teachers differs greatly across high and low track Chinese and Romanian high schools studied by Ding and Lehrer (2007) and Pop-Eleches and Urquiola (2011) respectively. Raudenbush and Nomi (2011) show that the effect of factors such as instructional strategies, resource allocations and institutional environments can be separated from peer effects only under very strong assumptions. In contrast to the existing tracking literature, I study a setting in which there are no substantial differences between tracks (dormitories) other than student characteristics. Students in different dormitories attend the same classes, are taught by the same instructors, and use the same libraries and computer rooms. There are some small differences in the institutional and physical features of the individual dormitories but I show that these features explain little of the treatment effect. Third, I engage seriously with the heterogeneous treatment effects that seem *ex ante* likely in this context. The nature of the treatment experienced by students under tracking depends inherently on their own characteristics, so it is important to pay close attention to how the treatment effect varies with both observed and unobserved student characteristics. I estimate not only mean treatment effects but also quantile and inequality treatment effects, which provides a richer and more comprehensive picture of the effects of tracking.

I begin in section 2 by describing the nature and setting of the experiment, including a detailed description of the two assignment regimes and the nature of the transition between them. I introduce the data in section 3. Section 4 develops a linear difference-in-differences strategy that compares the change in the grades of dormitory students between the tracking and the mixing periods to the change in the grades of non-dormitory students over the same time period. I lay out the assumptions necessary for identification of this effect, which informally require that students do not make strategic decisions about whether to attend the university, when to attend it and whether to live in a dormitory based on unobserved characteristics that also affect their expected outcomes under different assignment regimes. I then report estimates of the average treatment effect on the treated arising from this strategy under a variety of specifications. Section 5 explores heterogeneity in this average treatment effect over observed student characteristics by repeating this exercise for relevant subgroups of the student population. In section 6 I extend the analysis to explore heterogeneity on unobserved student characteristics. I use the nonlinear difference-in-differences estimator proposed by Athey and Imbens (2006) to estimate the full set of quantile treatment effects on the treated. I then use this counterfactual distribution to estimate inequality treatment effects on the treated in section 7, adapting methodology proposed by Firpo (2010). Section 8 concludes by exploring which restrictions on the matching process and educational production function are consistent with the estimated treatment effects. In particular, I show that these effects suggest that the gains from interacting with peers of high baseline academic performance are largest for individuals with low baseline performance, or that the cross-partial derivative of own achievement with respect to own and peer baseline performance is negative.

2 The experiment

I study a selective research university in a large South African city. The university teaches a class of approximately 4000 incoming freshmen each year, some of whom live in university dormitories and some of whom live in private accommodation. Up to and including the 2005 academic year the university tracked incoming first year students into dormitories, assigning students to live with peers who had obtained similar scores on a standardized high school graduation test. From the 2006 academic year onward, the university randomly assigned students into dormitories with the intention of achieving approximately academically balanced dormitories. As I explain below, the assignment regimes deviated from pure tracking and pure randomization in several small respects. However, the change in regime generated a large change in the characteristics of students' residential peers, conditional on their own characteristics, and it is this cross-regime variation that I use for identification.

During both the tracking and mixing periods the university's dormitory system was heavily overscribed, with more than two applications for every available place.[1] Given this excess demand, places were rationed on three criteria. First, students were normally permitted to live in undergraduate dormitories only during their first two years of study. Exceptions are granted for those with medical conditions that limit their ability to live in private accommodation and those employed by the dormitory as administrators. Second, students who live farther from the university were given priority for admission. In practice this means that only students who live outside the local city or in poor neighborhoods in the urban periphery were typically admitted.[2] Third, a small number of high achieving students who live near the university were exempted from the geographic rationing and admitted to the dormitories as a recruitment device. There is no formal record of the number of such exemptions but admissions officers report that the group was small and never more than a few percent of the incoming cohort.

In addition to the formal admissions criteria, the time at which a student was admitted to the

[1] This section draws heavily on interviews with the university's director of admissions, manager of residence life, and senior coordinator for residence life.

[2] The residential segregation imposed under *apartheid* created an urban geography very different to that in the USA. White suburbs were typically relatively close to the urban center, with black citizens forced to live in "townships" considerably farther away. This pattern has largely persisted into the post-apartheid period, particularly in the city in which this university is located. Hence, suburban students typically live too close to be admitted to the dormitories, while township students are eligible.

university migh have also affected their probability of assignment to a dormitory. Students submit applications to the university between July and October to start study in January of the following year. Academic admissions are made on a rolling basis and students are assigned to dormitories at the time of their admission to the university. Later applicants have a higher probability of assignment to "transit" accommodation. Students in transit live in temporary accommodation at the beginning of the academic year and are assigned to permanent dormitories once the university knows how many admitted students are "no-shows." Each year 5 - 10% of students admitted to the dormitory system are placed in transit accommodation and they are always assigned a permanent place within a few weeks. The available data show only students' final dormitory allocations, not whether they were initially assigned to transit accommodation. Students who apply to the dormitory system but are not admitted live in private accommodation; this typically means living with family in the local city but some also live in rented houses or apartments.

Students did not have the option of moving from one dormitory to another under either assignment regime. More than 90% of all rising second year students remained in the same dormitory they had lived in for their first year and the remainder moved into private accommodation. When the assignment mechanism changed from tracking to mixing, returning second year students were not reassigned to new dormitories. Hence, incoming first year students in 2006 were assigned to a mixed treatment regime: their residential peers of the same cohort were randomly assigned but the older students with whom they lived were tracked. Given this mixed treatment, I omit the 2006 incoming cohort from the main analysis and use only the 2004 and 2005 cohorts for the tracking group and the 2007 and 2008 cohorts for the mixing group. (Data on dormitory allocations for years prior to 2004 are not available.)

Under the tracking regime, students were assigned to dormitories based on their result in a standardized high school graduation examination, which was also used for admissions decisions. There were, however, three deviations from a perfect tracking assignment, in which dormitories partition the distribution of high school grades. First, the assignment regime included an affirmative action component, so black and coloured students required lower scores for assignment to high performance dormitories than white and Indian students.[3] Second, transit students were assigned to dormitories as spaces became available, without regard to their high school grades. Third, students were assigned when they were admission to the university based on their *expected* rank, as the full distribution of high school grades was not yet known to admissions officers.

[3] South Africa's population was divided into four groups under *apartheid*'s racial classification system: black, white, Indian and colored, a residual category including mixed race individuals, and descendants of Malaysian slaves and the Khoisan groups from the country's west and south coasts. Given the ongoing salience of racial divisions, these distinctions are still in widespread use in social science research, public discourse, and government policy and I make extensive use of them in this paper.

Under the mixing regime, students were randomly assigned to dormitories conditional on their race. At the time of their admission, students were assigned to a dormitory based on a race-blind random number generator. The staff member responsible for assignments then regularly checked the racial composition of each dormitory and if he believed that one race group was underrepresented, he manually assigned the next few applicants of the underrepresented race to that dormitory. The criteria for underrepresentation or the number of manual assignments required to correct it were not formally established and were entirely at the discretion of the relevant staff member.

This raises the possible concern that students were able to manipulate their dormitory allocation by lobbying staff members involved in the assignment process. When interviewed, the director of the admissions office acknowledged that this was a risk under both the tracking and mixing regimes but stated that "everyone involved in the process was instructed to present a united front ... that residence allocations were final." I cannot directly replicate the algorithm used to assign students to dormitories as I do not have access to some of the information used in assignment: the date of admission and whether the student was initially placed in transit. However, I can explore whether the observed assignment of students is broadly consistent with the official policy.

Figure 1 shows the mean and $5^{th}/95^{th}$ percentiles of high school grades for each dormitory in the tracking and the mixing period, separately for female-only, male-only, and mixed-gender dormitories. Note that the university assigns each applicant a score of 0 to 48 points based on their performance in high school graduation test, with almost all accepted applicants scoring at least 20 points.[4] The change from tracking to mixing is accompanied by a large increase in the spread of high school grades within dormitories and a reduction in the range of dormitory means. Furthermore, there is little evidence of rank persistence between the periods: the high track dormitories in which strong students lived under tracking have lower means under mixing than many of the low track dormitories.[5]

For an alternative illustration of the difference between the two regimes, consider the decomposition of the total variance of high school grades into within- and across-dormitory components:

$$\mathbb{V}[X] = \mathbb{E}[\mathbb{V}[X|D]] + \mathbb{V}[\mathbb{E}[X|D]].$$

[4] More specifically, the measure of high school performance used for both admission to the university and assignment to dormitories is based on a standardized content-based high school graduation test in six subjects. It ranges from 0 to 48 points, with essentially all admitted students obtaining more than 20 points. A modified algorithm is applied to results on A-level and International Baccalaureate examinations for international students.

[5] There is one clear outlying dormitory, with considerably lower average and tail grades under both assignment regimes. This dormitory was the only self-catered option available to incoming first year students and so charged a lower fee than the other, catered, dormitories. Students were allowed to request placement in this dormitory for financial reasons and the dormitory consists almost exclusively of black students with very low grades on the high school graduation test. My results are, however, entirely robust to excluding this dormitory, which accounts for less than 2% of the sample.

The across-dormitory variance is 10 (standard error .5) during the tracking period and 1.8 (standard error .3) during the mixing period. The standard errors are estimated using 1000 replications of a pairs bootstrap that resamples individual students, stratifying by year. This again illustrates a large reduction in the variation across dormitories during the tracking period. However, the within-dormitory variance is still substantial in the tracking period: 23.5 (standard error .6) compared to 30.6 (standard error .8) in the mixing period. This reflects three factors: the different tracking thresholds by race, the untracked assignment of transit students and the long left tail of the distribution of high school points (see figure 3). The last factor in particular results in high within-dormitory variance in low track dormitories and keeps the within-dormitory component of the variance moderately high under tracking. This can also be seen in figure 2, which illustrates the significant reduction in across-dormitory variance under the mixing regime and shows the high within-dormitory variance for low track dormitories. The four outliers in the left tail are the four year-level observations for the single self-catered dormitory to which students were non-randomly assigned. Both the variance decomposition and the two figures demonstrate that there was a major shift in the way students are assigned between the two periods.

Furthermore, any problems in the assignment of individual students to dormitories do not directly undermine the identification strategy used in this paper: instead they affect the interpretation of the treatment effects. As discussed in the next four sections, the identification strategy rests on assumptions about the relative trends in unobserved characteristics of dormitory and non-dormitory students between the two periods. The differential tracking thresholds by race and the non-tracked assignment of transit students are not direct threats to identification but rather mean that the treatment effects should be interpreted as effects of *partial*, rather than *complete* tracking. *Prima facie*, the this might be expected to attentuate the treatment effect toward zero relative to the treatment effect of complete tracking.

3 Data

Tables 1 and 2 show the distribution of demographic characteristics (gender, race, language and nationality) and high school graduation test scores for dormitory and non-dormitory students in each period. The tracking period pools together students entering the university in 2004 and 2005, while the mixing period pools together students entering the university in 2007 and 2008. As noted in the previous section, students entering the university in 2006 are omitted from the analysis sample. I focus this discussion on the changes in observed student characteristics between the two periods for dormitory and non-dormitory students, as these are most relevant to the identification strategy presented in detail in the next section. I show that there is limited evidence of changes across the two periods and in particular of differential changes across the two groups of students. The pattern of changes are not consistent with the most obvious form of selection: high achieving students prefering to live in dormitories under tracking and in private accommodation under mixing, with low achieving students displaying the opposite behavior. This provides some reassurance that such selection should not be a major concern for the research design.

