Econometrica, Vol. 81, No. 3 (May, 2013), 855–882

FROM NATURAL VARIATION TO OPTIMAL POLICY? THE IMPORTANCE OF ENDOGENOUS PEER GROUP FORMATION

BY SCOTT E. CARRELL, BRUCE I. SACERDOTE, AND JAMES E. WEST[1]

We take cohorts of entering freshmen at the United States Air Force Academy and assign half to peer groups designed to maximize the academic performance of the lowest ability students. Our assignment algorithm uses nonlinear peer effects estimates from the historical pre-treatment data, in which students were randomly assigned to peer groups. We find a negative and significant treatment effect for the students we intended to help. We provide evidence that within our "optimally" designed peer groups, students avoided the peers with whom we intended them to interact and instead formed more homogeneous subgroups. These results illustrate how policies that manipulate peer groups for a desired social outcome can be confounded by changes in the endogenous patterns of social interactions within the group.

KEYWORDS: Peer effects, social network formation, homophily.

0. INTRODUCTION

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.[2] 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.[3]

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) using pre-treatment data in which students were randomly assigned to peer groups (squadrons) of about 30 students. These estimates showed that low ability students benefited significantly from being with peers who have high SAT Verbal scores. We use these estimates to create optimally designed peer groups intended to improve academic achievement of the

[1]This article was completed under a Cooperative Research and Development Agreement with the U.S. Air Force Academy. This research was partially funded by the National Academy of Education, the National Science Foundation, and Spencer Foundation. Thanks to D. Staiger, R. Fullerton, R. Schreiner, B. Bremer, K. Silz-Carson, and K. Calahan.

[2]For 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).

[3]Unless the peer effects include a nonlinearity, there is no social gain to sorting individuals into peer groups. With a linear in means effect, a "good" peer taken from one group and placed into another group will have equal and offsetting effects on both groups. See Bénabou (1996) for a discussion of how moments other than the mean may be critical to determining outcomes.

 DOI: 10.3982/ECTA10168

bottom one-third of incoming students by academic ability while not harming achievement of students at other points in the distribution.[4] Using an experimental design, we sorted the incoming college freshman cohorts at USAFA into peer groups during the fall semesters of 2007 and 2008. 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 in the lowest third of the predicted grade point average (GPA) distribution. To do so, low ability students were placed into squadrons with a high fraction of peers with high SAT Verbal scores. We refer to these as bimodal squadrons. In the process, the sorting algorithm also created a set of treatment squadrons consisting largely of middle ability students. We call these homogeneous squadrons.

The reduced form coefficients (using the pre-treatment data) predicted a Pareto-improving allocation in which grades of students in the bottom third of the academic distribution would rise, on average, 0.053 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 -0.061 ($p = 0.055$). For the middle ability students, expected to be unaffected, we observe a positive and significant treatment effect of 0.082 ($p = 0.041$). High ability students are unaffected by the treatment.

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. Survey responses following the experiment show that, compared to the control group, low ability students in the treatment group were much more likely to sort into study (friendship) groups with other low ability students. For the middle ability students, evidence suggests that the positive treatment effect occurred because these students did not interact with low ability students after being placed into the homogeneous squadrons.

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, and Ridder (2009). The study is unusual in its use of historical pre-treatment data to infer optimal policy, implement, and then test the efficacy of the policy in a controlled experiment. Importantly, our results highlight both the significant role that peers play in the education production process and the theoretical difficulties in manipulating peers to achieve a desired policy outcome.

[4]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.

Well known difficulties exist in the application of policy to affect a desired outcome. General equilibrium responses as in Lucas (1976) or Acemoglu (2010) can undo effects predicted by more simple partial equilibrium models. Large policy interventions can also lead to political responses by actors and interest groups (Acemoglu (2010)). However, we see in our results a different mechanism at work; policy interventions can affect patterns of endogenous social interaction. As such, we believe that endogenous responses to large policy interventions are a major obstacle to foreseeing the effects of manipulating peer groups for a desired social outcome.

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 experimental design and expected treatment effects. Section 3 presents results from the experiment and explores potential simple reasons for the experiment's unexpected findings. Section 4 discusses the role of peer dynamics and endogenous peer group formations. Section 5 concludes.

1. DATA

1.1. *The Data Set*

Our pre-treatment data set 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 by *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.

We use grade point average (GPA) as our measure of freshman academic performance. 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 courses. 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 toward certain professors. Carrell, Fullerton, and West (2009) and Carrell and West (2010) provided 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 set of summary statistics is provided in Table I.

