NBER WORKING PAPER SERIES

FROM NATURAL VARIATION TO OPTIMAL POLICY? THE LUCAS CRITIQUE MEETS PEER EFFECTS

Scott E. Carrell
Bruce I. Sacerdote
James E. West

Working Paper 16865
http://www.nber.org/papers/w16865

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2011

The views expressed in this article are those of the authors and do not necessarily reflect the official policy or position of the United States Air Force, Department of Defense, the U.S. Government, or the National Bureau of Economic Research. This research was partially funded by the National Academy of Education and Spencer Foundation. Thanks to D. Staiger, R. Fullerton, R. Schreiner, B. Bremer, K. Silz-Carson. Note: An earlier version of this paper was circulated under the title, “Beware of Economists Bearing Reduced Forms? An Experiment in How Not To Improve Student Outcomes.”

From Natural Variation to Optimal Policy? The Lucas Critique Meets Peer Effects
Scott E. Carrell, Bruce I. Sacerdote, and James E. West
NBER Working Paper No. 16865
March 2011
JEL No. I2,J01

ABSTRACT

We take cohorts of entering freshmen at the United States Air Force Academy and assign half to peer groups with the goal of maximizing the academic performance of the lowest ability students. Our assignment algorithm uses peer effects estimates from the observational data. We find a negative and significant treatment effect for the students we intended to help. We show that within our "optimal" peer groups, students self-selected into bifurcated sub-groups with social dynamics entirely different from those in the observational data. Our results suggest that using reduced-form estimates to make out-of-sample policy predictions can lead to unanticipated outcomes.

Scott E. Carrell
Department of Economics
University of California, Davis
One Shields Avenue
Davis, CA 95616
and NBER
secarrell@ucdavis.edu

Bruce I. Sacerdote
6106 Rockefeller Hall
Department of Economics
Dartmouth College
Hanover, NH 03755-3514
and NBER
Bruce.I.Sacerdote@dartmouth.edu

James E. West
Department of Economics and Geosciences
U.S. Air Force Acdemy
2354 Fairchild Dr. #6K100
USAF Academy, CO 80840
Jim.West@usafa.edu

From Natural Variation to Optimal Policy? The Lucas Critique Meets Peer Effects*

Scott E. Carrell
UC Davis and NBER

Bruce I. Sacerdote
Dartmouth College and NBER

James E. West
USAF Academy

February 25, 2011

Abstract

We take cohorts of entering freshmen at the United States Air Force Academy and assign half to peer groups with the goal of maximizing the academic performance of the lowest ability students. Our assignment algorithm uses peer effects estimates from the observational data. We find a negative and significant treatment effect for the students we intended to help. We show that within our "optimal" peer groups, students self-selected into bifurcated sub-groups with social dynamics entirely different from those in the observational data. Our results suggest that using reduced-form estimates to make out-of-sample policy predictions can lead to unanticipated outcomes.

Peer effects have been widely studied in the economics literature due to the perceived importance peers play in workplace, educational, and behavioral outcomes. Previous studies in the economics literature have focused almost exclusively on the *identification* of peer effects and have only hinted at the potential policy implications of the results.[1] Recent econometric studies on assortative matching by Bryan S. Graham, Guido W. Imbens & Geert Ridder (2009), and Debopam Bhattacharya

*The views expressed in this article are those of the authors and do not necessarily reflect the official policy or position of the United States Air Force, Department of Defense, or the U.S. Government. This research was partially funded by the National Academy of Education and Spencer Foundation. Thanks to D. Staiger, R. Fullerton, R. Schreiner, B. Bremer, K. Silz-Carson. Note: An earlier version of this paper was circulated under the title, "Beware of Economists Bearing Reduced Forms? An Experiment in How Not To Improve Student Outcomes."

[1] For recent studies in higher education see: (Bruce L. Sacerdote 2001, David J. Zimmerman 2003, Ralph Stinebrickner & Todd R. Stinebrickner 2006, Scott E. Carrell, Richard L. Fullerton & James E. West 2009, Scott E. Carrell, Frederick V. Malmstrom & James E. West 2008, Gigi Foster 2006, David S. Lyle 2007).

(2009) have theorized that individuals could be sorted into peer groups to maximize productivity. However, unless measured peer effects are nonlinear across individuals, there is no social gain to sorting individuals into peer groups.[2]

This study takes a first step in determining whether student academic performance can be improved through the systematic sorting of students into peer groups. We first identify nonlinear peer effects at the United States Air Force Academy (USAFA) and create optimally designed peer groups. Using an experimental design, we sort the incoming college freshman cohorts at USAFA into peer groups during the fall semesters of 2007 and 2008 with the objective of improving (for the treatment group) the grades of the bottom one-third of incoming students by academic ability.[3] Half of the students were placed in the control group and randomly assigned to squadrons, as was done with preceding entering classes. The other half of students (the treatment group) were sorted into squadrons in a manner intended to maximize the academic achievement of the students predicted to be in the lowest third of first year grades. The reduced form coefficients predicted a Pareto-improving allocation in which grades of students in the bottom third of the academic distribution would rise, on average, 0.056 grade points while students with higher predicted achievement would be unaffected.

Despite this prediction, actual outcomes from the experiment yielded quite different results. For the lowest ability students we observe a negative and statistically significant treatment effect of -0.054. For the middle ability students, expected to be unaffected, we observe a positive and significant treatment effect of 0.067. High ability students were unaffected by the treatment.

Our results show the important role that peers play in the education production process; however, they also highlight the danger in using reduced form peer effects estimates to actively sort individuals into peer groups without a thorough understanding of the underlying mechanisms that drive the social interactions. The latter point brings to mind the Lucas Critique of the Phillips curve as an exploitable policy relationship due to changing structural parameters (Robert Lucas 1976), and the appendix to Milton Friedman & Anna J. Schwartz (1991), where Friedman recounts his experience as a statistician during World War II. On the basis of a multiple regression out-of-sample forecast and without any knowledge of metallurgy, he proposed the composition of a new alloy for use in high temperature applications that proved to be vastly inferior to those contained in the observational data.

[2] If peer effects are linear in means, a "good" peer taken from one group and placed into another group will have equal and offsetting effects on both groups.

[3] This objective function was determined by USAFA senior leadership who had a strong desire to reduce the academic probation rate, then at roughly 20 percent.

We explore possible explanations for this perverse finding. One hypothesis is that the negative treatment effect is simply due to sampling variation. A second hypothesis is that our original findings were spurious and perhaps biased by over-fitting of the observational data to a large number of possible peer effects variables and functional forms. A third hypothesis is that the data generating process changed in a fundamental way. However, the data point to a fourth hypothesis which is that our "optimally" sorted squadrons, withmore extreme variation in the proportion of high and low ability students (i.e. bifurcation), have a unique social dynamic not seen in the observational data that is counterproductive to the achievement of low ability students. That is, high and low ability students in the treatment squadrons appear to have segregated themselves into separate social networks, resulting in decreased beneficial social interactions among group members. For the middle predicted achievement students, evidence suggests that the positive treatment effect occurred because these students did not interact with low predicted achievement students and were placed into more homogeneous peer groups. This finding is consistent with recent evidence on ability grouping and tracing by Ester Duflo, Pascaline Dupa & Michael Kremer (2008).

Results from this study are significant for several reasons. We believe this is the first study in the literature that uses peer effects estimates to actively sort individuals into peer groups, implementing the recent econometric literature on assortative matching by Bhattacharya (2009) and Graham, Imbens & Ridder (2009). The study is unusual in its use of historical observational data to infer optimal policy, implement, and then test the efficacy of the policy in a controlled experiment. In addition, our results highlight the significant role that peers play in the education production process. Finally, the unexpected results of the experiment suggest that using reduced form peer effects estimates to conduct out-of-sample policy predictions may lead to unanticipated outcomes. Hence, further work in this area will require knowledge of the underlying mechanisms or structure that drive the social network.

The remainder of the paper proceeds as follows. Section 1 presents the data and estimates the nonlinear peer effects at USAFA. Section 2 describes the squadron sorting mechanism. Section 3 describes the experimental design and provides simulated results. Section 4 presents results from the experiment. Section 5 explores reasons for the experiments' unexpected findings. Section 6 concludes.

1 Data

1.1 The Dataset

Our pre-treatment (i.e. observational) dataset includes all students in the USAFA graduating classes of 2005 through 2010, while our experimental subjects are all members of the USAFA graduating classes of 2011 and 2012. The data contain individual-level demographic information as well as measures of student academic, athletic and leadership ability. Pre-treatment academic ability is measured as *SAT verbal* and *SAT math* scores and an *academic composite*. The composite is computed by the USAFA admissions office and is a weighted average of an individual's high school GPA, class rank, and the quality of the high school attended. Athletic aptitude is measured as a score on a fitness test required of all applicants prior to entrance. Leadership aptitude is measured as a weighted average of high school and community activities.

Freshman academic performance is measured as grade point average (GPA). GPA is a consistent measure of performance across all students in our sample because students at USAFA spend their entire freshman year taking required core courses with a common exam and do not select their own coursework. Students have no ability to choose their professors. Core courses are taught in small sections of approximately 20 students, with students from all squadrons mixed across classrooms. Faculty members teaching the same course use an identical syllabus and give the same exams during a common testing period. This institutional characteristic assures there is no self-selection of students into courses or towards certain professors. Carrell, Fullerton & West (2009) and Scott E. Carrell & James E. West (2010) provide detailed tests of the randomness of the peer group and classroom assignments at USAFA to ensure estimates are not biased by self-selection. A complete list of summary statistics is provided in Table 1.