The dormitories in both periods clearly contain a dispropotionately large percentage of the university's black students, foreign students and students who do not speak English as a home language. This is consistent with the rules used to ration places in the dormitories discussed in the previous section and the fact that the local city has disproportionately few black residents and disproportionately many English speaking residents.

Dormitory students in the tracking period are significantly more likely to be English-speaking international students than during the mixing period. Most of this difference reflects the considerably worse political and economic fortunes of Zimbabwe during the mixing period, which sharply reduced the number of English-speaking international students. This trend was also visible amongst non-dormitory students but as almost all international students live in dormitories, the magnitude of the change is considerably smaller.

Table 2 and figure 3 perform a similar exercise for the distribution of high school grades. Mean high school grades are slightly lower amongst dormitory students during the tracking period and this change appears to be driven by a fall in the lower tail of the distribution. This suggests that there are marginally more students with low baseline academic performance in dormitories under the tracking period than the mixing period, rather than the opposite pattern of selection that might be expected.

The relatively small changes through time in baseline demographic and academic characteristics suggest that changes in observed characteristics are not a significant concern for identification. This provides some reassurance that changes in unobserved student characteristics may also be small, though as Bruhn and McKenzie (2009) note, this is true only to the extent that unobserved and observed characteristics are closely correlated. The manner in which the assignment regimes were changed provides additional reassurance that there are unlikely to be substantial changes in unobserved student characteristics through time. First, interviews with university staff and faculty confirm that there were no significant changes in admissions or grading policies at the same time that the regime change occurred. Second, the regime change was not widely publicised. There was no official accouncement by the university, nor any coverage in local or national media outlets. The criteria for assignment to dormitories were explicitly stated in the information provided by the university to prospective students but this information was not prominently displayed and no mention was made of the fact that the regime had changed in 2006. This suggests that at the time of their application, relatively few students may have been aware of the regime during either period, or the fact that it had changed. Informal discussion with students indicated that some of them became aware of the assignment regime in place at the time only after spending several months at the university. Third, university staff reported that under both regimes almost all students from the local city lived in private accommodation and almost all other students lived in dormitories.[6] All of these factors suggest that there was relatively little scope for students to strategically choose whether to live in a dormitory based on their expected outcome from living in a tracked or mixed dormitory relative to private accommodation.

4 Mean effects

I begin the analysis with a standard difference-in-differences design that compares the difference in the outcomes of dormitory students under tracking and mixing to the difference in the outcomes of non-dormitory students under the two regimes. This strategy identifies the mean effect of tracking (relative to mixing) under the relatively weak restriction on the data generating process that the time trend in the outcomes of both dormitory and non-dormitory students would be identical if no change had taken place in the dormitory allocation regime. These results are arguably the most comparable to those explored in the existing literature, which focuses on mean effects, though my estimates are identified off cross-regime variation in peer characteristics, rather than cross-group within-regime variation.

More formally, the first strategy identifies the the average treatment effect on the treated, defined as

$$\Delta^{ATT} = \mathbb{E}\left[\tilde{Y}_i|D_i = 1, T_i = 1\right] - \mathbb{E}\left[Y_i|D_i = 1, T_i = 1\right] \tag{1}$$

The values of D_i and T_i indicate the treatment to which a student is actually exposed: $D_i = 1$ and $D_i = 0$ denote dormitory and non-dormitory students respectively, while $T_i = 1$ and $T_i = 0$ denote the tracking and mixing periods respectively. $\tilde{Y}_i$ denotes the outcome for student i if s/he were living in a dormitory under tracking (i.e. if s/he is exposed to the intervention of interest) and Y_i denotes the outcome for student i if s/he were either not living in a dormitory or living in a dormitory under mixing (i.e. if s/he is not exposed to the intervention). The first term in equation (1) is directly observed, while the second is identified only under suitable restrictions on the data generating process. Note that identification here requires restrictions only on the data generating process *in the absence of tracking* and allows us to remain agnostic as to the process generating the outcomes under tracking.[7]

[6] I have recently obtained a database of the high school each student attended and so can soon test this claim.

[7] This is an advantage of restricting attention to the average treatment effect on the treated. Identifying the average treatment effect requires knowledge of two counterfactuals: the outcome that would have prevailed in the absence of tracking and the outcome that would have prevailed if tracking were applied in both periods. Hence, a model of the outcomes under

The standard difference-in-differences strategy imposes the identifying assumptions that in the absence of tracking:

(A1) $Y_i^{DT} = h(D_i, T_i, \epsilon_i) = \alpha + \beta D_i + \gamma T_i + \epsilon_i$

(A2) $\mathbb{E}[\epsilon_i|D = 1, T = 1] - \mathbb{E}[\epsilon_i|D = 1, T = 0] - \mathbb{E}[\epsilon_i|D = 0, T = 1] - \mathbb{E}[\epsilon_i|D = 0, T = 0]$

where ϵ_i^{DT} is a scalar capturing the unobserved characteristics of student i. The first assumption requires that the outcomes be generated by a single index function that is additively separable in D, T, and ϵ. The second assumption requires that the change in the mean of the unobserved characteristics from the tracking to the mixing period be identical for dormitory and non-dormitory students if no change had taken place in the dormitory allocation regime. Note that these assumptions will not jointly hold for multiple monotonic but non-affine rescalings of the outcomes; for example, the assumptions cannot be true for an outcome measured both in levels and logs. Also note that these assumptions require that the structure of the data-generating process be identical for dormitory students under mixing and for non-dormitory students during both periods. This is a substantive restriction but is unavoidable in all difference-in-differences analyses, which are based on assuming some common structure for the outcomes in the non-intervention group (non-dormitory students) and in the intervention group in the absence of the intervention (dormitory students under mixing).

Under assumptions (A1) and (A2),

$$
\begin{aligned}
&\mathbb{E}\left[Y_i|D_i = 1, T_i = 0\right] + \mathbb{E}\left[Y_i|D_i = 0, T_i = 1\right] - \mathbb{E}\left[Y_i|D_i = 0, T_i = 0\right] \\
&\quad = \alpha + \beta + \mathbb{E}\left[\epsilon_i|D_i = 1, T_i = 0\right] + \alpha + \gamma + \mathbb{E}\left[\epsilon_i|D_i = 0, T_i = 1\right] - \alpha - \mathbb{E}\left[\epsilon_i|D_i = 0, T_i = 0\right] \\
&\quad = \alpha + \beta + \gamma + \mathbb{E}\left[\epsilon_i|D_i = 1, T_i = 0\right] + \mathbb{E}\left[\epsilon_i|D_i = 0, T_i = 1\right] - \mathbb{E}\left[\epsilon_i|D_i = 0, T_i = 0\right] \\
&\quad = \mathbb{E}\left[Y_i|D_i = 1, T_i = 1\right].
\end{aligned}
$$

This yields the familiar difference-in-differences result

$$
\begin{aligned}
\Delta^{ATT} &= \mathbb{E}\left[\tilde{Y}_i|D_i = 1, T_i = 1\right] - \mathbb{E}\left[Y_i|D_i = 1, T_i = 0\right] \\
&\quad - \mathbb{E}\left[Y_i|D_i = 0, T_i = 1\right] + \mathbb{E}\left[Y_i|D_i = 0, T_i = 0\right] \qquad (2)
\end{aligned}
$$

tracking is required to identify the average treatment effect. See Heckman and Robb (1985) for a more general discussion of the difference between these two parameters and the nature of the restrictions required to identify them.

which can be consistently estimated by sample analogues

$$\begin{aligned}\hat{\Delta}^{ATT} &= \frac{1}{N_{11}}\sum_i D_i T_i Y_i - \frac{1}{N_{10}}\sum_i D_i(1-T_i)Y_i \\ &-\frac{1}{N_{01}}\sum_i (1-D_i)T_i Y_i + \frac{1}{N_{00}}\sum_i (1-D_i)(1-T_i)Y_i \end{aligned} \tag{3}$$

where N_{DT} is the number of observations in group DT.

Alternatively, assumptions (A1) and (A2) can be relaxed in favor of the conditional assumptions:

(A3) $Y_i^{DT} = h(D_i, T_i, X_i, \epsilon_i) = \alpha + \beta D_i + \gamma T_i + X_i'\delta + \epsilon_i$

(A4) $\mathbb{E}[\epsilon_i|X_i, D_i = 1, T_i = 1] - \mathbb{E}[\epsilon_i|X_i, D_i = 1, T_i = 0] - \mathbb{E}[\epsilon_i|X_i, D_i = 0, T_i = 1] - \mathbb{E}[\epsilon_i|X_i, D_i = 0, T_i = 0]$

where X is a vector of observed student characteristics. Note that these conditions require merely common trends in the unobserved characteristics conditional on each possible realization of the observed characteristics, rather than unconditional common trends. A model may be conditionally identified but not unconditionally identified. Under these assumptions, estimating the regression model

$$Y_i = \mu_{11} D_i T_i + \mu_{10} D_i(1-T_i) + \mu_{01}(1-D_i)T_i + \mu_{00}(1-D_i)(1-T_i) + X_i'\delta + \epsilon_i.$$

generates a consistent estimator of the conditional average treatment effect on the treated:

$$\hat{\Delta}^{ATT} = \hat{\mu}_{11} - \hat{\mu}_{10} - \hat{\mu}_{01} + \hat{\mu}_{00}. \tag{4}$$

Abadie (2005) notes that neither theory nor institutional knowledge typically provide a justification for the additive separability between X_i and ϵ_i or the linear form $X_i\delta$. He shows that even if (A4) is relaxed to allow a more general and unknown function $f(X_i, \epsilon_i)$, the conditional difference-in-differences estimator is nonparametrically identified. Abadie considers both fully nonparametric estimation of the four conditional expectations in equation (2) and, for cases where the dimension of X is large, a two step semiparametric reweighting scheme. The first step estimates the propensity score, or probability of being in a particular group of students conditional on observed characteristics, using a semiparametric logistic model. The second step reweights each observation by a function of the propensity score to assign more weight to students whose observed characteristics are similar to those of students in the intervention group (in this case, students in dormitories under the tracking regime). Abadie's approach is focused on difference-in-difference models using panel data but his method can be generalized to allow for repeated cross-sectional data using the formula:[8]

$$\Delta^{ATT} = \mathbb{E}\Big[Y \times \frac{P\left(D=1,T=1|X\right)}{P\left(D=1,T=1\right)} \times \\ \left(\frac{DT}{P(D=1,T=1|X)} - \frac{D(1-T)}{P(D=1,T=0|X)} - \frac{(1-D)T}{P(D=0,T=1|X)} + \frac{(1-D)(1-T)}{P(D=0,T=0|X)}\right)\Big] \tag{5}$$

The indicators determine which weight function to assign to each observation, while the weights, informally, equalize the distribution of the observed characteristics across all four groups. Unlike Abadie's reweighting scheme, I need to estimate the conditional probabilities of four possible combinations of D and T. I do so by rewriting $P\left(D=1,T=1|X\right) = P\left(T=1|X,D=1\right) \times P\left(D=1|X\right)$ and estimating each term separately.[9] The resultant estimator is

$$\hat{\Delta}^{ATT} = \frac{1}{N_{11}}\sum_i Y_i - \frac{1}{N_{10}}\sum_i Y_i \frac{\hat{P}\left(T=1|X,D=1\right)}{\hat{P}\left(T=0|X,D=1\right)} - \frac{1}{N_{01}}\sum_i Y_i \frac{\hat{P}\left(T=1|X,D=1\right)\hat{P}\left(D=1|X\right)}{\hat{P}\left(T=1|X,D=0\right)\hat{P}\left(D=0|X\right)} \\ + \frac{1}{N_{00}}\sum_i Y_i \frac{\hat{P}\left(T=1|X,D=1\right)\hat{P}\left(D=1|X\right)}{\hat{P}\left(T=0|X,D=0\right)\hat{P}\left(D=0|X\right)}. \tag{6}$$

Reweighting estimators are now widely used in the program evaluation literature but a consensus does not appear to have developed on the best practice for implementing the semiparametric logistic model in the first stage of the estimators. Most theoretical papers follow Hirano, Imbens, and Ridder (2003) in recommending a power series estimator in which an indicator variable for intervention status is regressed on a polynomial in each observed covariate and a full set of interactions. Few papers discuss criteria other than cross-validation for choosing the tuning parameter (the order of the polynomial) and the literature on this choice is considerably less well developed than for the choice of the tuning parameter in density weighted estimation (the bandwidth). I follow Dehejia and Wahba (1999) in choosing the lowest order of the polynomial that balances the distribution of covariates across groups and then testing the robustness of the resultant inferences to higher and lower order polynomials.

