

From Natural Variation to Optimal Policy: A Cautionary Tale in How Not to Improve Student Outcomes *

Scott E. Carrell
UC Davis and NBER

Bruce I. Sacerdote
Dartmouth College and NBER

James E. West
USAF Academy

September 30, 2010

Abstract

We begin with peer effects estimates from randomly assigned peer groups at the United States Air Force Academy. We then take subsequent cohorts of entering freshmen and assign half of the students to peer groups (squadrons) in a way intended to maximize the academic performance of students with the lowest incoming ability. We find a *negative* and statistically significant treatment effect for the students we intended to help. We explore three possible explanations for this perverse finding including the hypothesis (which we can reject) that the original findings were spurious. We show that our “optimal” assignment mechanism created bifurcated squadrons that had a social dynamic entirely different from most squadrons seen in the observational data. Our results suggest that even in a well understood and self contained environment, using reduced-form estimates to make out-of-sample policy predictions can lead to unanticipated and potentially negative consequences.

1 Introduction

Peer effects have been widely studied in the economics literature due to the perceived importance peers play in the production of outcomes. Previous studies have focused almost exclusively on the

*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: This paper was formerly circulated under the title, “Beware of Economists Bearing Reduced Forms? An Experiment in How Not To Improve Student Outcomes.”

identification of peer effects and have only hinted at the potential policy implications of the results.¹ Recent econometric studies on assortative matching by Graham, Imbens, and Ridder (2009), and Bhattacharya (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. This policy matches the original USAFA method for assigning students to squadrons. The other half of students (the treatment group) were sorted into squadrons in a manner intended to maximize the academic performance 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 higher ability student's grades 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 . These results show the important role that peers play in the education production process; however, they also highlight the potential danger in using reduced form estimates to actively sort individuals into peer groups without a thorough understanding of the underlying mechanisms that drive the social interactions.

We explore three possible explanations for this perverse finding. One 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 second hypothesis is that

¹For recent studies in higher education see: (Sacerdote 2001, Zimmerman 2003, Stinebrickner and Stinebrickner 2006, Carrell, Fullerton, and West 2009, Carrell, Malmstrom, and West 2008, Foster 2006, Lyle 2007).

²If 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.

³This objective function was determined by USAFA senior leadership. With roughly a 20 percent academic probation rate at USAFA, the senior leadership's objective at USAFA was to use this study to try to reduce the number of freshman students who are placed on academic probation.

the data generating process changed in a fundamental way. The data point to a third hypothesis which is that our “optimally” sorted squadrons had more extreme variation in the percent of high and low ability students (i.e. bifurcation) and that such squadrons have a unique social dynamic rarely seen in the observational data. High and low ability students in the treatment squadrons appear to have segregated themselves into separate social networks. This likely occurred because the assignment algorithm sorted at the extremes of the observed data. For the middle ability students, evidence suggests that the positive treatment effect occurred because these students were protected or segregated from the low ability students and put into more homogeneous peer groups. This finding is consistent with recent evidence on ability grouping and tracing by Duflo, Dupla, and Kremer (2008).

Results from this study are significant for several reasons. First, to our knowledge, 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, and Ridder (2009). Second, the study is fairly unique in its use of historical observational data to infer optimal policy and then explicitly implement and test the efficacy of the policy in an experiment. Third, they highlight the significant role that peers play in the education production process. Finally, the unexpected results of the experiment suggest that using reduced form effects 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 2 presents the data and estimates the nonlinear peer effects at USAFA. Section 3 describes the squadron sorting mechanism. Section 4 describes the experimental design and provides simulated results. Section 5 presents results from the experiment. Section 6 explores reasons for the experiments’ unexpected findings. Section 7 concludes.

2 Data

2.1 The Dataset

As described in Carrell, Fullerton, and West (2009), we use the random assignment of USAFA students to a peer group (i.e. a squadron), to identify peer effects in academic performance. We estimate reduced form equations in which we regress individual outcomes on pre-treatment

variables. The use of pre-treatment variables avoids simultaneous equation bias or the reflection problem common in estimation of peer effects (Manski 1993, Sacerdote 2001).

Our pre-treatment (i.e. observational) dataset includes all students in the USAFA graduating classes of 2005 through 2010. 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, and West (2009) and Carrell and 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.

2.2 Methods

Carrell, Fullerton, and West (2009) found large and statistically significant reduced form peer effects at USAFA. Specifically, they found student academic performance increased significantly with the average peer SAT verbal scores in the squadron. Additionally, Carrell, Fullerton, and West (2009) found evidence of nonlinear effects in which low ability 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 incoming ability. Specifically, we estimate separate peer coefficients for each third of the own academic ability distribution.⁴

⁴Each student's academic ability is determined by computing in-sample *predicted* GPA. Low, medium, or high ability is based on which third of the academic distribution their predicted GPA falls.

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

Specifically, we estimate the following equation:

$$GPA_{isqrt} = \phi_0 + \phi_1 t \frac{\sum_{k \neq i} X_{ksqr}}{n_{sc} - 1} + \beta X_{isqrt} + \epsilon_{isqrt} \quad (1)$$

where GPA_{isqrt} is the freshman fall semester GPA for individual i in squadron s , graduating class c , semester r , and of academic ability t . $\frac{\sum_{k \neq i} X_{ksqr}}{n_{sc} - 1}$ are the fraction of peers (excluding individual i) in the top and bottom quartile of the SAT verbal distribution. X_{isqrt} 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. ϵ_{isqrt} is the error term. 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.