1.2 Methods

As described in Carrell, Fullerton & West (2009), we use the random assignment of USAFA students to peer groups (i.e. military squadrons), to identify peer effects in academic performance free of biases arising from self-selection.[4]

[4] Conditional on a few demographic characteristics the students in our study are randomly assigned to a peer group in which they live in adjacent dorm rooms, dine together, compete in intramural sports together, and study together. They have limited ability to interact with other students outside of their assigned peer group during their freshman year of study.

Consider a structural model of peer effects in academic achievement, where own achievement is a function of own pre-treatment characteristics, the simultaneous achievement of ones peers, and their pre-treatment characteristics,

$$GPA_{iscrt} = \alpha_0 + \alpha_1 X_{iscr} + \alpha_{2t}\overline{GPA}_{-iscr} + \alpha_{3t}\overline{X}_{-iscr} + \epsilon_{iscrt} \tag{1}$$

where GPA_{iscrt} is the freshman fall semester GPA for individual i in squadron s, graduating class c, semester r, and of academic ability t. X_{iscrt} is a vector of individual i's specific (pre-treatment) characteristics, including SAT math, SAT verbal, academic composite, fitness score, leadership composite, race/ethnicity, gender, recruited athlete, and whether they attended a military preparatory school. $\overline{GPA}_{-iscrt}$ is the average freshman fall semester GPA in squadron s excluding individual i. $\overline{X}_{-iscr}$ is likewise the average of pre-treatment characteristics is squadron s excluding individual i. ϵ_{iscrt} is the error term. Following Charles F. Manski (1993), α_1 represents the exogenous peer effect and α_2 is the endogenous peer effect.

Averaging over equation (1) to derive $\overline{GPA}_{-iscr}$ and consolidating, we derive the reduced form equation in the structural parameters.

$$\begin{aligned} GPA_{iscrt} &= \frac{\alpha_0}{1-\alpha_{2t}} + \frac{\alpha_1 - \alpha_1\alpha_{2t} + \alpha_{2t}\alpha_{3t}}{1-\alpha_{2t}} X_{iscr} + \frac{\alpha_1\alpha_{2t} + \alpha_{3t}}{1-\alpha_{2t}} \overline{X}_{-iscr} + \tilde{\epsilon}_{iscrt} \\ &= \beta_{0t} + \beta_{1t} X_{iscr} + \beta_{2t}\overline{X}_{-iscr} + \tilde{\epsilon}_{iscrt} \end{aligned} \tag{2}$$

We include graduating class (cohort) fixed effects and semester fixed effects to control for mean differences across years and semesters in GPA. Given the potential for error correlation across individuals within a given squadron and class, we cluster all standard errors at the squadron by graduating class level.

Carrell, Fullerton & West (2009) found large and statistically significant reduced form peer effects estimating equation (2) at USAFA. Specifically, they found student academic performance increased significantly with the average peer SAT verbal scores in the squadron. Additionally, Carrell, Fullerton & West (2009) found evidence of nonlinear effects in which low predicted achievement students benefit the most from the presence of high ability peers. To determine whether student outcomes can be improved through systematic sorting of individuals into peer groups, we take a similar approach and estimate a nonlinear model in which we allow the peer coefficients to vary by own predicted achievement. Specifically, we estimate separate peer coefficients for each third of the own predicted GPA distribution.

We estimate models using both mean peer ability and the *proportion* of peers in the group who have relatively high and low peer SAT scores.[5] Our definition of a "high" (low) score is any peer in the top (bottom) quartile of the year-cohort SAT verbal distribution.[6]

We estimate equation (2) using ordinary least squares (OLS) and results are shown in Table 2. Specification 1 estimates a single coefficient for each peer characteristic while Specification 2 allows separate coefficients for each third of the predicted GPA distribution. Overall, the nonlinear model in Specification 2 finds larger and more precisely estimated peer effects than Specification 1 or a traditional linear in means model as in Carrell, Fullerton & West (2009).[7] The results suggest several nonlinearities in the data. The model fit in Specification 2 rejects the restrictions in Specification 1 at the 0.01-level ($F = 3.57$) and the six peer variables are jointly significant at the 0.01-level ($F = 3.48$). The coefficient on the fraction of peers in the top quartile of the SAT verbal distribution is positive and significant for both low (0.481) and high (0.215) ability students and negative and insignificant for middle ability students. Across the three predicted GPA groups, the peer coefficients are significantly different from one another. The coefficient on the fraction of peers in the bottom quartile of the SAT verbal distribution is negative and statistically significant for the middle (-0.193) ability students and statistically insignificant for low and high ability students.

The results suggest that low predicted GPA students benefit most from having peers with high SAT verbal scores while middle ability students benefit from being separated from peers with low SAT verbal scores. These conclusions are supported by similar specifications using peer measures other than ones based on SAT verbal scores. However, of all peer variables, peer SAT verbal scores are the most statistically significant.

Under the direction of the Superintendant of the US Air Force Academy we used this model to sort the freshman students entering USAFA in the fall of 2007 and fall of 2008 (the graduating class of 2011 and 2012) into peer groups with the intent of improving the grades of the lowest one-third of incoming ability students.

[5] We also find qualitatively similar results when using the *number* of peers who have high or low scores in the pre-treatment variables.

[6] For example, for the class of 2010 the top quartile of the SAT verbal distribution was 670 and above and the bottom quartile was 570 and below. We also find qualitatively similar results when estimating the model using other points of the distribution such as thirds and deciles.

[7] For brevity we do not show results for the linear in means model. Results are available upon request.

2 Sorting Methodology

To optimally sort students into squadrons, we draw on recent work on assortative matching by Bhattacharya (2009) and Graham, Imbens & Ridder (2009). For each entering cohort, approximately 650 students in the treatment group were assigned to one of 20 squadrons. Let $p_{i,s}$ be the probability student i is allocated to squadron s, thus, $p_{i,s} \in \{0,1\}$. The allocation matrix is then

$$P = \begin{pmatrix} p_{1,1} & \cdots & p_{1,20} \\ p_{2,1} & \cdots & p_{2,20} \\ \vdots & \ddots & \vdots \\ p_{650,1} & \cdots & p_{650,20} \end{pmatrix}$$

Every student must be assigned to a squadron, thus,

$$\sum_{s=1}^{20} p_{i,s} = 1 \quad i = 1..650$$

Every squadron s must contain N_s students, thus

$$\sum_{i=1}^{650} p_{i,s} = N_s \quad s = 1..20$$

$$31 \leq N_s \leq 33$$

One's peers are the additional members of the squadron. The average peer attributes are thus

$$Z_{i,s} = \frac{1}{N_s} \sum_{j \neq i} p_{j,s} X_{j,s}$$

Student i, assigned to squadron s and of academic type t, has $GPA_{i,s,t}$, which is a function of own attributes, X_i, and peer attributes, $Z_{i,s}$. Peer coefficients vary by academic type of the student, t, (low, middle, or high predicted GPA) as shown in Table 2, Specification 2.

$$GPA_{i,s,t} = X_i\beta + Z_{i,s}\gamma_t + \epsilon_{i,s,t} \tag{3}$$

Since own effects do not change with squadron assignment, maximizing GPA for the lowest third of students is equivalent to maximizing the positive peer effects experienced by these students, $Z_{i,s}\gamma_l$.[8] Thus, we optimize

$$\max_{p_{i,s},\lambda_i,\delta_s} \left[\min_{i \in I_l} \left\{ \sum_{s=1}^{20} p_{i,s} \frac{1}{N_s} \sum_{j \neq i} p_{j,s} X_{j,s} \gamma_l - \lambda_i \left(1 - \sum_{s=1}^{20} p_{i,s} \right) - \delta_s \left(N_s - \sum_{i=1}^{650} p_{i,s} \right) \right\} \right] \tag{4}$$

[8] Upon the request of USAFA officials, our algorithm constrained each squadron to have a relatively even distribution of females, Hispanics, blacks, recruited athletes, and students who attended a military preparatory school.

We solved the constrained optimization problem using the nonlinear optimizer in XpressMP.[9] Given that membership in a squadron and squadron size are linear functions of $p_{i,s}$, our objective function is nonlinear in the choice variable $p_{i,s}$.

3 Experimental Design

The graduating classes of 2011 and 2012 entered USAFA with $1,314$ and $1,391$ students, respectively. Half of the incoming classes were randomly assigned to the control group and half to the treatment group.[10] Table 3 shows a regression of membership in the treatment group on the pre-treatment variables. Specification 1 shows results for the class of 2011, Specification 2 shows results for the class of 2012, and Specification 3 shows a combined regression. Results show no statistical differences in the observed attributes between the treatment and control groups. For example, the joint F statistic for the combined samples is 0.26 with a p-value of 0.99. Figure 1 shows the distribution of predicted grades (excluding any potential peer effects) for students in the treatment and control groups. A Wilcoxon rank-sum test fails to reject the null hypothesis that the treatment and control samples are random draws from a single population (p-value $= 0.64$).