Table 3 reports estimates of the average treatment effect of tracking on the treated for estimators (3), (4), and (6). I use first year grade point average (GPA) as the primary outcome of interest, which in this university ranges between 0 and 100 with a mean and standard deviation of 59 and 12 amongst

[8] Throughout this section I assume the existence of the relevant expectations and probability limits.

[9] For further background on the use of the propensity score in program evaluation, see Rosenbaum and Rubin (1983), Heckman, Ichimura, and Todd (1997), and Hirano, Imbens, and Ridder (2003). For an application to wage decompositions, see DiNardo, Fortin, and Lemiuex (1996). Note that an alternative to breaking each of the conditional probabilities into two terms is to use the generalized propensity score for comparisons across more than two groups (Imbens, 2000).

non-dormitory students. Each successive row of the table uses a more flexible polynomial in the demographic and academic controls, beginning with a linear model with no interactions and concluding with a cubic model with three-way interactions. The regression-adjusted estimates consistently show an average treatment effect on the treated of -1.2 to -1.3 points of GPA, which is equal to approximately 0.1 standard deviations of the outcome in the control group. These effects are precisely estimated, highly robust to the specification of the control vector, and almost entirely unaffected by the inclusion of dormitory fixed effects. The reweighted estimates are larger but slighly less precisely estimated and less robust across specifications. They suggest an average treatment effect on the treated of approximately 0.25 standard deviations, which falls to 0.2 standard deviations after controling for dormitory fixed effects. The robustness of these results suggests that neither individual covariates nor dormitory fixed effects are driving the estimated treatment effects. I interpret this as evidence that there is limited selection on student observed characteristics and that time-invariant physical and institutional features of the dormitories play little role in determining student outcomes. This in turn suggests that the differences between high and low tracks that are potentially conflated with peer effects in studies of classroom- or school-level tracking are not an important consideration in this context.

To understand the differences between the regression-adjusted and reweighted estimates, a brief discussion of the standard error estimators is required. The standard errors in this table, and all those reported in the remainder of the paper, are constructed using a mixed bootstrap algorithm. I resample dormitory-year clusters from the group of dormitory students and individual students from the non-dormitory group, stratifying by year in each case. I estimate the treatment effect for the resampled data and use the standard deviation of the 500 estimated effects as the standard error of the estimator. This estimator does not offer asymptotic refinement (i.e. it does not converge faster than the standard parametric rate - see Horowitz (2001)) but it does avoid the need to derive analytic expressions for the standard errors of the various estimators. These derivations are simple for the linear estimators discussed in this section but are extremely complex for the nonlinear estimators discussed in sections 6 and 7. It is particularly challenging to derive the formulae under the possibility of serial correlation across students in each dormitory.[10]

The large standard errors on the reweighted estimates with higher-order polynomials and dormitory fixed effects may occur in part because the logistic regression used to estimate the propensity scores in

[10] Cameron, Gelbach, and Miller (2008) suggest that when the number of clusters is relatively small, a cluster bootstrap such as this has less accurate coverage than a wild bootstrap estimator that resamples clusters while imposing a particular null hypothesis. However, the moderate number of clusters, 58, in this application mitigate their concerns about the finite sample performance of the cluster bootstrap. Furthermore, their approach only generates confidence intervals with improved coverage for a specific null hypothesis, rather than standard errors that can be used to test multiple hypothesis and Kline and Santos (2011) show that the superior Monte Carlo performance of the wild bootstrap is specific to a narrow class of data generating processes.

the reweighting function has convergence problems in some of the resampled datasets. In particular, some interaction terms are highly colinear with some dormitory indicators in some resamplings, so the likelihood function is relatively flat and the logistic regression converges only with relatively unrestrictive convergence thresholds. This creates a significant number of outliers in the bootstrap distribution of estimated treatment effects which in turn inflates the bootstrap standard error. I interpret these large standard errors as reflecting “real” variability in the reweighting estimators, rather than a problem with the resampling algorithm and so do not trim the bootstrap distribution before estimating the standard errors. Implementing a trimming algorithm reduces the standard errors of the reweighted estimates by up to 25%.

The results in table 3 suggest that inference on the average treatment effect is very robust to different specifications of the control vector and relatively robust to the choice of a regression-adjusted or reweighted estimator. This robustness remains true for the estimators employed in the remainder of the paper, so I report results only for the models with quadratic controls and pairwise interactions and cubic controls and three-way interactions (the fourth and sixth rows in table 3 respectively). The former is the sparsest specification that equalizes the mean change through time for dormitory students with that for non-dormitory students for all covariates and therefore satisfies the balance on observed characteristics criterion suggested by Dehejia and Wahba (1999). I report the latter specification as a demonstration of robustness.

It is worth noting that these effects are unusually large in the peer effects literature. They are not directly comparable to most existing studies of peer effects in higher education, which are based on across-dormitory or across-roommate variation under a given random assignment regime. However, most of these studies find relatively small effects of peer characteristics. Sacerdote (2001), Zimmerman (2003) and Stinebrickner and Stinebrickner (2006) find small effects for some students at Williams, Dartmouth and Berea universities respectively, while Foster (2006) and Han and Li (2009) find no evidence of peer effects at the University of Maryland and an unnamed Chinese university respectively. Only Carrell, Fullerton, and West (2009) find large peer effects, perhaps reflecting the fact that the groups they examine not only live together but also study and take classes together. Studies that compare high- and low-track students under a tracked assignment regime typically find larger effects but these may reflect variation in both peer quality and in school resources (Ding and Lehrer, 2007; Jackson, 2010; Pop-Eleches and Urquiola, 2011).

5 Observed heterogeneity

The linear difference-in-differences strategy used in section 4 does not impose any assumptions about the distribution of treatment effects and so allows any form of heterogeneity. However, the average treatment effect on the treated that it estimates averages across this heterogeneity rather than showing how treatment effects vary across students. Estimators (3), (4), and (6) can all be adapted to explore heterogeneity along observed dimensions. In particular, I estimate (4) separately for different groups of students defined by demographic characteristics and high school graduation test scores. This estimates the average treatment on the treated for members of each group and permits a test of whether these effects differ across groups. Note that this does not provide any information about heterogeneity with respect to unobserved student characteristics, a topic to which I turn in section 6.

Table 4 reports the average treatment effects on the treated separately for each gender, for each race group and for students in each tercile of the distribution of high school graduation test scores. The effects are slightly larger for men than women (0.12 standard deviations of the outcome relative to 0.09 standard deviations) but this difference is never significant. Black students are the worst affected by a large margin, with their grades more than 0.2 standard deviations lower under tracking than mixing. White students are also hurt by tracking but by a considerably smaller margin. This is consistent with the fact that the negative treatment effect is concentrated amongst students relatively low in the distribution of high school grades, who are disproportionately likely to be black. Students in the lowest tercile obtain grades that are 0.18 standard deviations lower under tracking, while the effect on students in the highest tercile is less than half as large (though still negative and marginally significant). The positive effect on coloured students is surprising, though it should be interpreted with caution as only 5% of dormitory students are members of this race group. The results are, once again, very robust to the specification of the control vector and the inclusion of dormitory fixed effects.

These results suggest, perhaps unsurprisingly, that the negative effects of tracking are largest for students with demographic and academic characteristics that would likely place them in the lower portion of the grade distribution. However, effects are negative even for students who obtained relatively high grades in high school graduation tests, the population who seem *ex ante* most likely to benefit from tracking. To explore this point further, I estimate a local linear regression of university GPA on high school graduation test scores separately for dormitory and non-dormitory students in each of the two periods. I then construct the difference-in-difference estimator,

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

where $\hat{\mu}_{DT}(X)$ is the local linear regression from group (D, T) evaluated at a score of X on the high school graduation test. This estimator is identified if the trend between the tracking and mixing periods in unobserved characteristics for students at every score on high school graduation test was identical for dormitory and non-dormitory students. The resultant estimates are shown in figure 4, with a 95% pointwise confidence interval.[11] The estimated effects are imprecisely estimated, particularly in the lower tail, but suggest that the effect of tracking is negative for all but the very highest achieving students. Only 14% of students have high school graduation test scores for which the estimated treatment effect is positive.

6 Unobserved heterogeneity

Estimating average effects for selected subgroups of students provides some information about the heterogeneity of the effects of tracking. However, it provides no information about heterogeneity on unobserved student characteristics and hence cannot speak to full distribution of outcomes in the absence of tracking. In this section I therefore turn to the nonlinear difference-in-differences strategy proposed by Athey and Imbens (2006) that identifies the full counterfactual distribution of outcomes if no change had taken place in the dormitory allocation regime. This stronger result identifies heterogeneity on both observed and unobserved student characteristics. However, it does so at the cost of a stronger assumption: that there would have been no change in the distribution of unobserved characteristics for either dormitory or non-dormitory students if no change had taken place in the dormitory allocation regime.

More formally, the Athey-Imbens strategy recovers the full set of quantile treatment effects on the treated (QTT), defined by

$$\tau^{QTT}(q) = F^{-1}_{\tilde{Y}^{11}}(q) - F^{CF-1}_{Y^{11}}(q) \tag{7}$$

where $F_{\tilde{Y}^{11}}$ is the observed distribution of university GPAs under the tracking treatment and $F^{CF}_{Y^{11}}$ is the counterfactual distribution of GPAs that the same group of students would have obtained in the absence of tracking. (Recall that I define $\tilde{Y}_i$ as an outcome for a dormitory student under tracking and Y_i as an outcome for a student in any other group. The inverse distribution functions are defined as $F^{-1}_Y(q) = \inf\ \{y \in \mathbb{Y} : F_Y(y) \geq q\}$. This strategy assumes that in the absence of tracking

(B1) $Y^{DT}_i = h(T_i, \epsilon^D_i)$ with $h()$ strictly increasing in the scalar unobserved characteristic ϵ_i for $T \in \{0, 1\}$

[11] The pointwise confidence intervals are constructed by taking the difference between percentiles 2.5 and 97.5 of the distribution of treatment effects from 500 iterations of the mixed pairs/cluster bootstrap described in the previous section. The percentiles are computed separately for each score on the high school graduation test, so the confidence intervals should be expected to have the correct coverage only pointwise and not uniformly.

(B2) $\epsilon_i \perp T_i | D_i$

(B3) $\varepsilon^1 \subseteq \varepsilon^0$, where ε^D is the support of ϵ_i^D for $D_i \in \{0,1\}$.