We estimate equation (1) 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 own incoming ability 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, and 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 incoming ability 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.

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

⁶For 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.

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

The results suggest that low ability 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 Superintendent 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. Our objective function was to improve the grades of the lowest one-third of incoming ability students.

3 Sorting Methodology

To optimally sort students into squadrons, we draw on recent work on assortative matching by Bhattacharya (2009) and Graham, Imbens, and 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) as shown in Table 2, Specification 2.

$$GPA_{i,s,t} = X_i\beta + Z_{i,s}\gamma_t + \epsilon_{i,s,t} \quad (2)$$

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

$$\max_{p_{i,s}, \lambda_i, \delta_s} \left[\min_{i \in I_l} \left\{ \sum_{s=1}^{20} p_{i,s} \frac{1}{N_s} \sum_{j \neq i} p_{j,s} X_{j,s} \gamma_l - \lambda_i \left(1 - \sum_{s=1}^{20} p_{i,s} \right) - \delta_s \left(N_s - \sum_{i=1}^{650} p_{i,s} \right) \right\} \right] \quad (3)$$

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 $p_{i,s}$.

4 Experimental Design

4.1 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).

⁸Upon 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.

⁹XpressMP was provided to us by FICO under their Academic Partners Program.

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

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

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 ability 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 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 power of the experiment for different possible levels of the key peer coefficient,

¹¹Specifically, 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.

¹²The 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 AOC’s, 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 first year of the experiment.

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 a stochastic error terms drawn from a normal distribution with mean zero and variance 0.269.¹³ We ran the simulation 1000 times and asked how often we detected the positive treatment effect for low ability students implied by Table 2 column 2. Results show the power of the experiment at a 0.05–level (one-tailed) of significance is 76.3. That is, in 1,000 draws, 763 of the treatment regressions were positive and statistically significant. Figure 4 plots the power curve at various estimated sizes of the key peer coefficient, namely the effect of percent high SAT Verbal on the low ability students. The vertical line represents our predicted effect given the peer coefficient of 0.481.

5 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 had the exact opposite effect intended. That is, 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 ability distribution.

One possibility is the negative treatment effect is simply due to sampling variation. That is, a positive treatment effect exists, but the noise in estimating GPA masked it. As one simple check of this hypothesis, we conducted a Monte Carlo simulation by adding a stochastic error term drawn from a normal distribution to determine the probability of detecting a negative and significant treatment effect when the predicted treatment effect actually exists. Results from this simulation found none of the 1,000 simulated draws to be negative and statistically significant at the 0.10–percent level. Hence, we conclude the effect is not likely due to sampling variation.

6 Why the Unexpected Results?

Given the unanticipated findings of the experiment, we next explore three possible explanations. First, 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

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

different functional forms. Second, 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? Because we have a control group we are able to ask whether peer effects in the randomly assigned control group resemble those found in the earlier cohorts in the observational data. Third, 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.

6.1 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 we allow for the proportion of peers in the top or bottom of these distributions to each have a separate effect. And we allow these six possible effects to vary by own ability (three groups) yielding a total of eighteen peer coefficients. We run this fully interacted specification to test whether the peer coefficients are jointly significant. Our logic is that a conservative test for the existence of peer effects is to require joint significance of all possible peer variables, as opposed to limiting the specification to right hand side variables already found to be statistically significant in other explorations of the data. 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 we estimated 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 artifact or the result of a failure to correct standard errors for multiple hypothesis tests.

6.2 Did the Process Change?

Although the peer effects in the observational data appear to be robust, another possibility is 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 student attitudes and behavior may have changed over time. To test this hypothesis, we examine the magnitude and significance of the reduced-form peer effects in the randomly assigned control group. For ease of hypothesis testing, we stack the data and include both the control group and the observational data in the regression and allow for separate peer effects by data source (control versus observational).¹⁴ Table 8 presents these findings. For low ability students, 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 in the control group. Importantly, the effect is similar to that seen in observational data (0.480) and these two effects are statistically indistinguishable ($p = 0.761$). 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.

As a second test, we estimate the endogenous peer effects model in which we regress own GPA on concurrent peer GPA. Naturally we recognize the endogenous model is subject to the reflection problem and other criticisms (e.g. correlated shocks) which make interpretation difficult. However, the endogenous model offers the possibility of more precision since it incorporates peer effects that are working through contemporaneous as well as background channels. Prior studies find that endogenous effects are generally larger and more precisely estimated than contextual (background) effects (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.

¹⁴We 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.

6.3 Did the Peer Dynamics in the Treatment Groups Change?

A third possible explanation for the observed negative treatment effect is that the extreme variation created in the treatment squadrons caused the peer dynamics in the treatment squadrons to change. As we previously showed 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 high and low ability students grouped together and 2) homogenous squadrons consisting of middle ability students.

Regression results in Table 10 tests this assertion more formally. We take the treatment and control groups and regress characteristics of the squadron on a dummy for treatment status. We interact the treatment dummy with the three groups for own ability. For low ability students, the treatment raised the fraction of high SAT verbal peers and low SAT verbal peers. In column 3, the dependent variable is the number of low ability students in one's squadron (ability is measured as predicted GPA). On average the treatment assigned each low ability student an extra 4.5 low ability peers (out of 30 peers) relative to the control. These low ability peers took the place of middle ability peers. For middle ability students, the treatment squadrons had significantly fewer low ability peers.