Students in the control group were *randomly* assigned to one of the 20 control squadrons according to an algorithm, which has been used by USAFA since the summer of 2000. The algorithm provides an even distribution of students by demographic characteristics.[11] Students in the treatment group were assigned to one of 20 treatment squadrons using the optimal sorting mechanism presented in the previous section. The algorithm maximized the positive peer effect experienced by the students who are in the bottom one-third of the incoming academic ability distribution. More specifically we maximized the minimum peer effect experienced by a low ability student.[12]

[9] XPressMP was provided to us by FICO under their Academic Partners Program.

[10] The random division was subject to the constraint that siblings were split between the treatment and control groups.

[11] Specifically, the USAFA admissions office implements a stratified random assignment process where females are first randomly assigned to squadrons. Next, male ethnic and racial minorities are randomly assigned, followed by male non-minority recruited athletes. Students who attended a military preparatory school are then randomly assigned. Finally, all remaining students are randomly assigned to squadrons. Students with the same last name, including siblings, are not placed in the same squadron. This stratified process is accomplished to ensure demographic diversity across peer groups.

[12] The random selection of the treatment and control squadrons was stratified across the four cadet "groups" which contain 10 squadrons each. It was also stratified with respect to new and returning "Air Officers Commanding" or AOCs, the officer in charge of military training within each squadron. This was done to eliminate any potential group or AOC-level common shocks to academic performance. We flipped the treatment and control squadrons after the

Figure 2 shows histograms of student characteristics in the treatment and control squadrons by student ability. We note the sorting mechanism created squadrons which are quite different in make-up compared to the historical observational data used to estimate the peer effects. Relative to randomly assigned squadrons, the optimal sorting mechanism assigned low predicted GPA students in the treatment group to squadrons with a much higher proportion of peers with SAT verbal scores in the top quartile. In the process, the algorithm also created a number of treatment squadrons with no low ability students. In contrast, for the classes of 2005-2010 there were no freshman squadrons containing zero low ability students while eleven such squadrons existed in the treatment group for the classes of 2011 and 2012. We intentionally allowed the algorithm to engage in extreme sorting to maximize the potential peer effects and the perceived statistical power of the experiment.

Table 4 shows predicted GPA and predicted treatment effect by student ability. For students in the bottom third of incoming academic ability the estimated treatment effect is a statistically significant 0.056 grade points. For students in the middle and top third of the academic distribution, the estimated treatment effects are positive, but statistically insignificant. Figure 3 plots the distribution of predicted GPA after the sort. These predictions imply that the optimal sorting mechanism predicts a Pareto-improving allocation relative to random assignment.

To estimate the likelihood of observing a positive treatment effect given the underlying variability of grades, we conducted a Monte Carlo simulation. Specifically we simulated the treatment effect for the bottom one-third of students as being equal to the fitted values from Column 2 in Table 2 plus two stochastic error terms, one with the statistical properties of student level grade variation and the other with properties of squadron level variation.[13]

Figure 4 plots the statistical power of the experiment for values of the key peer coefficient (percent of high SAT Verbal peers on low ability students) ranging from 0 to 1. At the vertical line, representing our estimated peer coefficient of 0.481, 630 of $1,000$ draws were positive and statistically significant at the 0.05 level.

4 Experimental Results

Actual results of the experiment are shown in Table 5 and Figure 5. There are two striking findings. First, the estimated treatment effect for the lowest ability students is negative and statistically significant. The magnitude of the effect (-0.054) indicates that the treatment was of the magnitude

first year of the experiment.

[13] The estimated variance of the error term was obtained from the observational data in predicting student grades.

predicted but the opposite sign, meaning that low ability students in the treatment group performed significantly worse than those in the control group. The second striking finding is the positive and statistically significant (0.067) treatment effect for students in the middle third of the predicted GPA distribution.

5 Why the Unexpected Results?

Given the unanticipated findings of the experiment, we next explore four possible explanations. First, we examine whether the effect could be due to sampling variation. Second, we test the robustness of the nonlinear reduced form peer effects that motivated the experiment. We ask whether our initial finding of reduced form peer effects may have been spurious and possibly a result of fitting the observational data to a large number of different peer variables and different functional forms. Third, we ask whether the data generating process changed fundamentally. Did something about the students or institution alter the process by which social interactions occur in the fall of 2007? Finally, we investigate whether the extreme sorting (and bifurcation) in the treatment groups created by our algorithm lead to unexpected peer dynamics in the treatment squadrons.

5.1 Is the Effect Due to Sampling Variation?

One possibility is that the negative treatment effect is simply due to sampling variation; meaning that a positive treatment effect exists, but that it was unobservable due to the statistical variation of GPA. To assess the likelihood of this event, we note that in a Monte Carlo power simulation, only in one draw out of $1,000$ was the treatment effect negative and significant at the 0.10-level. Hence, we conclude the negative and significant treatment effect is not likely due to sampling variation.

5.2 Did We Imagine the Peer Effects?

To test the robustness of the estimated peer effects, Table 6 shows results in the observational data when estimating the full set of possible peer coefficients in a flexible functional form. We use all three possible measures of academic ability (SAT verbal, SAT math, and academic composite) and allow for the proportion of peers in the top or bottom of these distributions to each have a separate effect. We further allow these six possible effects to vary by own predicted GPA (three groups) yielding a total of eighteen peer coefficients. Testing for the joint significance of all eighteen peer coefficients is a much more conservative test for the existence of peer effects. Results show that the full set of academic peer variables are jointly significant at the $0.10-$level and the coefficients for the SAT verbal variables are jointly significant at the $0.01-$level. Importantly, the magnitude and significance of the coefficient we used to sort students, the fraction of peers in the top quartile of the SAT verbal distribution for low ability students, is virtually unchanged compared to the restricted model of equation (1) reported in Table 2.

As a second robustness test, Table 7 shows results when splitting the sample across years. We do this to examine whether the significant peer effects were driven by a few (potentially spurious or unusual) years. In both subsamples, the fraction of peers in the top quartile of the SAT verbal distribution for low ability students remains positive and statistically significant at the $0.05-$level. Additionally, the magnitude of the effects is statistically indistinguishable across the two sets of years.

We conclude that the peer effects used to originally motivate the experiment are unlikely to be a statistical anomaly or the result of a failure to correct standard errors for multiple hypothesis tests.

5.3 Did the Process Change?

Although the peer effects in the observational data appear to be robust, another possibility is that the process by which peer interactions occur at USAFA changed around the time when the class of 2011 matriculated. This may be due to some unobserved policy or leadership change, or changing student attitudes and behaviors. To test this hypothesis, we examine the magnitude and significance of the reduced-form peer effects in the randomly assigned control group, in which students were assigned to squadrons according to the process used in the observational data. We combine the observational and control data, and test for structural change between the two groups.[14] Table 8 presents these findings. For low ability students in the control group, the coefficient on the fraction of peers in the top quartile of the SAT verbal distribution is positive and significant (0.593) at the 0.10-level. We fail to find evidence of structural change between the observational and control data, as this coefficient is statistically indistinguishable from its companion coefficient in the observational data ($F = 0.093$, $p = 0.761$). Furthermore the key non-linearity in which low ability students benefit more from high ability peers than do middle ability students is present in both the observational and the control groups.

[14] We do not estimate the reduced-form effects in the treatment group because there is virtually no variation in the fraction of peers in the top quartile of the SAT verbal for low ability students.

As a second test, we estimate the endogenous peer effects model in which we regress own GPA on concurrent peer GPA. Due to the reflection and common shocks problems, estimated coefficients are upward biased estimates of true contemporaneous peer effects. However, standard errors of estimated coefficients are much smaller than those estimated using unbiased estimation techniques such as two-stage least squares. In spite of biased estimates, the endogenous peer effects model can provide evidence of the existence of peer effects and has been utilized in prior studies (Sacerdote 2001, Lyle 2007). Results in Table 9 show large positive and statistically significant endogenous effects for all subgroups in both the observational and control groups. However, the effects are smaller and statistically insignificant in the treatment group. Most notably, the effect for the lowest ability students in the treatment group is negative (-0.015).

These results provide evidence that the process by which peer interactions occurred in the randomly assigned control squadrons was not likely different than what occurred in the pre-experiment observational squadrons. However, the results suggest that something very different may have occurred in the treatment squadrons. We explore this hypothesis in the next section.

5.4 Did the Peer Dynamics in the Treatment Groups Change?

A third possible explanation for the observed negative treatment effect is that the extreme variation in the treatment squadrons caused the peer dynamics in the treatment squadrons to change. As shown in Figure 2, the sorting algorithm created rather different squadrons than those previously observed under random assignment. Figures 6 and 7 provide more detail by showing the distribution of low SAT peers in the observational, treatment, and control groups. While low ability students in the treatment group were assigned an unusually large number of high ability peers (Figure 5), they were also assigned an unusually large number of low ability peers (Figure 6). This was achieved by removing the middle ability peers and placing them in homogenous squadrons of primarily middle ability peers. In other words the sorting procedure lead to a combination of 1) bifurcated squadrons with many low ability students grouped together with students with high SAT-Verbal scores and 2) homogenous squadrons consisting of middle and high ability students that earned lower SAT-Verbal scores.