TABLE I

SUMMARY STATISTICS[a]

Variables	(1) Pre-Treatment Group Mean (sd)	(2) Control Group Mean (sd)	(3) Treatment Group Mean (sd)
Grade Point Average	2.785	2.789	2.781
	(0.661)	(0.642)	(0.659)
Fraction of High SAT-V Peers	0.276	0.263	0.272
	(0.0742)	(0.0603)	(0.161)
Fraction of Low SAT-V Peers	0.236	0.242	0.244
	(0.0717)	(0.0584)	(0.0774)
SAT Verbal Score	6.342	6.327	6.323
	(0.682)	(0.661)	(0.667)
SAT Math Score	6.643	6.568	6.580
	(0.654)	(0.646)	(0.653)
HS Academic Composite	12.96	12.82	12.81
	(2.103)	(2.152)	(2.162)
HS Fitness Score	4.451	3.809	3.799
	(0.994)	(0.725)	(0.728)
HS Leadership Score	17.28	17.29	17.32
	(1.844)	(1.668)	(1.713)
Recruited Athlete	0.247	0.229	0.226
	(0.431)	(0.420)	(0.419)
Attended Prep School	0.195	0.172	0.175
	(0.396)	(0.377)	(0.380)
Student is Black	0.0459	0.0529	0.0558
	(0.209)	(0.224)	(0.230)
Student is Hispanic	0.0661	0.0831	0.0771
	(0.248)	(0.276)	(0.267)
Student is Asian	0.0666	0.0847	0.0853
	(0.249)	(0.279)	(0.279)
Student is Female	0.180	0.208	0.216
	(0.384)	(0.406)	(0.412)
Observations	7160	1228	1219

[a]This shows summary statistics for the analysis sample. Data include observations for all students during the fall semester of their freshman year. The pre-treatment group consists of students in the classes of 2005–2010. The treatment and control groups consist of students in the classes of 2011–2012. Fraction of High SAT-V Peers denotes the fraction of peers with an SAT Verbal score in the upper quarter of their entire class. High School Fitness Scores and Leadership Scores are calculated by the USAFA admissions office using data provided by applicants.

We categorize students as low, middle, or high predicted GPA. To calculate predicted GPA, we take students in the classes of 2005–2010 and regress freshman GPA on SAT Verbal, SAT Math, academic composite, leadership composite, fitness score, and dummy variables for black, Hispanic, Asian, female, recruited athlete, and preparatory school attendance. We then perform an out-of-sample forecast (i.e., we calculate predicted GPA) for students in the

graduating classes of 2011 and 2012 who comprise our treatment and control groups.

1.2. *Methods*

As described in Carrell, Fullerton, and 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 into squadrons.[5]

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 one's peers, and their pre-treatment characteristics,

$$GPA_{st} = X\alpha_1 + \overline{GPA}_{s-i}\alpha_{2t} + \overline{X}_{s-i}\alpha_{3t} + \varepsilon_{st}, \tag{1}$$

where GPA_{st} is a vector of individual students' freshman fall semester grade point average who are members of squadron s and of academic ability $t \in \{low, middle, high\}$. X is a matrix of each individual's pre-treatment characteristics, including SAT Math, SAT Verbal, academic composite, fitness score, leadership composite, race/ethnicity, gender, recruited athlete, and whether he or she attended a military preparatory school. $\overline{GPA}_{s-i}$ is a vector of average freshman fall semester GPA in squadron s excluding individual i. $\overline{X}_{s-i}$ is likewise the average of pre-treatment characteristics in squadron s excluding individual i. ε_{st} is the error term. Following Manski (1993), α_3 represents the exogenous peer effect and α_{2t} is the endogenous peer effect, which varies by academic ability.

Solving for the reduced form specification and taking the limit as the number of peers approaches infinity,

$$\begin{aligned} GPA_{st} &= X\alpha_1 + \overline{X}_{s-i}\frac{\alpha_{3t} + \alpha_{2t}\alpha_1}{1-\alpha_{2t}} + \frac{\alpha_{2t}}{1-\alpha_{2t}}\overline{\varepsilon}_{st} + \varepsilon_{st} \\ &= X\beta_{1t} + \overline{X}_{s-i}\beta_{2t} + \tilde{\varepsilon}_{st}. \end{aligned} \tag{2}$$

In estimating equation (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.

[5]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.