To get a better understanding of whether the bifurcation of low ability student in treatment squadrons may have caused the negative treatment effect, we next examine the effect of bifurcation in the pre-experiment observational data. In Table 11 we regress GPA for students with low predicted grades on various indicator variables for the presence of bifurcation. Across all four of our measures of bifurcation, low ability performed *better* than average, with three of the four measures significant at the 5-percent level. These results show the negative treatment effect we observe was not likely foreseeable in the observational data. More importantly, these results indicate that the role peers play in the education production process is a complex social process that is not sufficiently described by a simple reduced form model to enable reliable out-of-sample forecasts such as our experiment .

7 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 . Analysis suggests that the unexpected negative treatment effect for the lower-ability students may have occurred due to changes in peer dynamics as a result of the bifurcated squadrons created by our sorting algorithm. The positive treatment effect likely occurred because the middle ability students were “segregated” from the lowest ability student. A key next step for this research will be to find independent evidence to confirm that social dynamics in the treatment squadrons did in fact change when we altered the composition of the squadrons.

Results from this study show that using reduced form effects to conduct out-of-sample policy predictions may lead to unanticipated outcomes.

References

- BHATTACHARYA, D. (2009): “Inferring Optimal Peer Assignment from Experimental Data,” *Journal of the American Statistical Association*, 104(486), 486–500.
- CARRELL, S., AND J. WEST (2010): “Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors,” *Journal of Political Economy*, 118(3), 409–432.
- CARRELL, S. E., R. L. FULLERTON, AND J. E. WEST (2009): “Does Your Cohort Matter? Estimating Peer Effects in College Achievement,” *Journal of Labor Economics*, 27(3), 439–464.

- CARRELL, S. E., F. V. MALMSTROM, AND J. E. WEST (2008): “Peer Effects in Academic Cheating,” *Journal of Human Resources*, 43(1), 173–207.
- DUFLO, E., P. DUPA, AND M. KREMER (2008): “Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya,” (14475).
- FOSTER, G. (2006): “It’s not your peers, and it’s not your friends: Some progress toward understanding the educational peer effect mechanism,” *Journal of Public Economics*, 90(8-9), 1455–1475.
- GRAHAM, B. S., G. W. IMBENS, AND G. RIDDER (2009): “Complementarity and Aggregate Implications of Assortative Matching: A Nonparametric Analysis,” Working Paper 14860, National Bureau of Economic Research.
- LYLE, D. 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, C. F. (1993): “Identification and Endogenous Social Effects: The Reflection Problem,” *Review of Economic Studies*, 60(3), 531–42.
- SACERDOTE, B. L. (2001): “Peer Effects with Random Assignment: Results for Dartmouth Roommates,” *Quarterly Journal of Economics*, 116(2), 681–704.
- STINEBRICKNER, R., AND T. 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), 1435–54.
- ZIMMERMAN, D. 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

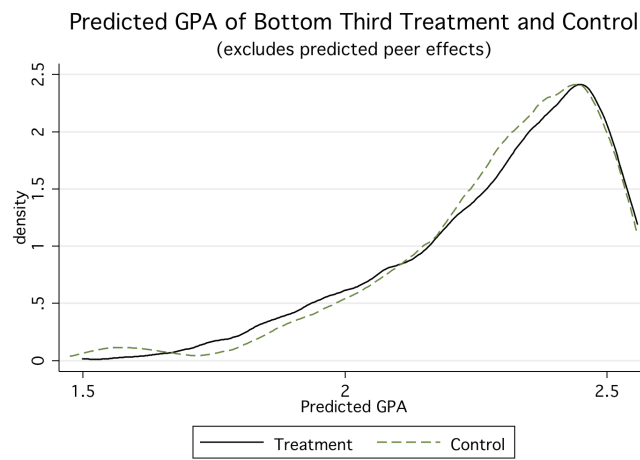
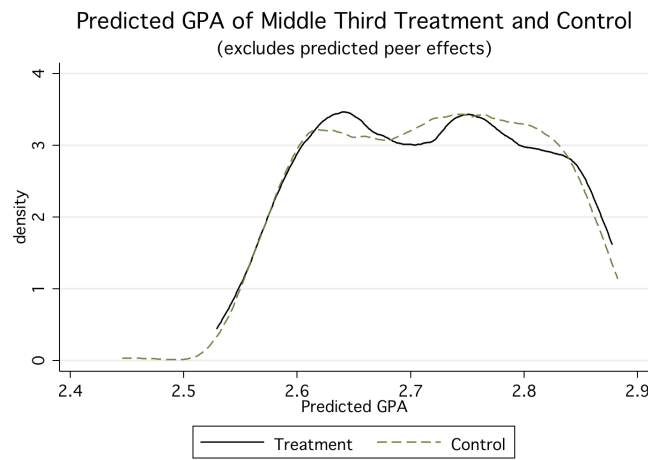
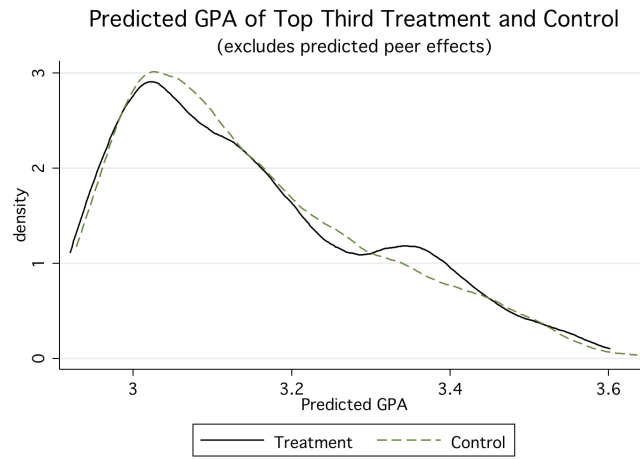