To build intuition for their strategy, consider two students from the mixing period with the same outcome Y_i, one from the dormitory group and one from the non-dormitory group. The assumption that $h()$ is invertible and common to the two groups means that they must have the same value of the scalar unobserved characteristic $\epsilon_i = h^{-1}(Y_i; T_i = 0)$. The assumption that $h()$ is common to non-dormitory students in both periods means that the non-dormitory student would have the outcome $h\left(h^{-1}\left(Y_i; T_i = 0\right), T_i = 1\right)$ in the tracking period. Comparing this outcome to Y_i reveals the counterfactual outcome that the dormitory student would have experienced during the tracking period in the absence of tracking. Replicating this analysis for all values of the outcome amongst dormitory students in the mixing period generates the full counterfactual distribution of outcomes, as assumption (B2) rules out changes in the distribution of the unobserved scalar through time. Finally, assumption (B3) requires that there is an appropriate non-dormitory student under mixing to which every dormitory student under mixing can be compared. If this final assumption is violated, then the model is identified only for those values of ϵ_i^1 that are contained in ε_i^0. Under these identifying assumptions, the counterfactual distribution is given by

$$F_{Y^{11}}^{CF}(y) = F_{Y^{10}}\left(F_{Y^{00}}^{-1}\left(F_{Y^{01}}\left(y\right)\right)\right) \tag{8}$$

so the treatment effect can be consistently estimated by sample analogues:

$$\begin{aligned} \hat{\tau}^{QTT}(q) &= \hat{F}_{\tilde{Y}^{11}}^{-1}(q) - \hat{F}_{Y^{11}}^{CF-1}(q) \\ &= \min\left\{y : \hat{F}_{\tilde{Y}^{11}} \geq q\right\} - \min\left\{y : \hat{F}_{Y^{10}}\left(\hat{F}_{Y^{00}}^{-1}\left(\hat{F}_{Y^{01}}\left(y\right)\right)\right) \geq q\right\} \end{aligned} \tag{9}$$

where the cumulative distribution functions are estimated by the empirical distribution, so $\hat{F}_{Y^{DT}}(y) = N_{DT}^{-1} \sum_i D_i T_i \mathbf{1}\{Y_i \leq y\}$.

There are two salient distinctions between this model and the standard difference-in-differences strategy. First, the Athey-Imbens model relaxes assumptions (A1) and (A3) that outcomes are a linear function of group membership, time, and unobserved characteristics, all of whose effects are additively separable. Relaxing this restriction allows for a richer class of data generating processes and in particular removes the sensitivity to monotonic but non-affine transformations of the outcomes. Second, the model now requires that there be no changes in the distribution of the unobserved characteristic through time, which is strictly stronger than the assumption of parallel mean time trends in the standard model. Note that both the standard and Athey-Imbens models impose the restriction that the structure of the data generating process is the same for non-dormitory students and dormitory students in the mixing period. This is again restrictive but unavoidable.

The key advantage of this second strategy is that it generates the full counterfactual distribution of grades that would have prevailed in the absence of tracking, or equivalently, the full set of quantile treatment effects on the treated. Note that these should be interpreted as treatment effects for each quantile of the grade distribution, rather than for any particular student. To identify treatment effects on individual students, we require the stronger assumption that all tracked students would have the same rank in the grade distribution in the absence of tracking. This assumption seems particularly implausible in this context, as the nature of the tracking treatment is different for students with different high school graduation test scores. Heckman, Smith, and Clements (1997) provide a more detailed discussion of this point.

The identifying assumptions for the nonlinear difference-in-difference model can be relaxed to hold conditional on observed student characteristics. Athey and Imbens (2006) suggest either applying their model within cells defined by particular values of the covariates or regressing the outcome on a flexible function of the covariates

$$Y_i = \mu_{11} D_i T_i + \mu_{10} D_i (1 - T_i) + \mu_{01} (1 - D_i) T_i + \mu_{00} (1 - D_i)(1 - T_i) + X_i' \delta + \epsilon_i$$

and applying their model to the group-specific estimated residuals

$$\hat{\epsilon}_i^{DT} = \hat{Y}_i - \hat{\mu}_{DT}.$$

As a middle ground, I propose to reweight students from the mixing period so that they have the same distribution of observed characteristics as students in the mixing period. I implement this procedure separately for dormitory and non-dormitory students, as this strategy does not require that the distribution of observed or unobserved characteristics is the same for dormitory and non-dormitory students. This procedure closely follows Firpo (2007) and his paper provides a more comprehensive discussion of the assumptions necessary for identification and estimation. I replace assumptions (B1) - (B3) with

(B4) $Y_i^{DT} = h(T_i, X_i, \epsilon_i^D)$ with $h()$ strictly increasing in the scalar unobserved characteristic ϵ_i for $T \in \{0, 1\}$ and for all $X_i \in \mathbb{X}$

(B5) $\epsilon_i \perp T_i | D_i, X_i$

(B6) $\varepsilon^{D,X} \subseteq \varepsilon^{D,X}$, where $\varepsilon^{D,X}$ is the support of ϵ_i for $D_i \in \{0, 1\}$ and for all $X_i \in \mathbb{X}$

(B7) $Pr\left(T_i = 1|D_i, X_i\right) < 1$ for $D_i \in \{0, 1\}$.

The first three assumptions are simply relaxations of the identifying assumptions (B1) - (B3). The fourth assumption requires the propensity score for the tracking peroid is less than one, or that there are no values of (D, X) that exist only in the tracking period and not the mixing period. This is a version of the common support condition widely used in the program evaluation literature that requires some degree of similarity on observed characteristics between the two groups being compared.[12] Under assumptions (B4) - (B7), the counterfactual distribution is given by

$$F^{CF}_{Y^{11}}(y) = F_{Y^{10}_{\omega}}\left(F^{-1}_{Y^{00}_{\omega}}\left(F_{Y^{01}}\left(y\right)\right)\right) \tag{10}$$

where $F_{Y^{D0}_{\omega}}(y)$ denotes the distribution function for grades in group D in the mixing period, reweighted to have the same distribution of observed characteristics as in group D in the tracking period. Under appropriate regularity conditions, this distribution can be consistently estimated by constructing

$$\begin{aligned}\hat{F}_{Y^{10}_{\omega}}(y) &= \frac{1}{N_{10}}\sum_i D_i(1 - T_i)\hat{\omega}_1(X_i)\mathbf{1}\{Y_i < y\}\\ \hat{F}_{Y^{00}_{\omega}}(y) &= \frac{1}{N_{00}}\sum_i D_i(1 - T_i)\hat{\omega}_0(X_i)\mathbf{1}\{Y_i < y\}\end{aligned}$$

where the reweighting term

$$\hat{\omega}_d(X_i) = \frac{\hat{P}\left(T_i = 1|X_i, D_i = d\right)}{1 - \hat{P}\left(T_i = 1|X_i, D_i = d\right)}$$

is estimated by a flexible logistic regression model. Hence, the quantile treatment effect on the treated is estimated by

$$\hat{\tau}^{QTT}_{\omega}(q) = \min\left\{y : \hat{F}_{\tilde{Y}^{11}} \geq q\right\} - \min\left\{y : \hat{F}_{Y^{10}_{\omega}}\left(\hat{F}^{-1}_{Y^{00}_{\omega}}\left(\hat{F}_{Y^{01}}\left(y\right)\right)\right) \geq q\right\} \tag{11}$$

for each quantile q. Figure 5 plot the distribution of propensity scores for dormitory and non-dormitory students and clearly indicate that these are bounded away from zero and one with most of the probability mass in the neighborhood of 0.5. This means that there is unlikely to be a problem with very large weights that introduce instability into the estimators and blow up their variance.

Figures 6 and 7 show the full set of quantile treatment effects on the treated for a variety of specifications. The left-hand graph in panel A of figure 6 shows the the distribution of university GPA for the

[12] Note that identification of treatment on the treated effects does not require that $Pr\left(T_i = 1|D_i, X_i\right) > 0$ for $D_i \in \{0, 1\}$, so I allow for the possibility of values of (D, X) found only in the mixing period. Students with these values are simply assigned a weight of zero. This additional restriction would be required to identify average treatment or treatment on the untreated effects.

treated group (dormitory students under tracking) and the counterfactual distribution that would have prevailed if that group were assigned to dormitories using mixing instead of tracking. The counterfactual distribution is constructed using estimator (9), not controling for observed student characteristics and so making no correction for changes in such characteristics through time. The quantile treatment effects on the treated (QTT) are defined as the horizontal distance between the two distribution functions at each quantile and are shown in the right-hand graph in panel A.[13] These effects in the lower tail of the distribution are very large: approximately -0.5 and -0.25 standard deviations at the 5^{th} and 10^{th}percentiles respectively. They are negative for all but the highest quantiles but are considerably smaller away from the bottom of the distribution, falling to -0.05 standard deviations at the median. These effects are, however, relatively imprecisely estimated and are significantly different from zero only near the 5^{th} percentile.[14]

Panels B and C of figure 6 use the reweighted estimator (11) to construct the counterfactual distribution, adding controls for observed student characteristics and dormitory fixed effects respectively. The estimates are relatively robust to adding individual controls, though the treatment effects are now slightly larger at the lower quantiles: 0.6, 0.3, and 0.15 standard deviations at the 5^{th}, 10^{th}, and 25^{th} percentiles respectively. As in sections 4 and 5, the primary effect of controling for observed student characteristics is to increase the precision of the QTT effects. The confidence intervals for the reweighted estimator are considerably narrower and the effects are significantly different to zero for all percentiles between 2 and 34. Adding dormitory fixed effects reduces the precision of the estimates slightly but has no effect on the point estimates.

The reweighted estimates in figure 6 were generated using a logistic regression with the second-order polynomial (quadratic functions of the controls and pairwise interactions) recommended by the Dehejia and Wahba (1999) balance criterion. To test the robustness of these results, I repeat the same analysis using cubic functions of the covariates and three-way interactions. These estimates are shown in figure 7 and are indistinguisable. Figure 8 shows the treatment effects using regression-adjusted grades for quadratic and cubic models with and without dormitory fixed effects. These results are consistent with those from the reweighting estimators, though the estimated treatment effects are slightly smaller in the lower tail and appear to be quite unstable in the upper tail.

Having constructed the entire counterfactual distribution of outcomes in the absence of tracking, I can

[13] The vertical distance between the distribution functions does not correspond to a standard treatment effect but can still be of interest in other economic applications. See Fortin, Lemiuex, and Firpo (2011) for a more detailed discussion on this point.

[14] I again estimate confidence intervals using the percentiles of the distribution of treatment effects estimated over 500 replications of the mixed cluster/pair bootstrap described above. The conditions under which the bootstrap is valid for the Athey-Imbens model have not been formally established. However, their simulation results suggest that bootstrap confidence intervals have better coverage properties than those based on analytic standard error formulae.

estimate the means of the treated and counterfactual outcomes and so estimate the average treatment effect on the treated. As the identifying assumptions required for the linear and nonlinear difference-in-differences models are slightly different, checking whether they yield comparable estimates provides an informal test of the sensitivity of the estimates. The first line of table 5 reports estimates of the mean for the treated and counterfactual grades and the ATT from the nonlinear model. The treatment effect ranges between 0.09 and 0.11 standard deviations and is robust to different combinations of individual and dormitory controls, which is entirely consistent with the estimates derived from the linear model in section 4.