Carrell, Fullerton, and 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, and 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.[6] Our definition of a "high" (low) score is any peer in the top (bottom) quartile of the year-cohort SAT Verbal distribution.[7]

We estimate equation (2) using ordinary least squares (OLS), and results are shown in Table II. Columns 1 and 2 estimate a single coefficient for each peer characteristic, while Columns 3 and 4 allow separate coefficients for each third of the predicted GPA distribution. The odd numbered columns include but do not report coefficients on peer SAT Math and peer academic composite. Overall, the nonlinear models in Columns 3 and 4 find larger and more precisely estimated peer effects than Columns 1 and 2 or a traditional linear in means model as in Carrell, Fullerton, and West (2009) and Lyle (2009).[8] The results suggest several nonlinearities in the data. The models fit in Column 3 and 4 reject the restrictions in Columns 1 and 2 at the 0.05 and 0.01-level, respectively, and the peer SAT Verbal variables are jointly significant at the 0.05-level in all Columns 3 and 4. For Column 4, the coefficient on the fraction of peers in the top quartile of the SAT Verbal distribution is positive and significant for low (0.474) ability students and insignificant for high (0.233) and middle ability students (−0.119). Across the three predicted GPA groups, the "High Verbal SAT" peer coefficients are significantly different between the middle ability group and the other two groups. Additionally, the coefficient on the fraction of peers in the bottom quartile of the SAT Verbal distribution is negative and insignificant (−0.230) for the middle ability students and small and statistically insignificant for low and high ability students.

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

[7]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 quintiles.

[8]In a very similar context, Lyle found positive peer effects from high math ability cadets at West Point (the U.S. Military Academy). For brevity, we do not show results for the linear in means model. For these results, see Carrell, Fullerton, and West (2009).

TABLE II

PEER EFFECTS IN THE PRE-TREATMENT GROUP[a]

Variables	(1) GPA	(2) GPA	(3) GPA	(4) GPA
Fraction of High SAT-V Peers	0.181[d]	0.190[c]		
	(0.094)	(0.096)		
Fraction of Low SAT-V Peers	−0.050	−0.061		
	(0.095)	(0.094)		
Fraction of High SAT-V Peers × High $\widehat{GPA}$			0.222	0.233
			(0.156)	(0.151)
Fraction of High SAT-V Peers × Middle $\widehat{GPA}$			−0.136	−0.119
			(0.136)	(0.137)
Fraction of High SAT-V Peers × Low $\widehat{GPA}$			0.464[b]	0.474[b]
			(0.150)	(0.152)
Fraction of Low SAT-V Peers × High $\widehat{GPA}$			0.026	0.009
			(0.144)	(0.147)
Fraction of Low SAT-V Peers × Middle $\widehat{GPA}$			−0.219	−0.230
			(0.145)	(0.142)
Fraction of Low SAT-V Peers × Low $\widehat{GPA}$			0.065	0.061
			(0.141)	(0.140)
Observations	14,024	14,024	14,024	14,024
R^2	0.345	0.345	0.346	0.345
F All Peer Variables	0.994		1.089	
p-value All	0.430		0.365	
F All Peer SAT Verbal Variables	2.304	2.627	2.412	2.464
p-value SAT Verbal	0.102	0.075	0.028	0.025
F High SAT Verbal Peers (High $\widehat{GPA}$ v Middle $\widehat{GPA}$)			3.068	3.135
p-value H v M			0.081	0.078
F High SAT Verbal Peers (High $\widehat{GPA}$ v Low $\widehat{GPA}$)			1.598	1.665
p-value H v L			0.208	0.198
F High SAT Verbal Peers (Low $\widehat{GPA}$ v Middle $\widehat{GPA}$)			9.850	9.441
p-value L v M			0.002	0.002

[a]We run our baseline peer effects specifications in the pre-treatment group. 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. Odd specifications additionally control for peer SAT Math and peer academic composite variables. Low, Middle, and High groups are based on the distribution of predicted GPA using own pre-treatment characteristics. Robust standard errors in parentheses are clustered by class by squadron.

[b] $p < 0.01$,

[c] $p < 0.05$,

[d] $p < 0.1$.

These results suggest that low predicted GPA students benefit most from having peers with high SAT Verbal scores, while middle ability students may benefit from being separated from peers with low SAT Verbal scores. Zimmerman (2003) also found that SAT Verbal scores matter more for peer effects than SAT Math scores. High math scores are even more prevalent among

USAFA students than high verbal scores. Thus, within the squadrons studied here, high verbal ability peers may be a more scarce resource and have a higher marginal productivity for own outcomes.

Under the direction of the Superintendent of the U.S. Air Force Academy, we used the nonlinear peer effects results in Column 3 of Table II to sort the freshman students entering USAFA in the fall of 2007 and fall of 2008 (the graduating classes of 2011 and 2012) into peer groups with the goal of improving the grades of the lowest one-third of incoming ability students using a controlled experimental design.