Figure 2: Squadron Characteristics by Student Ability

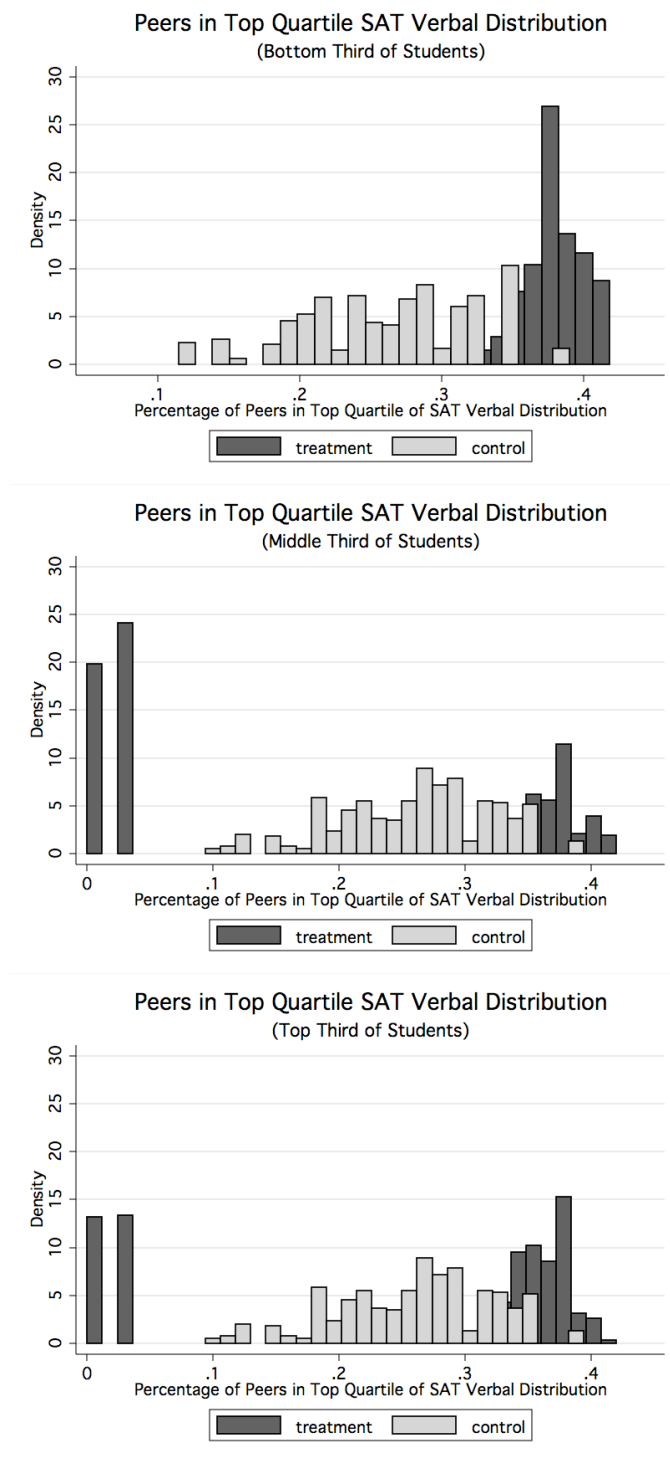


Figure 3: Distribution of Post-treatment Predicted GPA

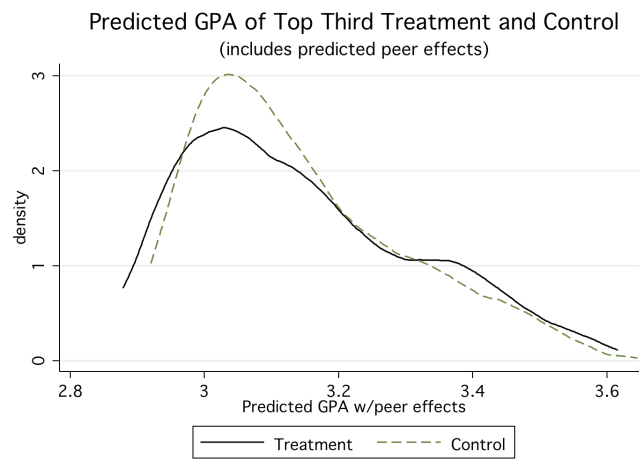
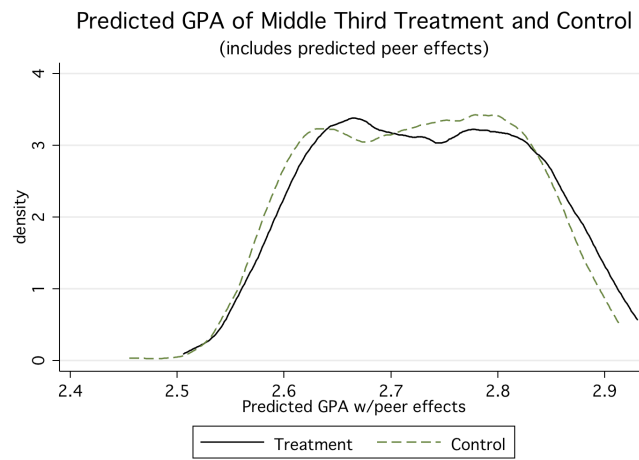
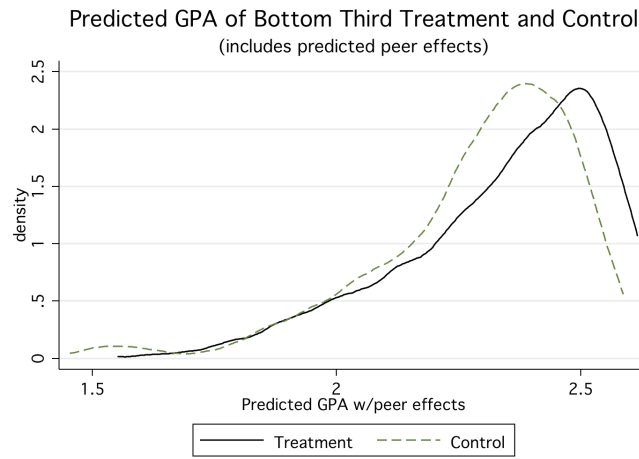