Although the extreme type of bifurcation our algorithm created in the treatment squadrons was not present in the observational data, more limited bifurcation did occasionally occur as a result of random sampling variation. In Table 10, we test to see if various indicators of bifurcation had any effect on the academic achievement of low predicted GPA students in the pre-experimental observational data. Across all four indicators of bifurcation, low predicted GPA students in more bifurcated squadrons performed *better* than average, with three of the four measures significant at the 10-percent level. On the basis of these results, our predicted treatment effect of 0.056 grade points was too low for omitting the beneficial effects of bifurcation observed in the observational data.

As a second look at the effects of bifurcation, we examine roommate matching. In their first semester, students at USAFA are not permitted to choose their own roommates. However, in the second semester, this prohibition is relaxed. This affords us an opportunity to test whether different social structures evolved in treatment versus control squadrons. Table 11, Panel A reports the regression of own predicted GPA for bottom third predicted GPA students on the predicted GPA of her/his roommate(s), and the endogenous regression of own first semester GPA on roommate(s) first semester GPA. In all specifications, no selection effects were found in the first semester. Panel B reports the similar exogenous and endogenous models of roommate selection for the second semester with very different results. In the control group, no evidence of selection is found. However in the treatment group, we find evidence of strong positive selection, meaning that within the treatment group those below the mean are more likely to select a roommate whose GPA is also below the mean.[15]

As a further test of whether different social structures evolved in treatment versus control squadrons, we conducted a survey of all experimental subjects in the spring of their sophomore and junior years. In this survey, we asked students to name up to five students with whom they studied as a freshman and up to five students with whom they spent free time as a freshman. We received usable responses from approximately 25 percent of the experimental subjects. Table 12 reports various measurements of social structures inferred from the survey data. In columns 1 through 3, we regress the numbers of low, medium, and high predicted GPA study partners respectively on various subgroups within our data. Results show that low ability students in the treatment group report having 0.524 more low ability study partners and 1.105 fewer middle ability study partners than those in the control group. Additionally, we find no significant difference in the number of high SAT-Verbal study partners relative to the control group.[16]

These results provide compelling evidence of why our experiment likely failed to produce its intended positive treatment effect. While our sorting algorithm placed low predicted GPA students in peer groups with a large number of students with high SAT-Verbal scores, they were no more likely to study with these types of students. Instead, low ability students in the treatment group

[15] Roommate data were only available for a subset of students in the sample in the class of 2012.

[16] Results show a similar pattern for friendship formations. On average, low ability students report having 0.658 more low ability friends relative to control.

opted to study with other low ability students. We find that in the choice of roommates, study partners, and friends, there is empirical evidence that different social structures evolved in the treatment versus control groups.

6 Conclusion

This study set out to examine whether a fixed set of students could be sorted into peer groups in a way that would improve either aggregate student academic performance or at least the performance of the lowest ability students. To do so, we identified nonlinear peer effects in academic performance at the United States Air Force Academy (USAFA) and created "optimally" designed peer groups based on the reduced form effects in the observational data. We sorted the entire freshman cohorts for the classes of 2011 and 2012. A randomly chosen half of the incoming freshman were randomly assigned to the control squadrons while the other half were sorted into the treatment squadrons. The reduced form coefficients predicted a Pareto-improving allocation in which students' grades in the bottom third of the academic distribution would rise, on average, 0.056 grade points while higher ability student's grades would be unaffected.

Despite this prediction, results from the experiment yielded a rather different outcome. For the lowest ability students, we observed a negative and statistically significant treatment effect of -0.054. For the middle ability students, predicted to be unaffected, we observed a positive and statistically significant treatment effect of 0.067.

We find evidence in the choice of roommates, study partners, and friends that social structures evolved in the treatment group that were not observed in the pre-treatment observational data used to infer our "optimal policy". We conclude that using reduced form peer effects estimates is not sufficiently descriptive of peer group formation to allow reliable implementation of "optimal policy". These findings bear similarity to Lucas (1976) and Friedman & Schwartz (1991).

References

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

Carrell, Scott E., and James E. West. 2010. "Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors." *Journal of Political Economy*, 118(3): 409–432.

Carrell, Scott E., Frederick V. Malmstrom, and James E. West. 2008. "Peer Effects in Academic Cheating." *Journal of Human Resources*, 43(1): 173–207.

Carrell, Scott E., Richard L. Fullerton, and James E. West. 2009. "Does Your Cohort Matter? Estimating Peer Effects in College Achievement." *Journal of Labor Economics*, 27(3): 439–464.

Duflo, Ester, Pascaline Dupa, and Michael Kremer. 2008. "Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya." National Bureau of Economic Research Working Paper 14475.

Foster, Gigi. 2006. "It's not your peers, and it's not your friends: Some progress toward understanding the educational peer effect mechanism." *Journal of Public Economics*, 90(8-9): 1455–1475.

Friedman, Milton, and Anna J. Schwartz. 1991. "Alternative Approaches to Analyzing Economic Data." *The American Economic Review*, 81(1): 39–49.

Graham, Bryan S., Guido W. Imbens, and Geert Ridder. 2009. "Complementarity and Aggregate Implications of Assortative Matching: A Nonparametric Analysis." National Bureau of Economic Research Working Paper 14860.

Lucas, Robert. 1976. "Econometric Policy Evaluation: A Critique." In *The Phillips Curve and Labor Markets.* , ed. Karl Brunner and Allan H. Melzer, 19–46. American Elsevier.

Lyle, David S. 2007. "Estimating and Interpreting Peer and Role Model Effects from Randomly Assigned Social Groups at West Point." *Review of Economics and Statistics*, 89(2): 289–299.

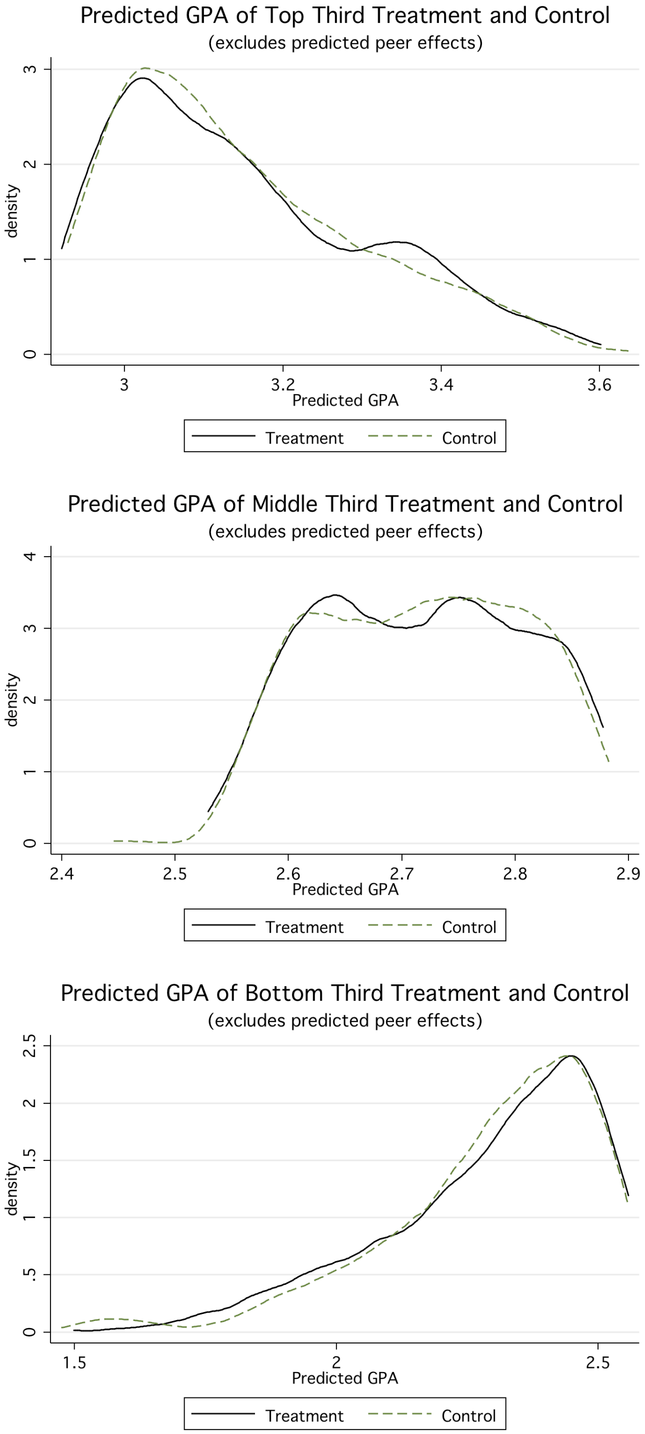
Manski, Charles F. 1993. "Identification and Endogenous Social Effects: The Reflection Problem." *Review of Economic Studies*, 60(3): 531–42.

Sacerdote, Bruce L. 2001. "Peer Effects with Random Assignment: Results for Dartmouth Roommates." *Quarterly Journal of Economics*, 116(2): 681–704.

Stinebrickner, Ralph, and Todd R. Stinebrickner. 2006. "What can be learned about peer effects using college reoomates? Evidence from new survey data and students form disadvantaged backgrounds." *Journal of Public Economics*, 90(8-9): 1435–54.

Zimmerman, David J. 2003. "Peer Effects in Academic Outcomes: Evidence From a Natural Experiment." *The Review of Economics and Statistics*, 85(1): 9–23.