2. EXPERIMENTAL DESIGN AND SORTING METHODOLOGY

The graduating classes of 2011 and 2012 entered USAFA with 1314 and 1391 students, respectively. Half of the incoming classes were randomly assigned to the control group and half to the treatment group.[9]

Table III shows a regression of membership in the treatment group on the pre-treatment variables. Column 1 shows results for the class of 2011, Column 2 shows results for the class of 2012, and Column 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 0.277 with a p-value of 0.99. Figure 1, Column 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 that has been used by USAFA since the summer of 2000. The algorithm provides an even distribution of students by demographic characteristics. 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.

[9] 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 first year of the experiment. Additionally, the random division was subject to the constraint that siblings were split between the treatment and control groups.

TABLE III

RANDOMIZATION CHECK[a]

Variables	(1) Class 2011	(2) Class 2012	(3) All Classes
SAT Verbal Score	0.018	0.005	0.009
	(0.026)	(0.024)	(0.018)
SAT Math Score	−0.001	0.030	0.014
	(0.034)	(0.031)	(0.023)
HS Academic Composite	−0.004	0.012	0.004
	(0.011)	(0.011)	(0.008)
HS Fitness Score	−0.017	0.014	−0.002
	(0.020)	(0.020)	(0.014)
HS Leadership Score	0.013	0.002	0.008
	(0.008)	(0.008)	(0.006)
Recruited Athlete	−0.001	0.023	0.011
	(0.037)	(0.036)	(0.025)
Attended Prep School	0.063	−0.019	0.018
	(0.054)	(0.044)	(0.033)
Student is Black	0.008	0.020	0.015
	(0.067)	(0.064)	(0.046)
Student is Hispanic	−0.009	0.003	−0.004
	(0.056)	(0.051)	(0.037)
Student is Asian	0.001	0.038	0.017
	(0.049)	(0.054)	(0.036)
Student is Female	−0.005	0.010	0.006
	(0.036)	(0.035)	(0.025)
$\widehat{GPA}$ in Lowest Third of Class	0.013	0.051	0.028
	(0.051)	(0.051)	(0.036)
$\widehat{GPA}$ in Highest Third of Class	0.039	−0.069	−0.014
	(0.048)	(0.048)	(0.034)
Class of 2011			−0.005
			(0.020)
Observations	1287	1366	2653
R^2	0.004	0.003	0.001
F All Variables	0.380	0.280	0.277
p-value	0.964	0.990	0.990

[a]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. SAT, Academic Composite, Fitness, and Leadership scores have been divided by 100.

Students in the treatment group were sorted into one of 20 treatment squadrons with the objective of raising the grades of students in the bottom one-third of predicted GPA. Drawing on recent work on assortative matching by Bhattacharya (2009) and Graham, Imbens, and Ridder (2009), we set our objective function to maximize the minimum peer effect (i.e., fraction of peers with high SAT Verbal scores) for each low ability student subject to the constraint that squadrons are of equal size and contain an even split of females,