Figure 4: Power of the Experiment

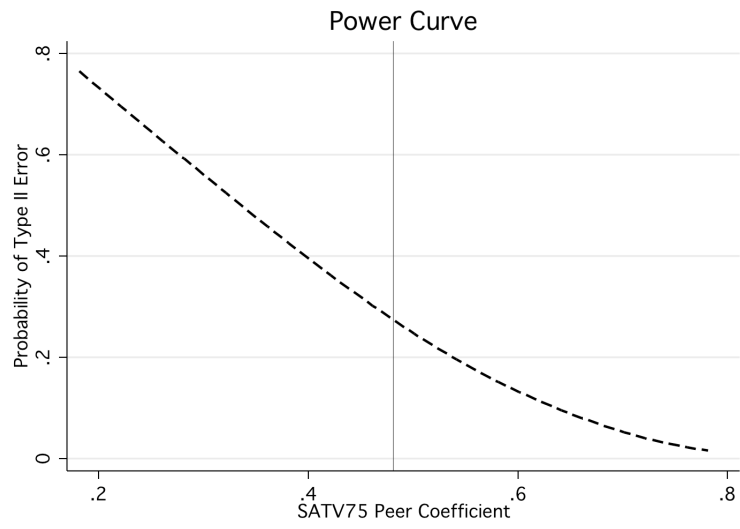


Figure 5: Distribution of Post-treatment Actual GPA

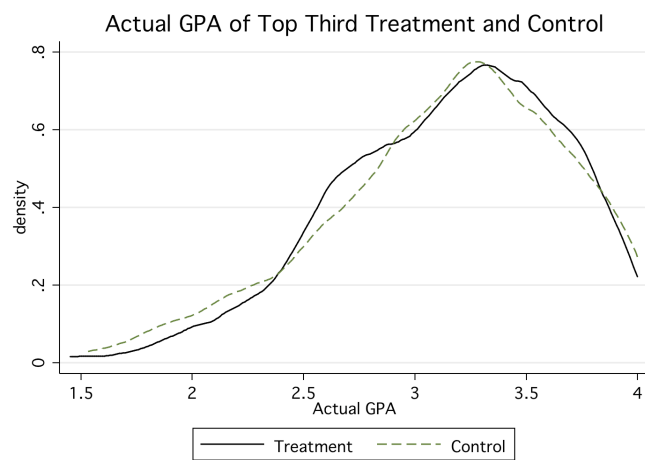
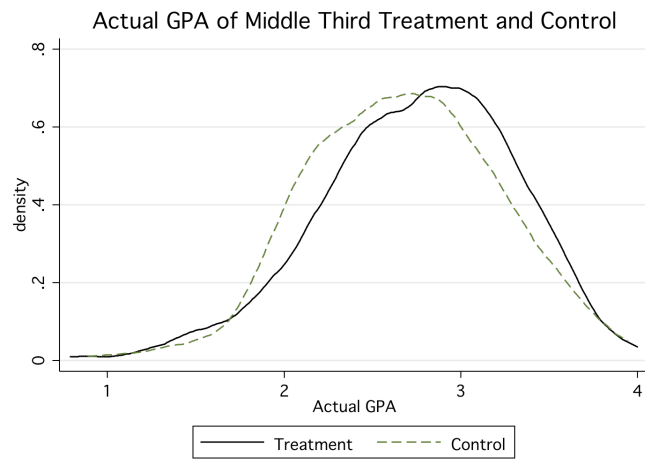
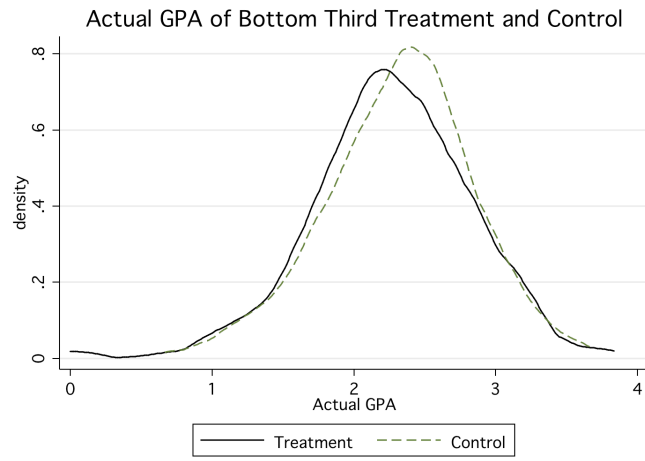