Figure 1: Distribution of Pre-treatment Predicted GPA



Predicted GPA of Top Third Treatment and Control
(excludes predicted peer effects)

Predicted GPA of Middle Third Treatment and Control
(excludes predicted peer effects)

Predicted GPA of Bottom Third Treatment and Control
(excludes predicted peer effects)

Figure 2: Squadron Characteristics by Student Ability

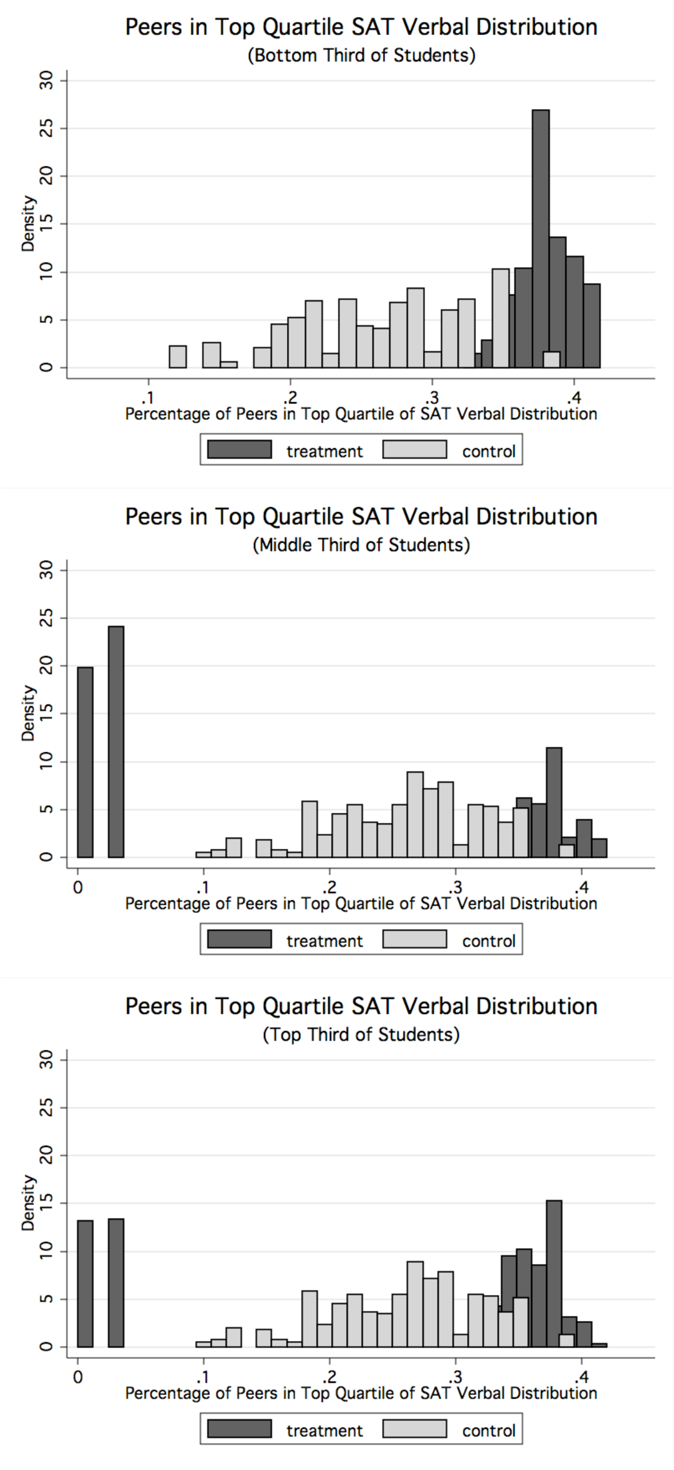


Figure 3: Distribution of Post-treatment Predicted GPA

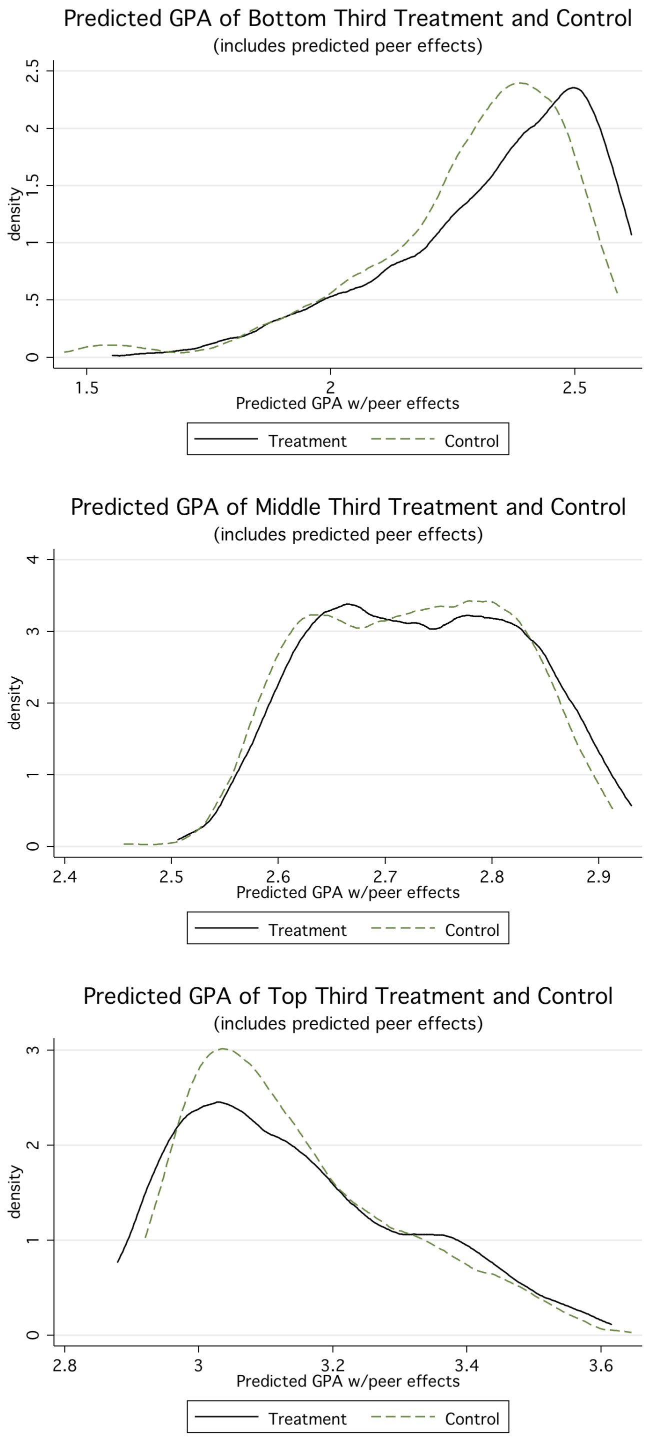


Table 1: Summary Statistics

Variable	Observational (2005-2010)	Treatment (2011-2012)	Control (2011-2012)
Grade Point Average	2.78	2.76	2.76
	(0.64)	(0.64)	(0.65)
Fraction Peers w. SAT Verbal Above 75th Percentile	0.28	0.26	0.27
	(0.08)	(0.06)	(0.16)
Fraction Peers w. SAT Verbal Below 25th Percentile	0.24	0.23	0.23
	(0.07)	(0.07)	(0.07)
SAT Verbal Score	634.40	633.00	632.60
	(68.20)	(66.00)	(67.00)
SAT Math Score	664.70	657.00	658.10
	(65.40)	(64.40)	(65.30)
Academic Composite Score	13.00	12.80	12.80
	(2.10)	(2.20)	(2.20)
Fitness Score	445.50	381.00	380.10
	(99.30)	(72.30)	(72.80)
Leadership Composite Score	17.30	17.30	17.30
	(1.80)	(1.70)	(1.70)
Recruited Athlete	0.25	0.23	0.23
	(0.43)	(0.42)	(0.42)
Attended Military Preparatory School	0.20	0.17	0.17
	(0.40)	(0.38)	(0.38)
Black	0.05	0.05	0.06
	(0.21)	(0.22)	(0.23)
Hispanic	0.07	0.08	0.08
	(0.25)	(0.28)	(0.27)
Asian	0.07	0.08	0.09
	(0.25)	(0.28)	(0.28)
Female	0.18	0.21	0.22
	(0.39)	(0.41)	(0.41)
Observations	14,024	2,422	2,412

Notes: Data include all students except those who left USAFA prior to the end of the first semester.