7 Inequality

Knowledge of the counterfactual distribution of grades in the absence of tracking also allows me to explore inequality treatment effects on the treated, or the change in the level of inequality amongst the tracked students caused by the tracking treatment. In particular, I estimate the interquartile range (75^{th} percentile - 25^{th} percentile), interdecile range (90^{th} percentile - 10^{th} percentile), coefficient of variation (standard deviation divided by the mean) and Gini coefficient for the treated grades and for the counterfactual grades. These are all standard measures of outcome inequality and Firpo (2010) shows that under weak assumptions, these measures are identified, can be consistently estimated by sample analogues and have sampling distributions that can be approximated by standard bootstrap algorithms.[15] I define the inequality treatment effect on the treated (ITT) as the difference between observed inequality in the treatment group (dormitory students under tracking) and inequality for the counterfactual distribution of grades that would have prevailed in the absence of tracking.

Table 5 reports the estimated inequality levels for the observed distribution of grades under treatment (column 1) and for the counterfactual distribution of grades in the absence of treatment without controls (column 2), with a quadratic control specification with and without fixed effects (columns 4 and 6 respectively), and with a cubic control specification with and without fixed effects (columns 8 and 10 respectively). Next to each inequality estimate for the counterfactual distribution I show the inequality treatment effect on the treated.

The interdecile range and interquartile range are both increased by tracking, with the former measure rising by 2.6 to 3.4 points of GPA (0.22 to 0.28 standard deviations) and the latter measure by 0.9 to 1.2

[15] In particular, he requires that the mean grade is non-zero for identification of the coefficient of variation and the Gini coefficient and that grades take on only positive values for identification of the Gini coefficient. Additional restrictions are required on the upper tail of the distribution of grades for the Gini coefficient to be Hadamard differentiable and hence for the validity of bootstrapped standard errors. These are automatically satisfied in this application as grades are bounded above by 100.

points of GPA (.08 to .1 standard deviations). These effects are very robust to different specifications, though only the former effect is consistently significantly different to zero. It is difficult to obtain a benchmark for the substantive magnitude of these effects, as very few papers analyze the inequality effects of education interventions. One way to understand their magnitude is to consider that tracking increases the interdecile range by more than 10% in each specifications and increase the interquartile range by 7 to 10%. These are clearly large increases in inequality, particularly when recalling that the 75^{th} and 90^{th} percentiles are actually marginally lower under tracking than they otherwise would be. This provides a sharp illustration of how damaging tracking is to the lower tail of the GPA distribution.

The coefficient of variation, a normalized measure of the spread of GPA, is also consistently higher for the tracked students than would be the case in the absence of tracking. The coefficient increases by 0.005 and although this effect is statistically significant, it amounts to an increase of less than 1% in the counterfactual coefficient of variation. In contrast, the Gini coefficient increases from approximately 0.09 in the absence of tracking to 0.104 with tracking, an increase of 15%. To understand the magnitude of this effect, note that the Gini coefficient equals the expected difference between the grades of any two students selected randomly from the sample, divided by twice the sample mean (Deaton, 1997). Increasing this difference by 15% is clearly a huge increase in inequality and this emphasizes the inequality enhancing effects of tracking.

8 Conclusion

This study provides evidence that the tracked dormitory assignment regime had large damaging effects on the distribution of achievement at this South African university. Average GPA was approximately 0.1 standard deviations lower under tracking, with GPA falling by up to 0.5 standard deviations in the lower tail. The concentration of the treatment effect in the lower tail increased measured inequality by up to 15%, demonstrating that tracking was a highly regressive intervention. The research design does not allow me to identify the treatment effects on individual students. However, combining the evidence from sections 5 and 6 suggests that not only was the lower tail disproportionately affected by tracking but also that students whose demographic and academic characteristics made them likely to be in the lower tail were the ones most affected by the tracked assignment regime.

This pattern of results provides some information about the nature of the GPA production process. Consider an analogy to the canonical marriage market model, in which market participants form matches with partners of varying “quality” in order to maximize the surplus they obtain from the marriage (Becker, 1991; Becker and Murphy, 2000). The tracking regime can be considered as an intervention that imposes some cost to forming heterogeneous matches, perhaps because it is more difficult to identify desirable matches outside one's own dormitory. Becker (1991) shows that if the cross-partial derivative of match output with respect to own and partner quality is positive, then positive assortative matching will occur: all students will form matches with partners of similar baseline performance and the distribution of outcomes would be the same under tracking and mixing.[16] In constrast, a negative cross-partial derivative is sufficient for negative assortative matching, in which students maximize the joint output of the match by forming heterogeneous matches. Tracking would substantially disrupt this matching process by making the desired heterogeneous matches difficult to create.

The large negative effect on the lower tail of the distribution suggests that negative assortative matching would occur in the absence of tracking. However, Becker and Murphy (2000) note that negative assortative matching also requires that the surplus from the match be transferable. These transfers allow students with low baseline performance to "bribe" students with high baseline performance into forming matches through surplus transfers. Otherwise, students with high baseline performance can maximize their individual gains by forming homogeneous matches, even though these produce a smaller joint surplus. In the context of a marriage market, where matches are long-term and include negotiation over multiple dimensions, such transfers are typically regarded as plausible. In the current context, where the gains from matches are in terms of higher GPA, direct transfers are clearly not possible. However, the observed results can be rationalized if there is scope for compensating transfers along some other dimension from students with low baseline performance to their higher performance partners. Such transfers might include the purchase of higher social status or assistance with routine domestic chores such as laundry. Under such a model, the fact that the upper tail is largely unaffected by tracking can be explained by substitution. Students with high baseline performance spend more time on academic work in the mixing regime than the tracking regime as they assist their lower performing partners but they receive non-academic transfers in exchange for doing so. Taken literally, the model suggests that their utility is lower under tracking but that their losses are on non-academic dimensions that are not observed in my data.

Hence, a matching model in which students form heterogeneous matches in order to maximize the joint surplus from the match and low ability students transfer this surplus to their high ability partners through non-academic dimensions. If the cost of forming matches across dormitories is small relative to the gains from negative assortative matches, then outcomes should be unaffected by the dormitory assignment

[16] More formally, the distributions will be identical if there is a continuum of students placed in a countable number of dormitories by a tracking regime that perfectly partitions the distribution of baseline performance. If any of these conditions are violated, either tracking or mixing may disrupt the perfect positive assortative matching that would occur without division into dormitories.

regime. The magnitudes of the treatment effects and the fact that students do have opportunities to meet outside their dormitories suggest that the gains from negative assortative matching must be large to generate the observed treatment effects.

The magnitude and direction of the treatment effects that I estimate are perhaps surprising in light of the previous literature. Duflo, Dupas, and Kremer (2011) is the only previous study using across-regime variation, rather than across group variation within a given regime, and so has the most comparable research design. They find moderate positive effects of tracking for both high- and low-track students, which they attribute to the ability of teachers to target their instruction better in homogeneous classes. My results suggest that the pure peer effect of tracking is negative, which is consistent with findings from Ding and Lehrer (2007), Jackson (2010) and Pop-Eleches and Urquiola (2011) that high track students outperform low track students on a range of outcomes. Their designs, however, cannot directly separate pure peer effects from the effects of differences in resources across schools.

The direction of my results is broadly consistent with the findings from research designs based on random assignment to peer groups. Early studies by Sacerdote (2001) and Zimmerman (2003) found significant effects on student outcomes of roommates's prior academic performance. However, many papers in this literature have found relatively small effects that operate only through certain measures of baseline performance or ability, leading Foster (2006) to question the robustness of most research designs that identify peer groups using exogeneous variation in spatial proximity. A smaller set of papers that explore heterogeneous effects of assignment to different types of peers also finds that the tails matter more than the middle of the distribution (Lavy, Silva, and Weinhardt, 2009). These papers do not, however, explore heterogeneity across both observed and unobserved characteristics and so may miss important details that are captured by my research design.

The differences between my results and the prior literature emphasize the important insights available from studying peer effects using across-regime variation. I find peer effects that amongst the largest in the existing literature and I provide a comprehensive discussion of the heterogeneity of these effects. I also show that these results provide some insights into the nature of the GPA production process. Own and peer baseline performance appear to be substitutes and this motivates negative assortative matching amongst students, provided low performing students are able to transfer surplus to their high performing partners. An important direction for future research is to understand the conditions under which such transfers are feasible and take place, as this would provide some indication of the effects tracking might have in other settings.

References

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

Angrist, J., and K. Lang (2004): "Does school integration generate peer effects? Evidence from Boston's Metco program," 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," 74(2), 431–497.

Becker, G. (1991): *A treatise on the family.* Harvard University Press, Cambridge, MA.

Becker, G., and K. Murphy (2000): *Social economics: Market behavior in a social environment.* Harvard University Press, Cambridge, MA.

Bertrand, M., E. Luttmer, and S. Mullainathan (2000): "Network effects and welfare cultures," *Quarterly Journal of Economics*, pp. 1019–1055.

Bruhn, M., and D. McKenzie (2009): "In Pursuit of Balance: Randomization in Practice in Development Field Experiments," *American Economic Journal: Applied Economics*, 1(4), 200–232.

Cameron, C., J. Gelbach, and D. Miller (2008): "Bootstrap-based Improvement for Inference with Clustered Errors," *Review of Economics and Statistics*, 90(3), 414–427.

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

Carrell, S., B. Sacerdote, and J. West (2011): "From natural variation to optimal policy? The Lucas critique meets peer effects," Discussion Paper 16865, National Bureau of Economic Research.

Deaton, A. (1997): *The Analysis of Household Surveys.* Johns Hopkins University Press, Baltimore, MD.

Dehejia, R., and S. Wahba (1999): "Propensity score-matching methods for nonexperimental causal studies," *Review of Economics and Statistics*, 84(1), 151–161.

DiNardo, J., N. Fortin, and T. Lemiuex (1996): "Labor market institutions and the distribution of wages, 1973 - 1992: A semiparametric approach," 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," forthcoming.

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

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

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, 1455–1475.

Fowler, J., and N. Christakis (2008): "Estimating Peer Effects on Health in Social Networks," *Journal of Health Economics*, 27(5), 1400–1405.

Glewwe, P. (1997): "Estimating the impact of peer group effects on socioeconomic outcomes: Does the distribution of peer group characteristics matter?," *Economics of Education Review*, 16(1), 39–43.

Han, L., and T. Li (2009): "The Gender Difference of Peer Influence in Higher Education," *Economics of Education Review*, 28, 129–134.

Heckman, J., H. Ichimura, and P. Todd (1997): "Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme," *Review of Economic Studies*, 64(4), 605–654.

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

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

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

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

Imbens, G. (2000): “The role of the propensity score in estimating dose-response functions,” *Biometrika*, 87(3), 706–710.

Jackson, K. (2010): “Do students benefit from attending better schools? Evidence from rule-based student assignments in Trinidad and Tobago.,” *Economic Journal*, forthcoming.

Kline, P., and A. Santos (2011): “Higher order properties of the wild bootstrap under misspecification,” mimeo, University of California at Berkeley.

Lavy, V., O. Silva, and F. Weinhardt (2009): “The good, the bad and the average: Evidence on the scale and nature of ability peer effects in schools,” Discussion Paper 15600, National Bureau of Economic Research.

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

Mas, A., and E. Moretti (2009): “Peers at work,” *American Economic Review*, 99(1), 112–145.

Pop-Eleches, C., and M. Urquiola (2011): “The consequences of going to a better school,” mimeo, Columbia University.

Raudenbush, S., and T. Nomi (2011): “Using site-by-treatment to identify the joint effects of academic course-taking and classroom peers on student achievement,” mimeo, University of Chicago.

Rosenbaum, P., and D. Rubin (1983): “The central role of the propensity score in observational studies for causal effects,” *Biometrika*, 70, 41–55.

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

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.