Figure 6: Distribution of Low Ability Peers

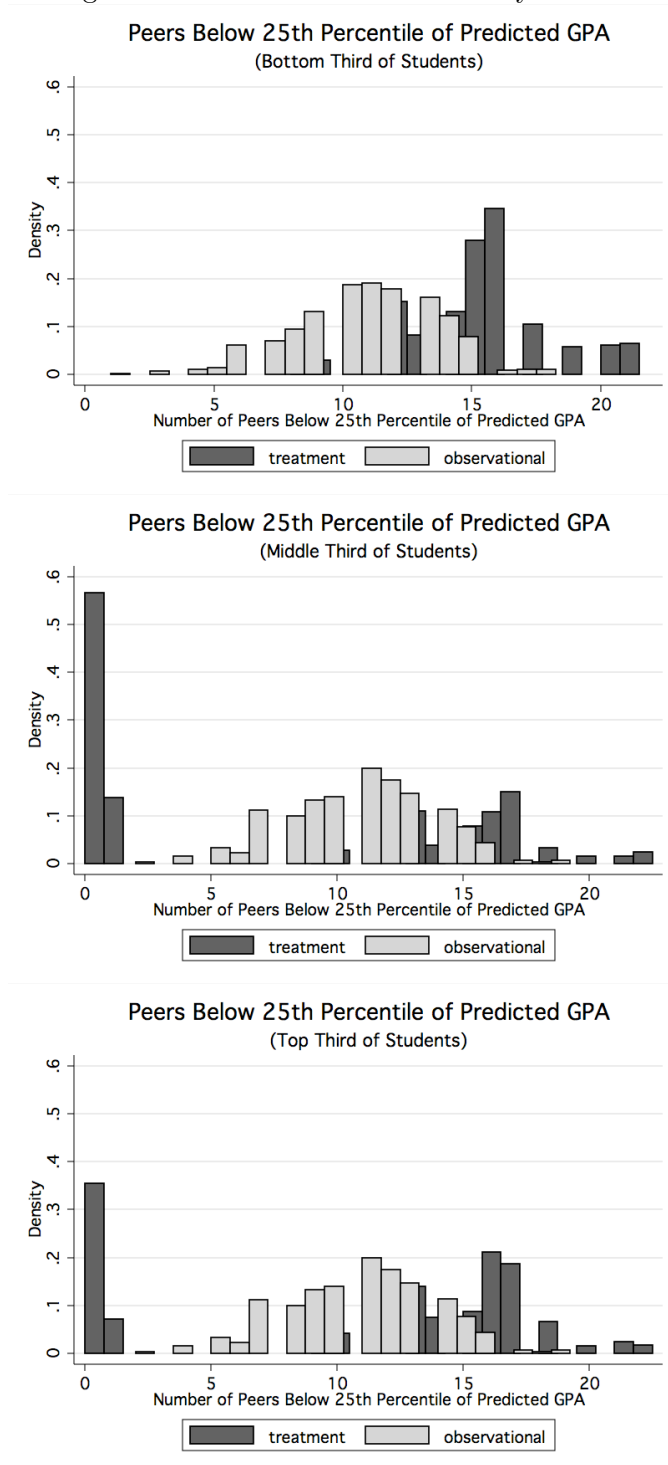


Figure 7: Distribution of Peer Ability

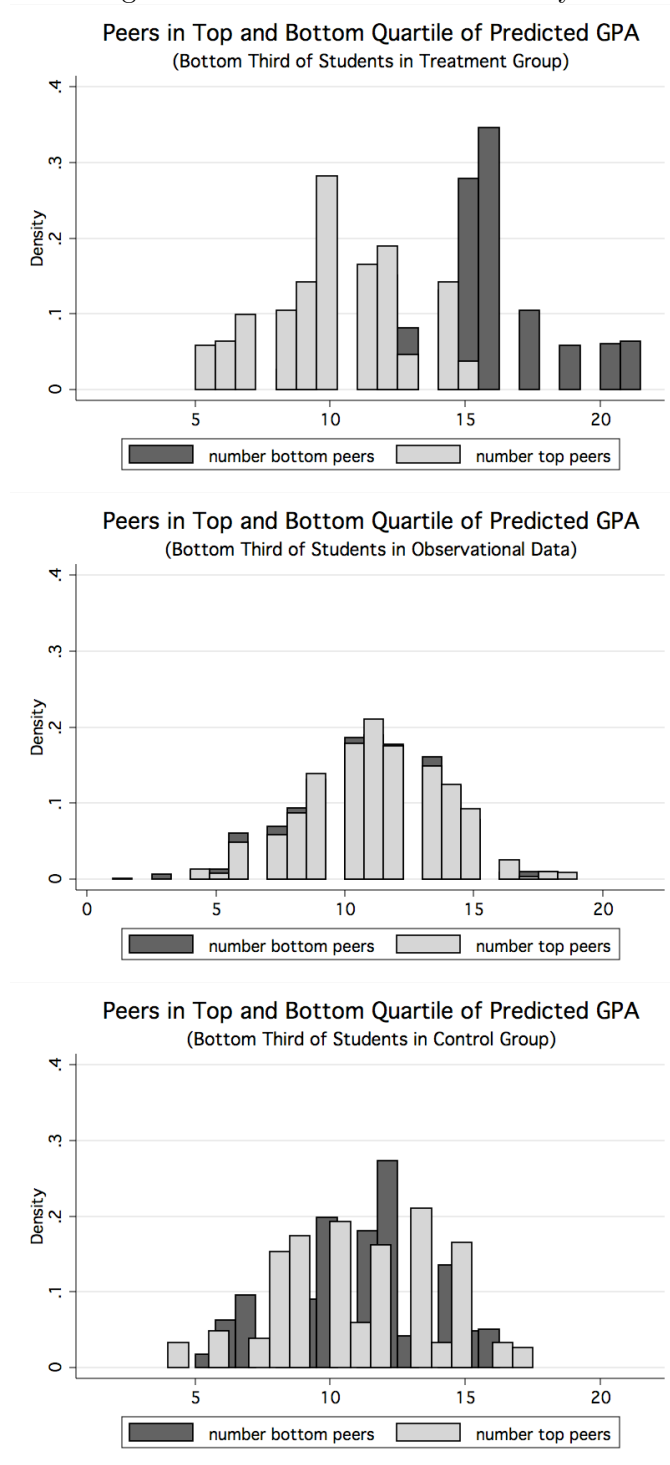


Table 1: Summary Statistics

Variable	Observational (2005-2010)	Treatment (2011- 2012)	Control (2011-2012)
Grade Point Average	2.78 (0.64)	2.76 (0.64)	2.76 (0.65)
Fraction Peers w. SAT Verbal Above 75th Percentile	0.28 (0.08)	0.26 (0.06)	0.27 (0.16)
Fraction Peers w. SAT Verbal Below 25th Percentile	0.24 (0.07)	0.23 (0.07)	0.23 (0.07)
SAT Verbal Score	634.40 (68.20)	633.00 (66.00)	632.60 (67.00)
SAT Math Score	664.70 (65.40)	657.00 (64.40)	658.10 (65.30)
Academic Composite Score	13.00 (2.10)	12.80 (2.20)	12.80 (2.20)
Fitness Score	445.50 (99.30)	381.00 (72.30)	380.10 (72.80)
Leadership Composite Score	17.30 (1.80)	17.30 (1.70)	17.30 (1.70)
Recruited Athlete	0.25 (0.43)	0.23 (0.42)	0.23 (0.42)
Attended Military Preparatory School	0.20 (0.40)	0.17 (0.38)	0.17 (0.38)
Black	0.05 (0.21)	0.05 (0.22)	0.06 (0.23)
Hispanic	0.07 (0.25)	0.08 (0.28)	0.08 (0.27)
Asian	0.07 (0.25)	0.08 (0.28)	0.09 (0.28)
Female	0.18 (0.39)	0.21 (0.41)	0.22 (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		
<u>Predicted Academic Ability</u>	<u>All</u>	<u>Bottom</u>	<u>Middle</u>	<u>Top</u>
Fraction Peers w. SAT Verbal Above 75th Percentile	0.190** (0.081)	0.481*** (0.131)	-0.112 (0.111)	0.215* (0.117)
Fraction Peers w. SAT Verbal Below 25th Percentile	-0.062 (0.081)	0.048 (0.126)	-0.193* (0.116)	-0.017 (0.120)
Observations	14,024	14,024		
R ²	0.344	0.345		