Table 2: Nonlinear Peer Effects: Pre-experimental Data

Variable	1	2		
Predicted Academic Ability	All	Bottom	Middle	Top
Fraction Peers w. SAT Verbal Above 75th Percentile	0.190**	0.481***	-0.112	0.215*
	(0.081)	(0.131)	(0.111)	(0.117)
Fraction Peers w. SAT Verbal Below 25th Percentile	-0.062	0.048	-0.193*	-0.017
	(0.081)	(0.126)	(0.116)	(0.120)
Observations	14,024	14,024		
R^2	0.344	0.345		
F-statistic: Restrictions	NA	3.562		
P-value		0.010		
F-statistic: Peer variables	3.797	3.484		
P-value	0.023	0.002		
F-statistic: Peer Effect 75th Top v Middle	NA	4.844		
P-value		0.028		
F-statistic: Peer Effect 75th Top v Bottom	NA	2.889		
P-value		0.090		
F-statistic: Peer Effect 75th Middle v Bottom	NA	14.820		
P-value		0.000		

We regress student level GPA for the semester on peer variables plus additional controls as follows:year and semester fixed effects and individual-level controls for students who are black, Hispanic, Asian, female, recruited athlete, and attended a preparatory school. Bottom, Middle, and Top groups are based on the distribution of predicted GPA using own pre-treatment characterisics. Data are for the two semesters of students' first year. Data are the observational data from the classes of 2005-2010. Robust standard errors in parentheses are clustered by class by squadron. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Table 3: Treatment and Control Randomization Checks

Variable	1	2	3
Sample	Class of 2011	Class of 2012	Classes of 2011 & 2012
SAT Verbal Score	0.021	0.001	0.009
	(0.03)	(0.02)	(0.02)
SAT Math Score	-0.003	0.024	0.01
	(0.03)	(0.03)	(0.02)
Academic Composite Score	-0.577	0.88	0.128
	(1.13)	(1.10)	(0.78)
Fitness Score	-0.021	0.014	-0.003
	(0.02)	(0.02)	(0.01)
Leadership Composite Score	1.088	0.31	0.729
	(0.81)	(0.83)	(0.58)
Recruited Athlete (0-1)	0.005	0.024	0.013
	(0.04)	(0.04)	(0.03)
Attended Military Preparatory School	0.061	-0.014	0.02
	(0.05)	(0.04)	(0.03)
Cadet is Black (0-1)	0.026	0.02	0.024
	(0.07)	(0.06)	(0.05)
Cadet is Hispanic (0-1)	0.003	0.023	0.012
	(0.06)	(0.05)	(0.04)
Cadet is Asian (0-1)	-0.002	0.045	0.018
	(0.05)	(0.05)	(0.04)
Female (0-1)	0.000	0.008	0.007
	(0.04)	(0.03)	(0.03)
Predicted GPA in Lowest 3rd of Class	0.002	0.04	0.017
	(0.05)	(0.05)	(0.04)
Predicted GPA in Top 3rd of Class	0.037	-0.051	-0.006
	(0.05)	(0.05)	(0.03)
Graduating Class is 2011	NA	NA	0.000
			(0.02)
Observations	1,314	1,391	2,705
R^2	0.004	0.003	0.001
F-statistic: All Variables	0.398	0.28	0.264
P-value	0.957	0.99	0.992

Notes: Data are the experimental cohorts of the classes of 2011-2012. We regress an indicator for treatment (versus control) group on a large set of pre-treatment variables. Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. SAT, Academic Composite, Fitness, and Leadership scores have been divided by 100

Table 4: Predicted Treatment Effects

	Predicted GPA		
Group	Bottom Third	Middle Third	Top Third
Treatment Group	2.342	2.734	3.145
	(0.206)	(0.095)	(0.171)
Control Group	2.287	2.725	3.143
	(0.206)	(0.092)	(0.153)
Predicted Treatment Effect	0.055***	0.009	0.001
(Treatment - Control)	(0.014)	(0.008)	(0.017)
Observations	903	901	901

We use the regression coefficients in Table 2 Column 2 to form predicted GPAs for the students in the treatment and control groups. The latter are in the classes of 2011-2012. Means and differences in means are reported above. Robust standard errors in parentheses are clustered by class by squadron. *** p<0.01, ** p<0.05, * p<0.1.

Table 5: Observed Treatment Effects

	1	2	3
			Top
Variables	Bottom Third	Middle Third	Third
Treatment Group Dummy	-0.054**	0.067**	-0.004
	(0.03)	(0.03)	(0.03)
SAT Verbal Score	0.105***	0.124***	0.111***
	(0.03)	(0.03)	(0.02)
SAT Math Score	0.274***	0.312***	0.243***
	(0.03)	(0.05)	(0.03)
Academic Composite Score	0.105***	0.105***	0.137***
	(0.01)	(0.02)	(0.01)
Fitness Score	0.051***	0.131***	0.102***
	(0.02)	(0.02)	(0.02)
Leadership Composite Score	0.014	-0.019**	0.006
	(0.01)	(0.01)	(0.01)
Recruited Athlete (0-1)	0.009	-0.013	-0.038
	(0.03)	(0.04)	(0.04)
Attended Military Preparatory School	-0.176***	-0.184***	-0.072
	(0.04)	(0.06)	(0.06)
Cadet is Black (0-1)	-0.096**	0.038	0.082
	(0.04)	(0.06)	(0.10)
Cadet is Hispanic (0-1)	-0.093**	0.019	-0.087
	(0.04)	(0.05)	(0.06)
Cadet is Asian (0-1)	-0.075	0.095**	-0.023
	(0.05)	(0.05)	(0.04)
Female (0-1)	-0.013	-0.025	-0.033
	(0.03)	(0.03)	(0.03)
Graduating Class is 2011	0.021	-0.031	-0.104***
	(0.03)	(0.03)	(0.03)
Observations	1,563	1,631	1,640
R^2	0.139	0.071	0.155

Notes: We take the experimental group (classes of 2011 and 2012) and regress own first and second semester GPA on a dummy for treatment status and own incoming characteristics. We stratify the sample by predicted GPA. The treatment was intended to raise the GPA of the least able students by assigning them to squadrons with a high fraction of peers with high verbal SAT scores. All regressions include class year and semester effects. Standard errors are clustered at the Class by Squadron level. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 6: Fully Interacted Peer Model

Variable		1	
Predicted Academic Ability	Bottom	Middle	Top
Fraction Peers w. SAT Verbal Above 75th Percentile	0.468***	-0.123	0.204*
	(0.131)	(0.109)	(0.120)
Fraction Peers w. SAT Verbal Below 25th Percentile	0.053	-0.181	0.001
	(0.126)	(0.118)	(0.119)
Fraction Peers w. SAT Math Above 75th Percentile	0.067	-0.089	-0.015
	(0.120)	(0.107)	(0.101)
Fraction Peers w. SAT Math Below 25th Percentile	-0.020	-0.130	-0.130
	(0.142)	(0.123)	(0.119)
Fraction Peers w. Academic Composite Above 75th Percentile	0.022	0.146	-0.088
	(0.133)	(0.126)	(0.116)
Fraction Peers w. Academic Composite Below 25th Percentile	0.073	0.103	-0.11
	(0.138)	(0.122)	(0.115)
Observations		14,024	
R^2		0.345	
F-statistic: All Peer variables		1.518	
P-value		0.079	
F-statistic: SAT Verbal Peer Variables		3.309	
P-value		0.003	
F-statistic: SAT Math Peer Variables		0.581	
P-value		0.745	
F-statistic: Academic Composite Peer Variables		0.433	
P-value		0.881	

We take the observational data from the classes of 2005-2010. We regress first or second semester GPA on six peer variables interacted with three categories of own incoming ability (predicted GPA). Robust standard errors in parentheses are clustered by class by squadron. *** p<0.01, ** p<0.05, * p<0.1. All specifications include year and semester fixed effects and individua-level controls for students who are black, Hispanic, Asian, female, recruited athlete, and attended a preparatory school. Bottom, Middle, and Top groups are based on the distribution of predicted GPA using own pre-treatment characterisics.