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

Table 1: Baseline demographic characteristics

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Dormitory students			Non-dormitory students			
	Track	Mix	Δ_{dorm}	Track	Mix	Δ_{non}	$\Delta_{dorm} - \Delta_{non}$
Female	.5	.52	-.017	.52	.52	.008	-.025
			(.012)			(.012)	(.017)
Black	.49	.51	-.016	.11	.1	.009	-.025*
			(.01)			(.01)	(.014)
Coloured	.05	.05	-.002	.23	.28	-.053***	.051***
			(.008)			(.008)	(.012)
Indian	.08	.07	.007	.11	.09	.018*	-.011
			(.007)			(.007)	(.01)
White	.36	.36	.007	.53	.52	.02*	-.012
			(.012)			(.012)	(.017)
English	.59	.53	.056***	.86	.88	-.018*	.074***
			(.01)			(.01)	(.015)
Other language	.41	.47	-.056***	.14	.12	.018*	-.074***
			(.01)			(.01)	(.015)
South African	.78	.93	-.144***	.88	.97	-.085***	-.059***
			(.009)			(.009)	(.011)
Other nationality	.22	.07	.144***	.12	.03	.085***	.059***
			(.009)			(.009)	(.011)

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

Table 2: Baseline academic characteristics (high school grades).

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Dormitory students			Non-dormitory students			
	Track	Mix	Δ_{dorm}	Track	Mix	Δ_{non}	$\Delta_{dorm} - \Delta_{non}$
Mean	40.29	40.76	-.47***	39.58	39.67	-.09	-.38**
			(.14)			(.13)	(.19)
Std dev.	5.72	5.62	.1	5.43	5.67	-.24**	.34**
			(.11)			(.1)	(.15)
5th pctile	31	31	0	31	30	1	-1
			(.55)			(.66)	(.88)
10th pctile	33	33	0	33	32	1***	-1**
			(.24)			(.32)	(.41)
25th pctile	36	37	-1***	36	36	0	-1***
			(.27)			(.03)	(.27)
50th pctile	41	41	0	40	40	0	0
			(.5)			(.05)	(.5)
75th pctile	45	45	0	44	44	0	0
			(.3)			(.27)	(.39)
90th pctile	48	48	0	47	47	0	0
			(.45)			(.47)	(.64)
95th pctile	48	48	0	48	48	0	0
			(0)			(.06)	(.06)
Sample size	3156	3147		3569	3456		

Notes: Standard errors in parentheses estimated by stratified pairs bootstrap (1000 reps). ***, ** and * denote significance at 1%, 5% and 10% levels respectively.

Table 3: Linear difference-in-difference results

	(1)	(2)	(3)	(4)
No controls	-1.22			
	(.94)			
Linear controls	-1.16**	-1.15**	-1.8**	-1.78
& no interactions	(.56)	(.56)	(.88)	(1.08)
Linear controls	-1.25**	-1.25**	-2.9***	-2.37**
& pairwise interactions	(.53)	(.54)	(.79)	(1.12)
Quadratic controls	-1.21**	-1.21**	-2.9***	-2.36**
& pairwise interactions	(.48)	(.49)	(.78)	(1.11)
Quadratic controls	-1.29***	-1.3**	-2.89***	-2.29
& three-way interactions	(.47)	(.49)	(.82)	(2.31)
Cubic controls	-1.33***	-1.34***	-2.94***	-2.38
& three-way interactions	(.47)	(.48)	(.81)	(2.44)
Estimator	Regression	Regression	Reweighting	Reweighting
Individual-level controls	×	×	×	×
Dormitory indicators		×		×

Notes: Standard errors in parentheses estimated by mixed pairs/cluster bootstrap (500 reps).
***, ** and * denote significance at 1%, 5% and 10% levels respectively.
Control variables are high school grades, gender, race, language and nationality.
Dependent variable has mean 59 and standard deviation 12 in the control group.

Table 4: Linear difference-in-difference results for subgroups

	(1)	(2)	(3)	(4)	(5)
All	-1.22	-1.21**	-1.21**	-1.33***	-1.34***
	(.94)	(.48)	(.49)	(.47)	(.48)
Female	-1.25	-1.12*	-1.07*	-1.17*	-1.13*
	(1.15)	(.64)	(.6)	(.62)	(.59)
Male	-1.12	-1.38*	-1.37*	-1.52**	-1.51**
	(1.44)	(.75)	(.77)	(.75)	(.77)
Black	-3.69***	-2.61**	-2.58**	-2.65**	-2.62**
	(1.15)	(1.06)	(1.06)	(1.06)	(1.05)
Coloured	1.83	2.02	2.48*	1.8	1.97
	(1.75)	(1.35)	(1.41)	(1.31)	(1.42)
Indian	.77	-.02	.01	-.14	-.09
	(2.38)	(1.72)	(1.69)	(1.67)	(1.64)
White	-1.64	-1.58**	-1.76***	-1.66**	-1.83***
	(1.29)	(.64)	(.66)	(.63)	(.66)
Lowest tercile of high school grades	-2.09**	-2.16***	-2.13**	-2.22***	-2.17**
	(.82)	(.81)	(.84)	(.82)	(.85)
Middle tercile of high school grades	-1.2	-1.36*	-1.39*	-1.28*	-1.31*
	(.8)	(.73)	(.75)	(.74)	(.75)
Highest tercile of high school grades	-.43	-.92	-.95*	-.95*	-.98*
	(.87)	(.56)	(.56)	(.55)	(.55)
Quadratic controls & pairwise interactions		×	×		
Dormitory fixed effects			×		
Cubic controls & threeway interactions				×	×
Dormitory fixed effects					×

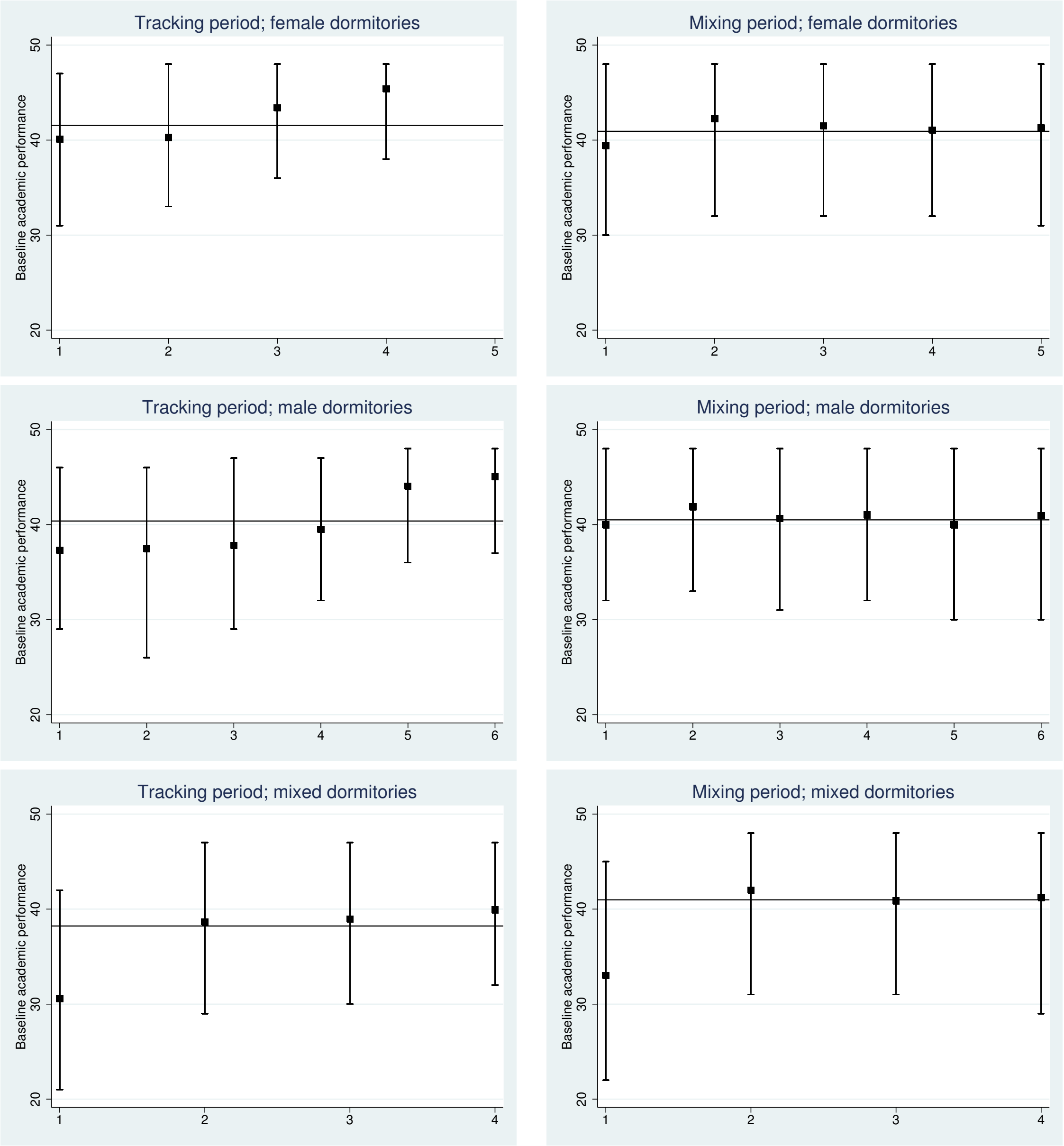
Notes: Standard errors in parentheses estimated by mixed pairs/cluster bootstrap (500 reps).
***, ** and * denote significance at 1%, 5% and 10% levels respectively.
Control variables are high school grades, gender, race, language and nationality.
Dependent variable has mean 59 and standard deviation 12 in the control group.

Table 5: Nonlinear difference-in-difference results: inequality and average treatment effects

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
	Treated	CF	$\Delta^{\text{Treatment}}$	CF	$\Delta^{\text{Treatment}}$	CF	$\Delta^{\text{Treatment}}$	CF	$\Delta^{\text{Treatment}}$	CF	$\Delta^{\text{Treatment}}$
Mean	59.95	61.03	-1.09	61.3	-1.35***	61.24	-1.3*	61.31	-1.36***	61.2	-1.25
	(.73)	(.56)	(.91)	(.61)	(.54)	(.74)	(.77)	(.66)	(.54)	(.87)	(.86)
Interdecile range	28.75	26.13	2.62*	25.38	3.38***	25.8	2.95**	25.47	3.28**	25.94	2.81*
	(.95)	(.98)	(1.36)	(.99)	(1.28)	(1.27)	(1.42)	(1.2)	(1.35)	(1.38)	(1.46)
Interquartile range	13.61	12.69	.92	12.41	1.19	12.44	1.17	12.42	1.19	12.4	1.21
	(.55)	(.48)	(.73)	(.53)	(.73)	(.55)	(.72)	(.62)	(.82)	(.99)	(1.12)
Gini coefficient	.104	.092	.012***	.089	.015***	.09	.013***	.089	.014***	.09	.013***
	(.003)	(.003)	(.004)	(.003)	(.004)	(.004)	.004	(.004)	(.004)	(.005)	(.005)
Coefficient of variation	1.018	1.014	.004***	1.013	.005***	1.013	.005***	1.013	.005***	1.013	.005**
	(.001)	(.001)	(.001)	(.001)	(.001)	(.001)	.001	(.001)	(.001)	(.002)	(.002)
Quadratic controls & pairwise interactions					×				×		
Cubic controls & three-way interactions							×				×
Dormitory fixed effects									×		×

Notes: Standard errors in parentheses estimated by mixed pairs/cluster bootstrap (500 reps).
***, ** and * denote significance at 1%, 5% and 10% levels respectively.
Control variables are high school grades, gender, race, language, and nationality.
Controls and dormitory fixed effects are applied by reweighting students in the mixing period to have the same distribution of covariates as in the tracking period.