F-statistic: Restrictions	NA	3.562		
P-value	NA	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	NA	0.028		
F-statistic: Peer Effect 75th Top v Bottom	NA	2.889		
P-value	NA	0.090		
F-statistic: Peer Effect 75th Middle v Bottom	NA	14.820		
P-value	NA	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 characteristics. 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 (0.03)	0.001 (0.02)	0.009 (0.02)
SAT Math Score	-0.003 (0.03)	0.024 (0.03)	0.01 (0.02)
Academic Composite Score	-0.577 (1.13)	0.88 (1.10)	0.128 (0.78)
Fitness Score	-0.021 (0.02)	0.014 (0.02)	-0.003 (0.01)
Leadership Composite Score	1.088 (0.81)	0.31 (0.83)	0.729 (0.58)
Recruited Athlete (0-1)	0.005 (0.04)	0.024 (0.04)	0.013 (0.03)
Attended Military Preparatory School	0.061 (0.05)	-0.014 (0.04)	0.02 (0.03)
Cadet is Black (0-1)	0.026 (0.07)	0.02 (0.06)	0.024 (0.05)
Cadet is Hispanic (0-1)	0.003 (0.06)	0.023 (0.05)	0.012 (0.04)
Cadet is Asian (0-1)	-0.002 (0.05)	0.045 (0.05)	0.018 (0.04)
Female (0-1)	0.000 (0.04)	0.008 (0.03)	0.007 (0.03)
Predicted GPA in Lowest 3rd of Class	0.002 (0.05)	0.04 (0.05)	0.017 (0.04)
Predicted GPA in Top 3rd of Class	0.037 (0.05)	-0.051 (0.05)	-0.006 (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

Group	Predicted GPA		
	Bottom Third	Middle Third	Top Third
Treatment Group	2.342 (0.206)	2.734 (0.095)	3.145 (0.171)
Control Group	2.287 (0.206)	2.725 (0.092)	3.143 (0.153)
Predicted Treatment Effect (Treatment - Control)	0.055*** (0.014)	0.009 (0.008)	0.001 (0.017)
Observations	1,352	1,353	2,705