Figure 4: Power of the Experiment

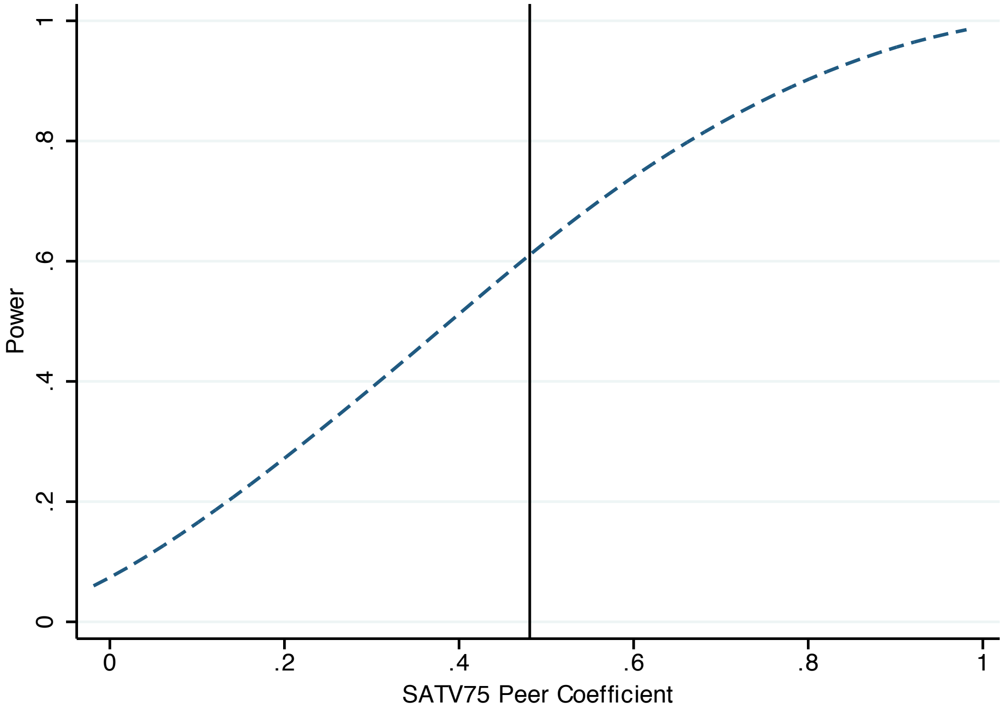
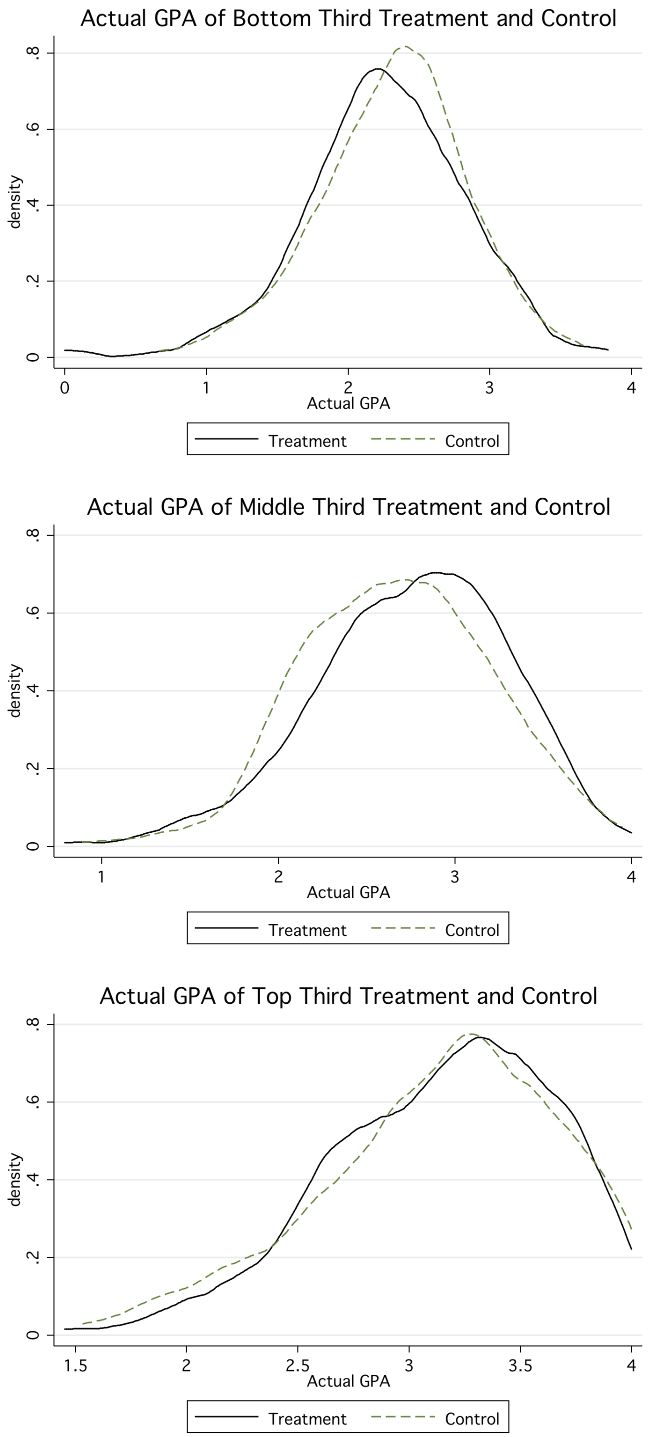


Figure 5: Distribution of Post-treatment Actual GPA



Actual GPA of Bottom Third Treatment and Control

Actual GPA of Middle Third Treatment and Control

Actual GPA of Top Third Treatment and Control

Figure 6: Distribution of Low Ability Peers

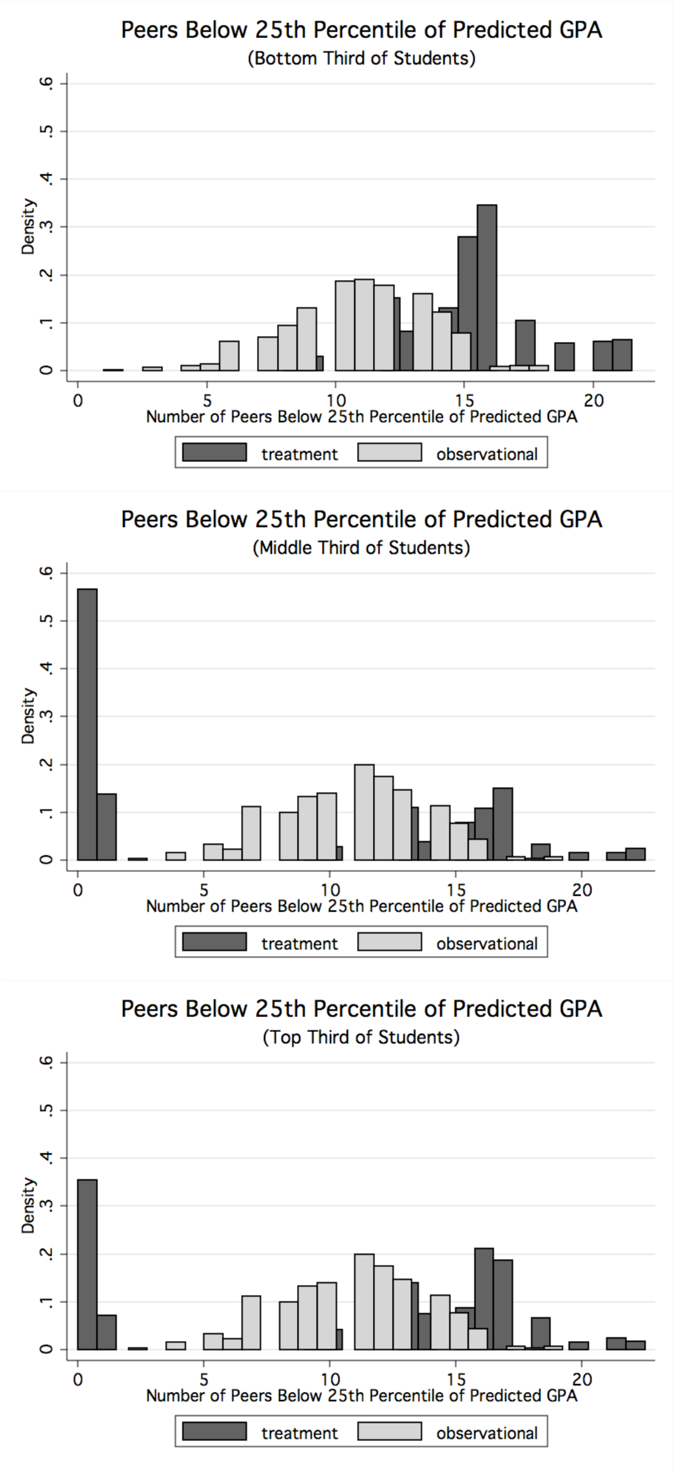
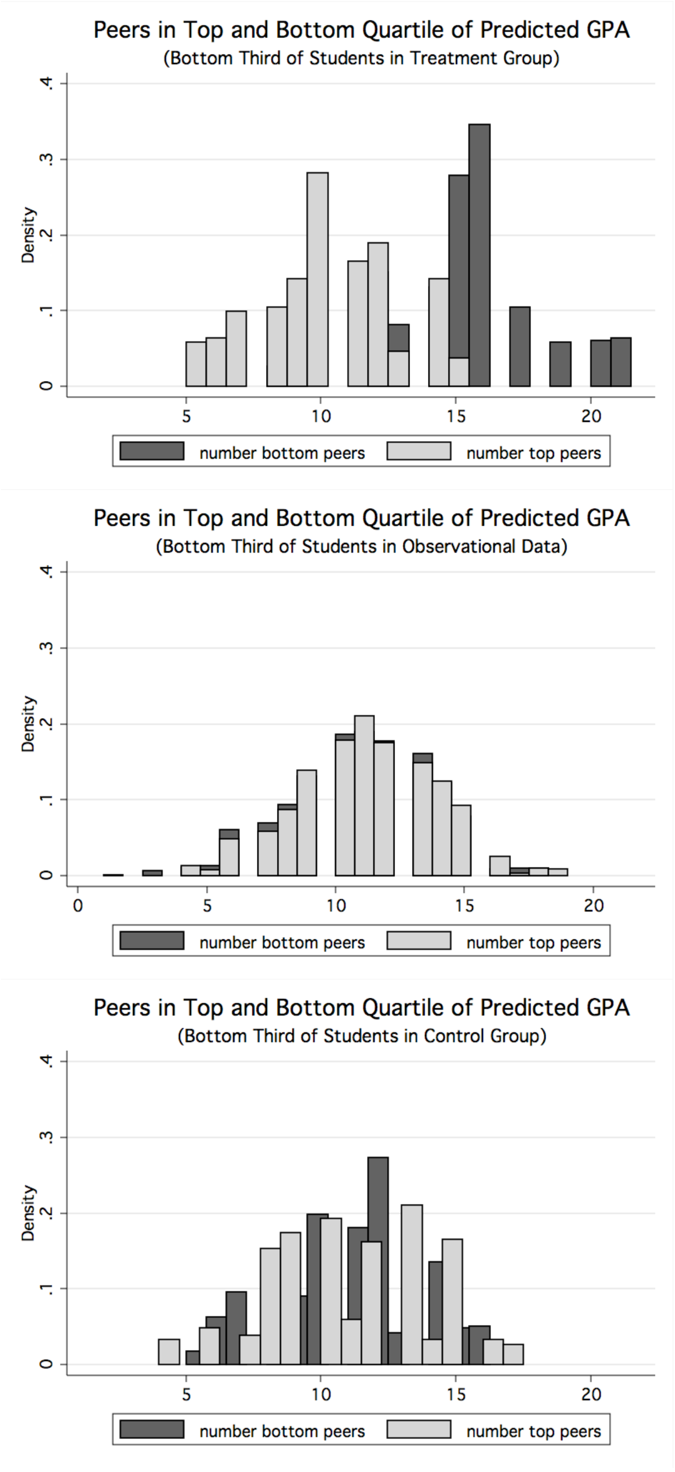