Figure 1: Mean and $5^{th}/95^{th}$ percentiles of high school grades by dormitory under tracking and mixing



Notes: The first, second, and third rows show female-only, male-only and mixed gender dormitories respectively.
The fifth female-only dormitory was created after the tracking period ended.
Dormitories are listed from lowest to highest mean high school grade under tracking.
The same order is used for the two periods, so dormitory 1 in the mixing period had the lowest mean high school grade in the tracking period.

Figure 2: Mean and variance of high school grades by dormitory under tracking and mixing

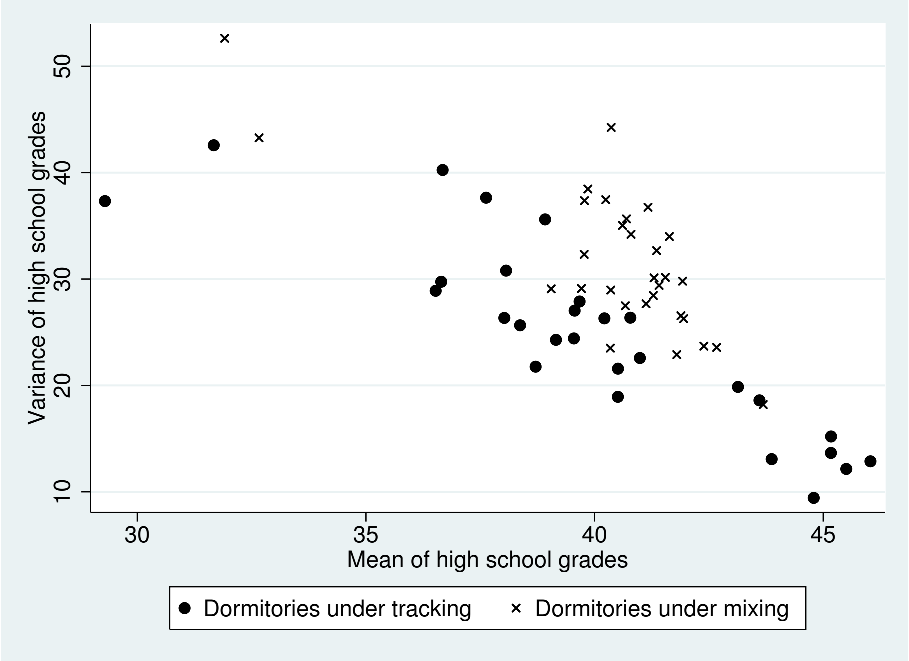


Figure 3: CDF of baseline academic characteristics (high school grades)

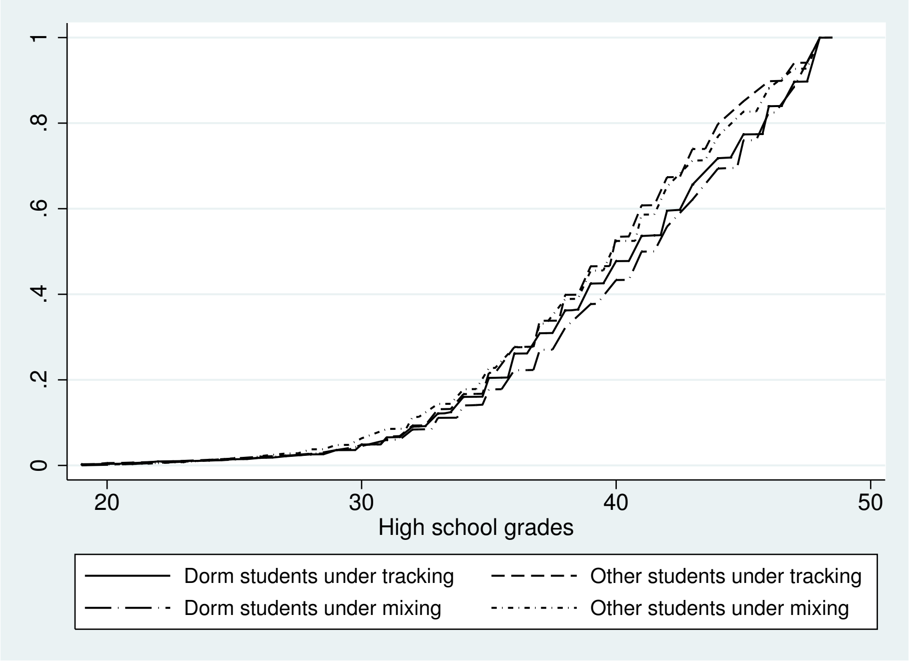
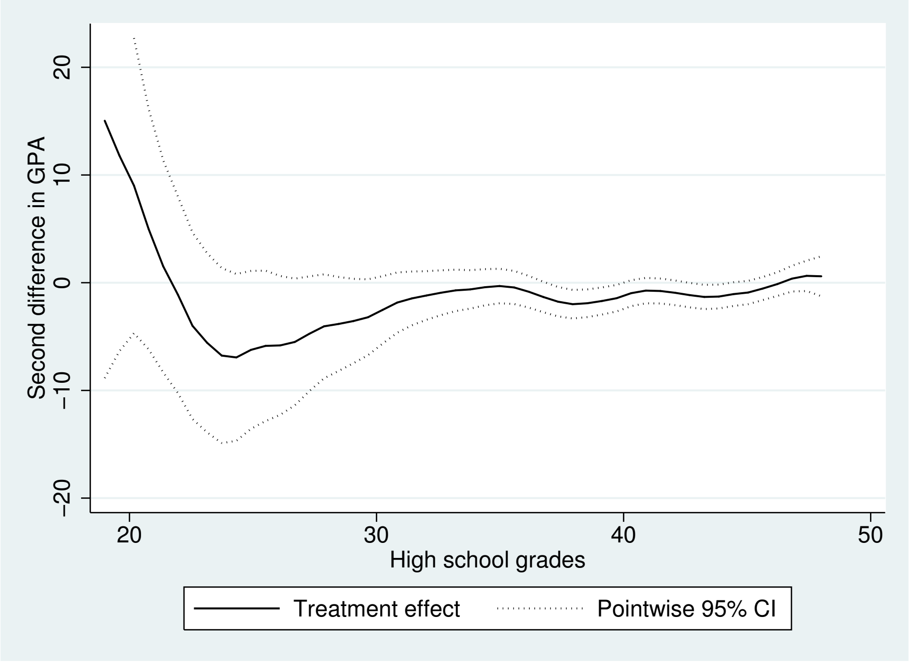
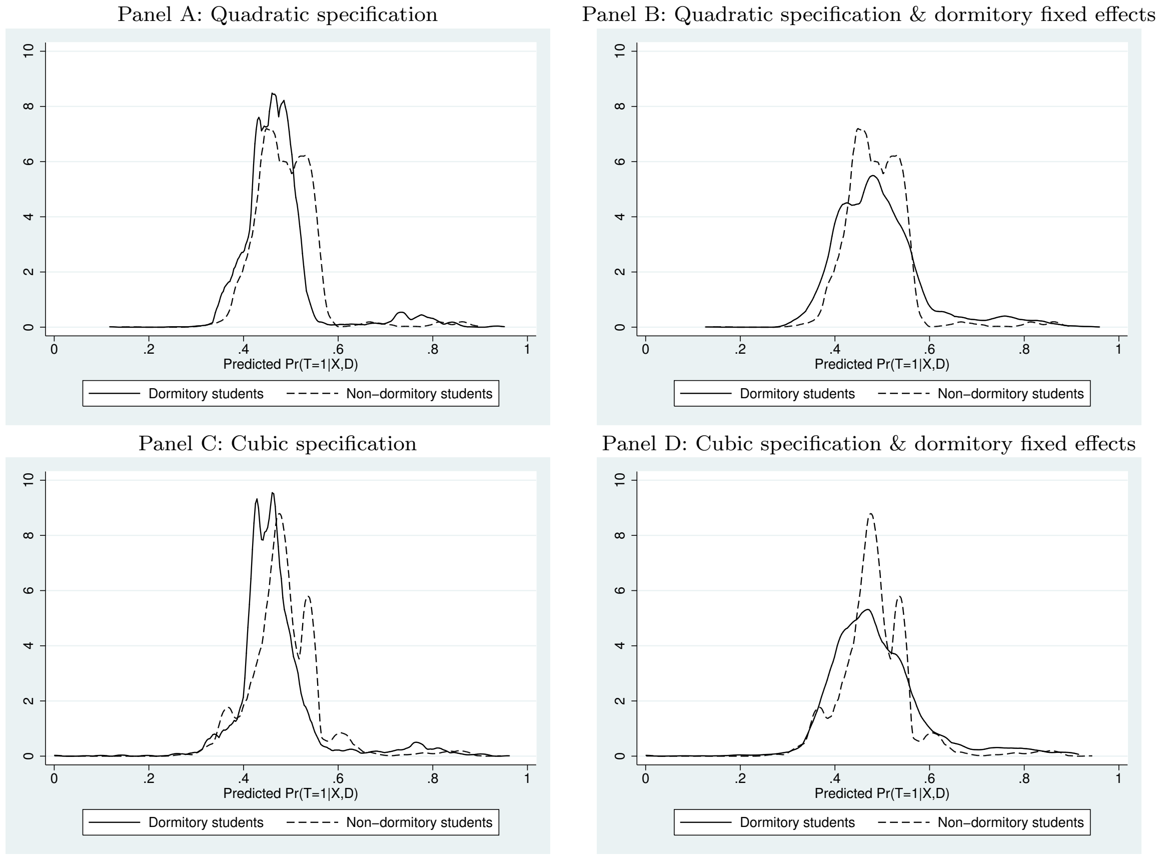


Figure 4: Difference-in-difference estimates from local linear regression of GPA on high school grades



Notes: The pointwise confidence intervals are the difference between percentiles 2.5 and 97.5 of the distribution of treatment effects from 500 iterations of a mixed pairs/cluster bootstrap.