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
Variables	Bottom Third	Middle Third	Top Third
Treatment Group Dummy	-0.054** (0.03)	0.067** (0.03)	-0.004 (0.03)
SAT Verbal Score	0.105*** (0.03)	0.124*** (0.03)	0.111*** (0.02)
SAT Math Score	0.274*** (0.03)	0.312*** (0.05)	0.243*** (0.03)
Academic Composite Score	0.105*** (0.01)	0.105*** (0.02)	0.137*** (0.01)
Fitness Score	0.051*** (0.02)	0.131*** (0.02)	0.102*** (0.02)
Leadership Composite Score	0.014 (0.01)	-0.019** (0.01)	0.006 (0.01)
Recruited Athlete (0-1)	0.009 (0.03)	-0.013 (0.04)	-0.038 (0.04)
Attended Military Preparatory School	-0.176*** (0.04)	-0.184*** (0.06)	-0.072 (0.06)
Cadet is Black (0-1)	-0.096** (0.04)	0.038 (0.06)	0.082 (0.10)
Cadet is Hispanic (0-1)	-0.093** (0.04)	0.019 (0.05)	-0.087 (0.06)
Cadet is Asian (0-1)	-0.075 (0.05)	0.095** (0.05)	-0.023 (0.04)
Female (0-1)	-0.013 (0.03)	-0.025 (0.03)	-0.033 (0.03)
Graduating Class is 2011	0.021 (0.03)	-0.031 (0.03)	-0.104*** (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		
	Bottom	Middle	Top
<u>Predicted Academic Ability</u>			
Fraction Peers w. SAT Verbal Above 75th Percentile	0.468*** (0.131)	-0.123 (0.109)	0.204* (0.120)
Fraction Peers w. SAT Verbal Below 25th Percentile	0.053 (0.126)	-0.181 (0.118)	0.001 (0.119)
Fraction Peers w. SAT Math Above 75th Percentile	0.067 (0.120)	-0.089 (0.107)	-0.015 (0.101)
Fraction Peers w. SAT Math Below 25th Percentile	-0.020 (0.142)	-0.130 (0.123)	-0.130 (0.119)
Fraction Peers w. Academic Composite Above 75th Percentile	0.022 (0.133)	0.146 (0.126)	-0.088 (0.116)
Fraction Peers w. Academic Composite Below 25th Percentile	0.073 (0.138)	0.103 (0.122)	-0.11 (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 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 characteristics.

Table 7: Split Samples

Variable	1			2		
	Classes 2005-2007			Classes 2008-2010		
<u>Predicted Academic Ability</u>	<u>Bottom</u>	<u>Middle</u>	<u>Top</u>	<u>Bottom</u>	<u>Middle</u>	<u>Top</u>
Fraction Peers w. SAT Verbal Above 75th Percentile	0.528** (0.203)	0.020 (0.195)	0.401** (0.174)	0.423*** (0.150)	-0.207* (0.124)	0.122 (0.153)
Fraction Peers w. SAT Verbal Below 25th Percentile	-0.290 (0.181)	-0.312* (0.173)	-0.098 (0.162)	0.294* (0.158)	-0.107 (0.144)	0.081 (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 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 characteristics.

Table 8: Peer Effects in the Control Group

Variable	1		
	Bottom	Middle	Top
<u>Predicted Academic Ability</u>			
Fraction Peers w. SAT Verbal Above 75th Percentile * Observational	0.480*** (0.132)	-0.111 (0.111)	0.215* (0.116)
Fraction Peers w. SAT Verbal Above 75th Percentile * Control Group	0.593* (0.346)	0.001 (0.314)	0.483* (0.270)
Fraction Peers w. SAT Verbal Below 25th Percentile * Observational	0.054 (0.127)	-0.186 (0.115)	-0.013 (0.120)
Fraction Peers w. SAT Verbal Below 25th Percentile * Control Group	-0.155 (0.256)	-0.507* (0.302)	0.495 (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 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 characteristics.

Table 9: Endogenous Peer Effects Model

Variable	1		
	Bottom	Middle	Top
Predicted Academic Ability			
Peer GPA * Observational	0.474*** (0.051)	0.346*** (0.050)	0.342*** (0.048)
Peer GPA * Control	0.209** (0.096)	0.439*** (0.107)	0.377*** (0.129)
Peer GPA * Treatment	-0.015 (0.169)	0.146 (0.157)	0.219* (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 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 characteristics.

Table 10: Distribution of Peer Characteristics

Variable	1	2	3	4
Dependent Variable	Fraction Peers In 75th Percentile SAT Verbal	Fraction Peers In 25th Percentile SAT Verbal	Number Peers in Bottom Third of Predicted GPA	Number Peers in Top Third of Predicted GPA
Predicted GPA Top Third * Treatment Group	-0.011 (0.018)	-0.002 (0.011)	-0.203 (0.800)	-0.328 (0.391)
Predicted GPA Middle Third * Treatment Group	-0.079*** (0.021)	-0.010 (0.011)	-3.521*** (0.921)	0.716* (0.419)
Predicted GPA Bottom Third * Treatment Group	0.115*** (0.008)	0.021* (0.011)	4.519*** (0.446)	-0.826* (0.474)
Observations	5,412	5,412	5,412	5,412
R ²	0.293	0.083	0.265	0.056

We regress peer group characteristics on the treatment dummy interacted with three categories for own ability (predicted GPA). We include only the control and treatment groups which are from the classes of 2011-2012. 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 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 characteristics.

Table 11: 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 individual-level controls for students who are black, Hispanic, Asian, female, recruited athlete, and attended a preparatory school.