Figure 7: Distribution of Peer Ability



Peers in Top and Bottom Quartile of Predicted GPA
(Bottom Third of Students in Treatment Group)

Peers in Top and Bottom Quartile of Predicted GPA
(Bottom Third of Students in Observational Data)

Peers in Top and Bottom Quartile of Predicted GPA
(Bottom Third of Students in Control Group)

Table 7: Split Samples

Variable	1			2		
	Classes 2005-2007			Classes 2008-2010		
Predicted Academic Ability	Bottom	Middle	Top	Bottom	Middle	Top
Fraction Peers w. SAT Verbal Above 75th Percentile	0.528**	0.020	0.401**	0.423***	-0.207*	0.122
	(0.203)	(0.195)	(0.174)	(0.150)	(0.124)	(0.153)
Fraction Peers w. SAT Verbal Below 25th Percentile	-0.290	-0.312*	-0.098	0.294*	-0.107	0.081
	(0.181)	(0.173)	(0.162)	(0.158)	(0.144)	(0.181)
Observations	6,674			7,350		
R^2	0.348			0.351		

We take the observational data from the classes of 2005-2010. We regress own GPA on peer variables interacted with three categories of own ability (terciles of predicted GPA based on own characteristics). We split the sample into the earlier and later years of the data. Robust standard errors in parentheses are clustered by class by squadron. *** p<0.01, ** p<0.05, * p<0.1. All specifications include year and semester fixed effects and individua-level controls for students who are black, Hispanic, Asian, female, recruited athlete, and attended a preparatory school. Bottom, Middle, and Top groups are based on the distribution of predicted GPA using own pre-treatment characterisics.

Table 8: Peer Effects in the Control Group

Variable	1		
Predicted Academic Ability	Bottom	Middle	Top
Fraction Peers w. SAT Verbal Above 75th Percentile * Observational	0.480***	-0.111	0.215*
	(0.132)	(0.111)	(0.116)
Fraction Peers w. SAT Verbal Above 75th Percentile * Control Group	0.593*	0.001	0.483*
	(0.346)	(0.314)	(0.270)
Fraction Peers w. SAT Verbal Below 25th Percentile * Observational	0.054	-0.186	-0.013
	(0.127)	(0.115)	(0.120)
Fraction Peers w. SAT Verbal Below 25th Percentile * Control Group	-0.155	-0.507*	0.495
	(0.256)	(0.302)	(0.327)
Observations	16,446		
R^2	0.343		
F-statistic Peer 75th for Bottom Group: Observational v Control	0.093		
P-value	0.761		

We stack the observational data and control data and run our baseline peer effects specification as a single regression. The purpose is to test whether the peer effects coefficients differ between the observational group and control group. Robust standard errors in parentheses are clustered by class by squadron. *** p<0.01, ** p<0.05, * p<0.1. All specifications include year and semester fixed effects and individua-level controls for students who are black, Hispanic, Asian, female, recruited athlete, and attended a preparatory school. Bottom, Middle, and Top groups are based on the distribution of predicted GPA using own pre-treatment characterisics.

Table 9: Endogenous Peer Effects Model

Variable		1	
Predicted Academic Ability	Bottom	Middle	Top
Peer GPA * Observational	0.474***	0.346***	0.342***
	(0.051)	(0.050)	(0.048)
Peer GPA * Control	0.209**	0.439***	0.377***
	(0.096)	(0.107)	(0.129)
Peer GPA * Treatment	-0.015	0.146	0.219*
	(0.169)	(0.157)	(0.124)
Observations			18,858
R^2			0.353
F-statistic: Observational v Treatment	7.666	1.466	0.848
P-value	0.006	0.227	0.357
F-statistic: Control v Treatment	1.339	2.402	0.766
P-value	0.248	0.122	0.382

We stack the observational, control, and treatment data. We run the endogenous peer effects model (eg own outcome on peers' average outcomes). The purpose is to allow a test of whether the data generating process changed among the three different samples. Robust standard errors in parentheses are clustered by class by squadron. *** p<0.01, ** p<0.05, * p<0.1. All specifications include year and semester fixed effects and individua-level controls for students who are black, Hispanic, Asian, female, recruited athlete, and attended a preparatory school. Bottom, Middle, and Top groups are based on the distribution of predicted GPA using own pre-treatment characterisics.

Table 10: Effects of Bifurcation in the Observational Group

Variable	1	2	3	4
Fewer than 6 Middle Predicted GPA Students in Squadron	0.099*			
	(0.059)			
Fraction Peers in Bottom Predicted GPA Pred > 0.40 and Fraction Peers in Top of Predicted GPA > 0.40		0.143*		
		(0.081)		
Greater than 15 Low Predicted GPA Students in Squadron			0.020	
			(0.039)	
Fraction Peers in Bottom Predicted GPA in Fourth Quartile and Fraction Peers with high SAT Verbal in Fourth Quartile				0.150***
				(0.045)
Observations	4,638	4,638	4,638	4,638
R^2	0.096	0.095	0.095	0.097

We regress own GPA on indicators for various measures of bifurcation for students with low predicted GPA in the observational group. Robust standard errors in parentheses are clustered by class by squadron. *** p<0.01, ** p<0.05, * p<0.1. All specifications include year and semester fixed effects and individua-level controls for students who are black, Hispanic, Asian, female, recruited athlete, and attended a preparatory school.

Table 11: Evidence of Bifurcation in the Treatment Group: Roommate Choices

Panel A. Randomly Assigned First Semester Roommates	*Treatment Group*	*Control Group*	*Treatment Group*	*Control Group*
Dependent variable	Own Predicted GPA in Bottom Third		Own First Semester GPA	
	1	2	3	4
Roommate Predicted GPA in Bottom Third	-0.036	-0.031		
(0-1)	(0.096)	(0.086)		
Roommate First Semester GPA			0.045	-0.029
(average if two roommates)			(0.077)	(0.095)
Observations	335	468	329	458
R-squared	0.039	0.042	0.068	0.051

Panel B. Self-Selected Second Semester Roommates	*Treatment Group*	*Control Group*	*Treatment Group*	*Control Group*
Dependent variable	Own First Semester GPA		Own Second Semester GPA	
	1	2	3	4
Roommate First Semester GPA	0.162*	-0.004		
(average if two roommates)	(0.091)	(0.113)		
Roommate Second Semester GPA			0.289***	-0.054
(average if two roommates)			(0.092)	(0.098)
Observations	344	428	342	476
R-squared	0.064	0.027	0.104	0.049

We regress own attributes on roommate attributs separately for the treatment and control group. All specifications include squadron, year, and semester fixed effects and individua-level controls for students who are black, Hispanic, Asian, female, recruited athlete, and attended a preparatory school. Robust standard errors in parentheses are clustered by class by squadron. *** p<0.01, ** p<0.05, * p<0.1. Data come from USAFA room assignment files and are only available for the graduating class of 2012.

Table 12: Evidence of Bifurcation in the Treatment Group: Study Partner Survey

	1	2	3	4	5	6
	Low GPA	Middle GPA	High GPA	High SAT	Low GPA	High SAT
	Study	Study	Study	Verbal Study	Friends	Verbal
	Partners	Parnters	Parnters	Parnters		Friends
Treatment Group* Bottom 3rd Predicted GPA	0.524*	-1.105***	-0.001	0.124	0.658**	-0.230
	(0.291)	(0.207)	(0.273)	(0.246)	(0.274)	(0.225)
Treatment	-0.119	0.260**	0.036	-0.074	-0.222	-0.010
	(0.128)	(0.115)	(0.131)	(0.144)	(0.147)	(0.128)
Predicted GPA in Lowest 3rd of Class	0.147	0.224	0.151	0.189	0.275	0.411**
	(0.194)	(0.187)	(0.239)	(0.228)	(0.231)	(0.187)
Predicted GPA in Top 3rd of Class	-0.044	-0.136	-0.168	-0.257	-0.095	-0.092
	(0.144)	(0.179)	(0.188)	(0.169)	(0.146)	(0.149)
Observations	559	559	559	559	559	559
R-squared	0.127	0.119	0.040	0.169	0.136	0.149

We regress self identified study partner and friends characteristics on whether the individual is in the treatment group. Robust standard errors in parentheses are clustered by class by squadron. *** p<0.01, ** p<0.05, * p<0.1. All specifications include year and semester fixed effects and individua-level controls for students who are black, Hispanic, Asian, female, recruited athlete, and attended a preparatory school. Data come from a retrospective survey conducted at USAFA during the spring term of 2010. The survey asked each student to name up to five study partners and five friends. Reponse rate was approximately 25 percent.