Figure 5: Propensity scores for students in the mixing period



Panel A: Quadratic specification

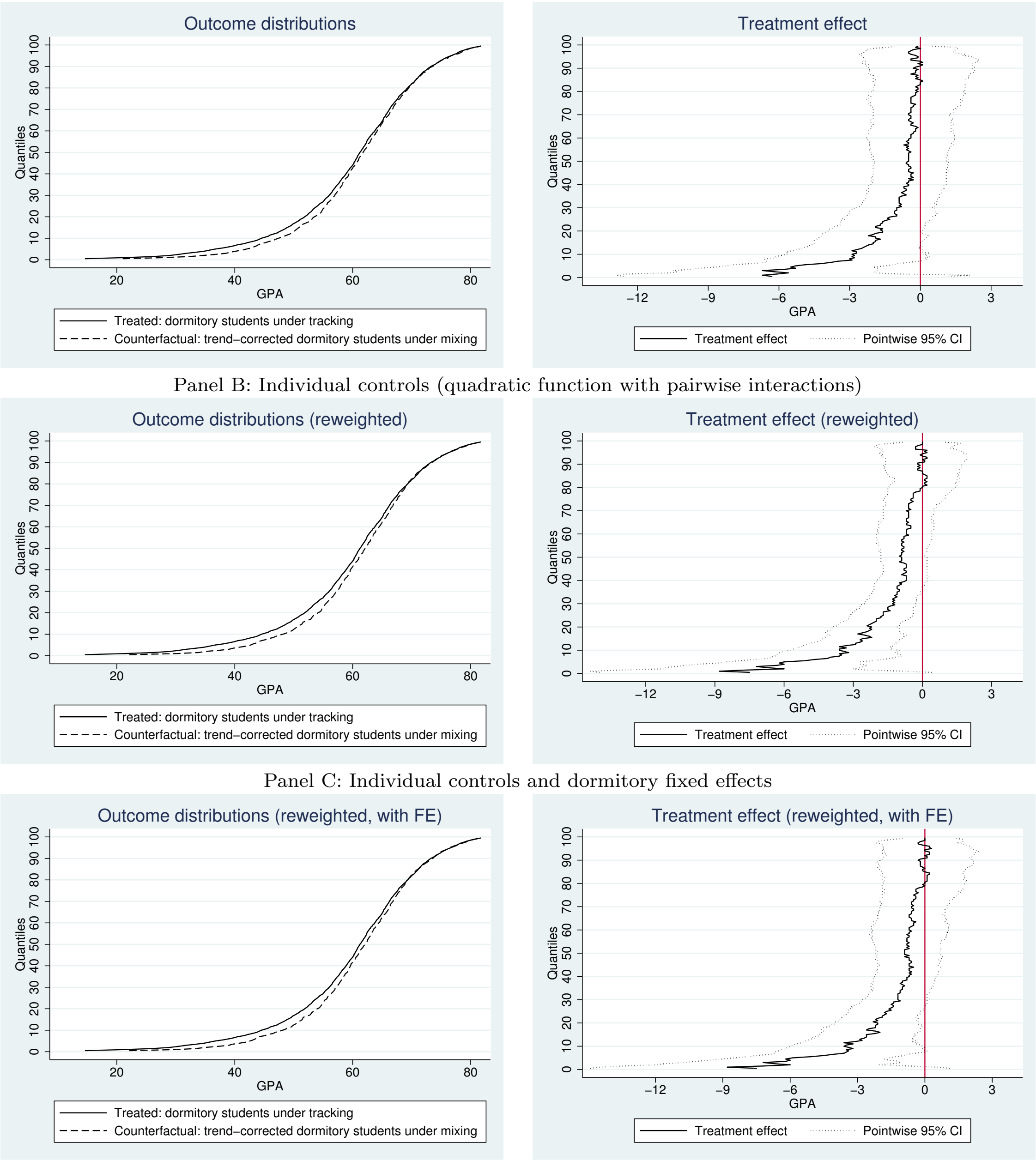
Panel B: Quadratic specification & dormitory fixed effects

Panel C: Cubic specification

Panel D: Cubic specification & dormitory fixed effects

Notes: Propensity scores are estimated using high school grades, gender, race, language, and nationality. Quadratic specification includes a linear and quadratic term for each control and all pairwise interations. Cubic specification includes a linear, quadratic and cubic term for each control and all pairwise interations.

Figure 6: Reweighted nonlinear DD estimates of the counterfactual distribution and treatment effects



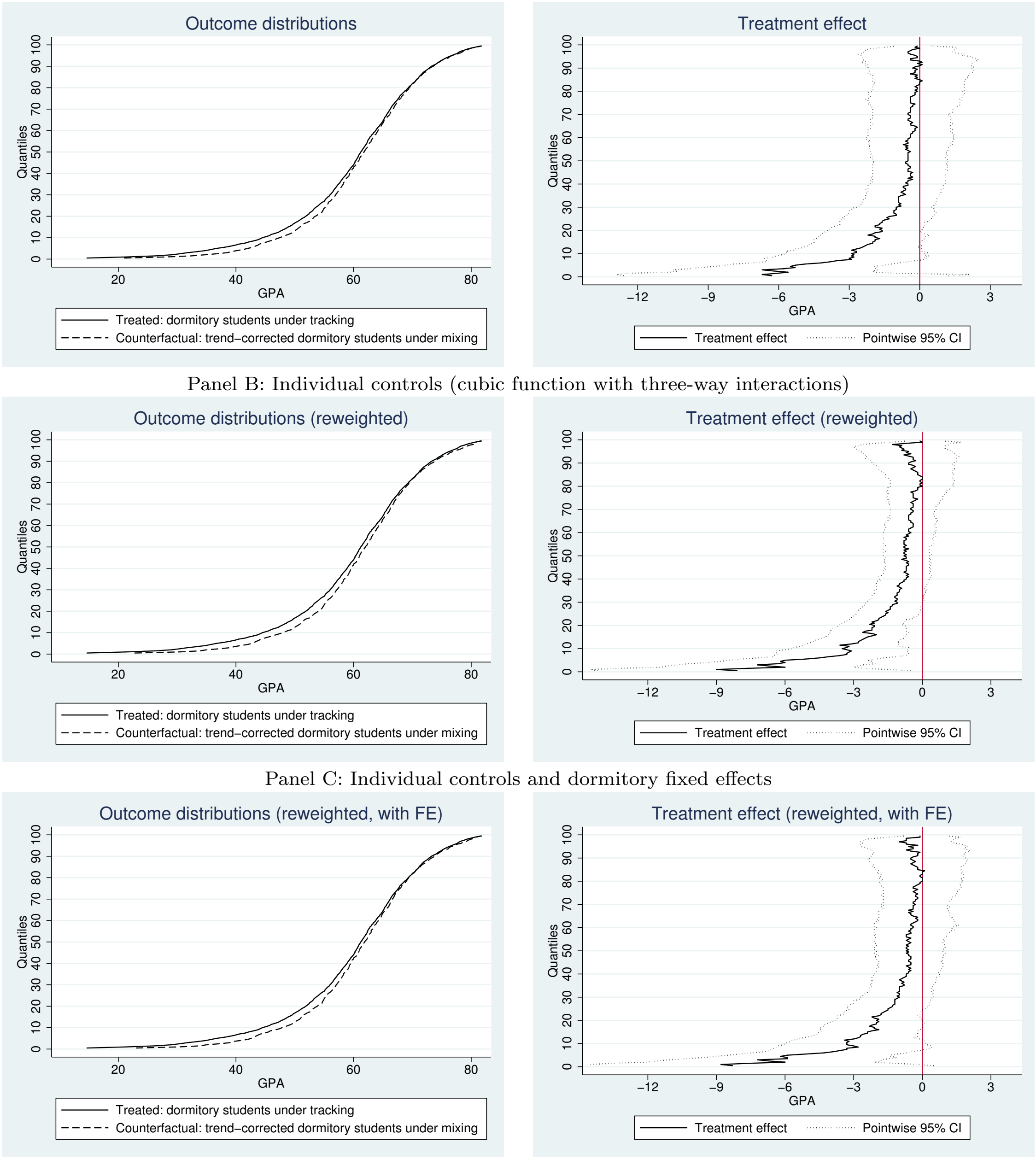
Panel A: No controls

Panel B: Individual controls (quadratic function with pairwise interactions)

Panel C: Individual controls and dormitory fixed effects

Notes: Controls are high school grades, gender, race, language, and nationality.
Controls and dormitory fixed effects are applied by reweighting students in the mixing period to have the same distribution of covariates as in the tracking period.
Reweighting functions are estimated as quadratic function of the controls with pairwise interactions.
The pointwise confidence intervals are the difference between percentiles 2.5 and 97.5 of the distribution of treatment effects from 500 iterations of a mixed pairs/cluster bootstrap.

Figure 7: Reweighted nonlinear DD estimates of the counterfactual distribution and treatment effects



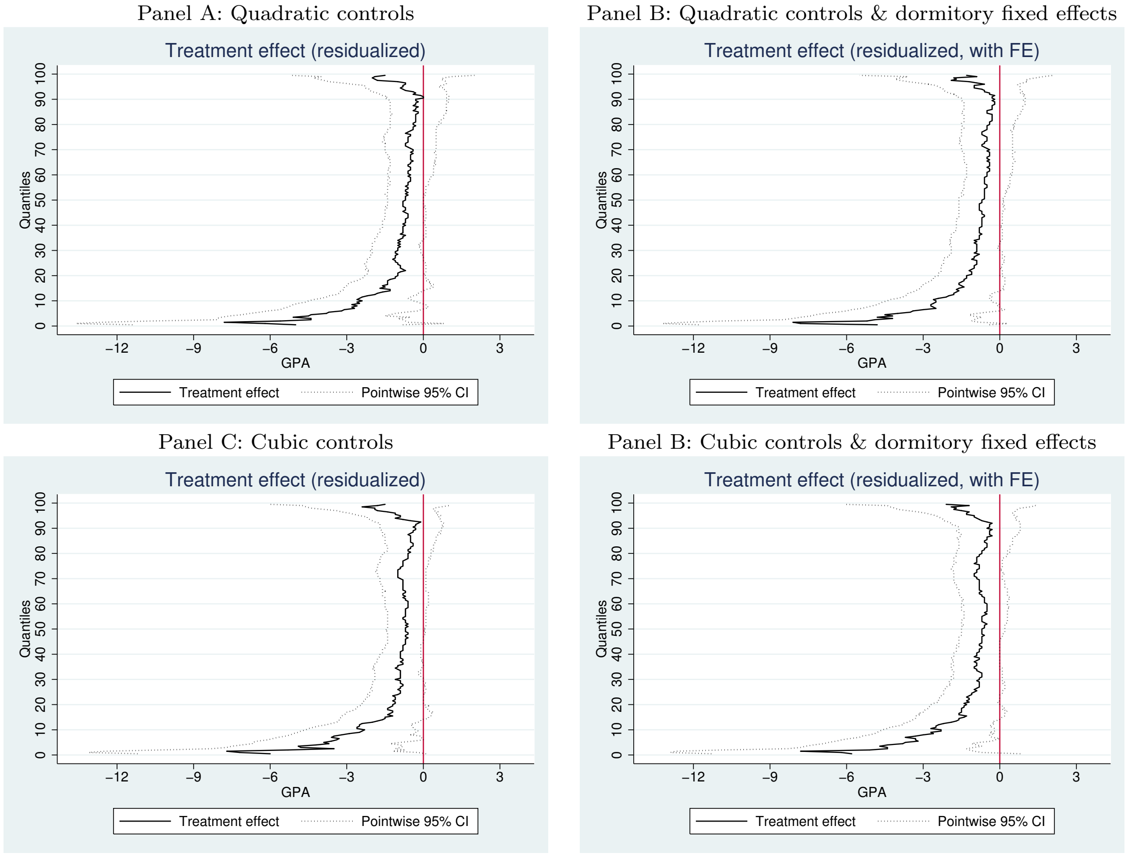
Panel A: No controls

Panel B: Individual controls (cubic function with three-way interactions)

Panel C: Individual controls and dormitory fixed effects

Notes: Controls are high school grades, gender, race, language, and nationality.
Controls and dormitory fixed effects are applied by reweighting students in the mixing period to have the same distribution of covariates as in the tracking period.
Reweighting functions are estimated as cubic function of the controls with three-way interactions.
The pointwise confidence intervals are the difference between percentiles 2.5 and 97.5 of the distribution of treatment effects from 500 iterations of a mixed pairs/cluster bootstrap.

Figure 8: Residualized nonlinear DD estimates of the counterfactual distribution and treatment effects



Panel A: Quadratic controls

Panel B: Quadratic controls & dormitory fixed effects

Panel C: Cubic controls

Panel B: Cubic controls & dormitory fixed effects

Notes: Controls are high school grades, gender, race, language, and nationality.

Controls and dormitory fixed effects are applied by applying the Athey-Imbens estimator to the residuals from a regression of university GPA on the controls, pooling data from all four groups but allowing for group-specific intercepts.

The pointwise confidence intervals are the difference between percentiles 2.5 and 97.5 of the distribution of treatment effects from 500 iterations of a mixed pairs/cluster bootstrap.