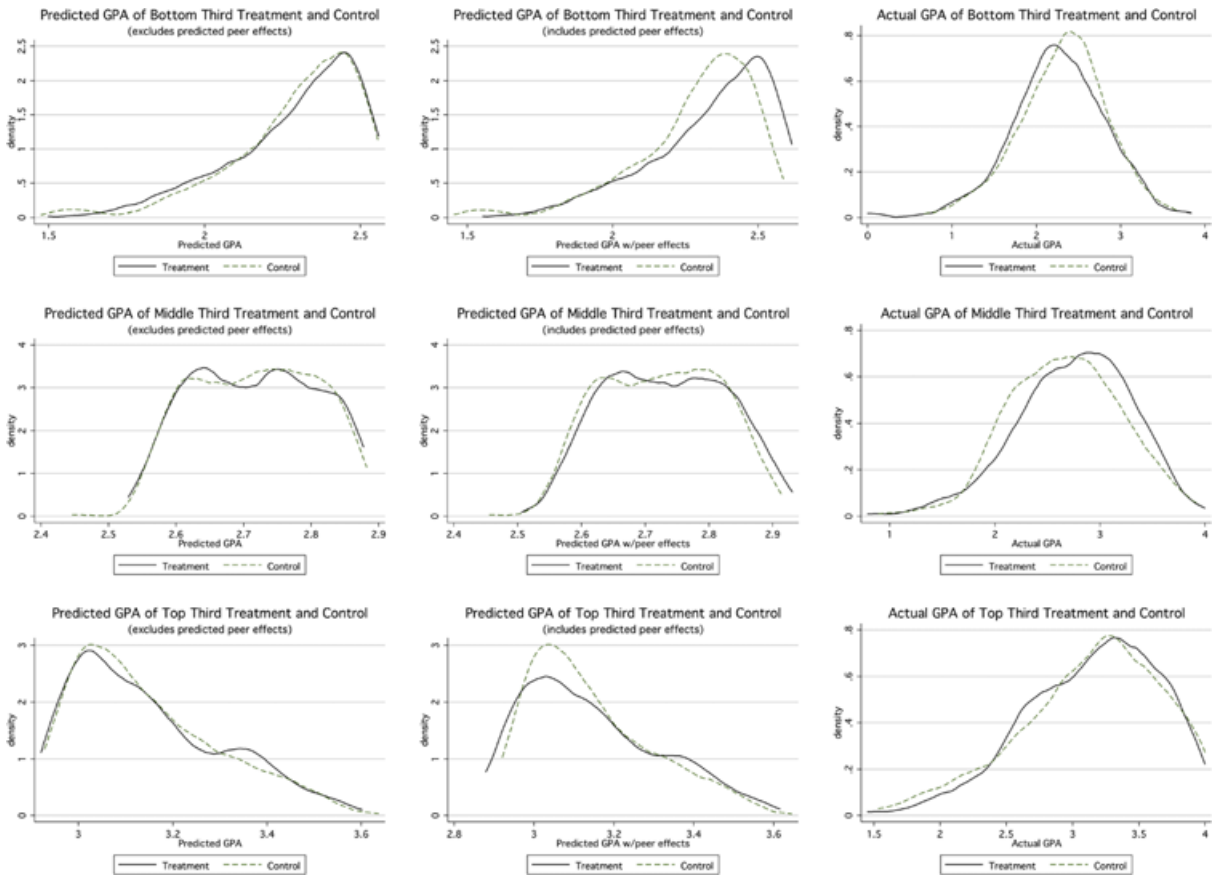

FIGURE 1.—Distribution of predicted and actual GPA for treatment and control by student ability.

athletes, racial and ethnic minorities, and students who attended a preparatory school. Thus, we solve the following objective function using integer programming[10]:

$$(3)\qquad \max_{m_{s,i}} \min_{i\in\mathbb{S}_{\text{low}}} \begin{bmatrix} m_{1,i} \\ m_{2,i} \\ \vdots \\ m_{20,i} \end{bmatrix}' \begin{bmatrix} g_1 & 0 & \cdots & 0 \\ 0 & g_2 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & g_{20} \end{bmatrix}^{-1} \times \begin{bmatrix} m_{1,1} & \cdots & m_{1,i-1} & 0 & m_{1,i+1} & \cdots & m_{1,n} \\ m_{2,1} & \cdots & m_{2,i-1} & 0 & m_{2,i+1} & \cdots & m_{2,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ m_{20,1} & \cdots & m_{20,i-1} & 0 & m_{20,i+1} & \cdots & m_{20,n} \end{bmatrix} X\widehat{\beta}_{2\text{low}},$$

where $\mathbb{S}_{\text{low}}$ is the set of low ability students. $m_{s,i} \in \{0, 1\}$ represents the membership of student i in one of the 20 treatment squadrons, s, while g_s is the number of students in each squadron s other than student i, and n is the overall number of students in the treatment group. X is a $(1 \times n)$ vector of indicator variables representing whether each student has a high SAT Verbal score, and $\widehat{\beta}_{2\text{low}}$ is the estimated peer SAT Verbal coefficient of 0.464 from equation (2) for low ability students, as shown in Table II, Column 3. Since it is unlikely that a low predicted GPA student also has a high SAT Verbal score, the choice of which squadron to assign student i largely does not affect the proportion of students in each squadron with high SAT Verbal scores, as determined by $g_s^{-1}\sum_{j\neq i} m_{sj}X_j$.

Prior to sorting the actual treatment group, we repeatedly sorted simulated treatment groups drawn from the pre-treatment data by maximizing equation (3) subject to constraints, and found that our algorithm consistently created two types of squadrons. The first type of squadron groups low ability students with a large number of peers with high SAT Verbal scores. We refer to these as bimodal squadrons. The second type of treatment squadron consists largely of middle ability students without high SAT Verbal scores, which we call homogeneous squadrons. We intentionally allowed the algorithm to engage in extreme sorting to maximize the potential peer effects and the statistical power of the experiment. Assuming that students in the treatment group would choose among available study partners in the same way that students in the pre-treatment groups did, changing the composition of a low ability student's squadron to include a larger proportion of students with high SAT Verbal scores should be beneficial. Ex ante, this did not seem unreasonable. For

[10]The software used to perform the integer program maximization, XPressMP, was provided to us by FICO under their Academic Partners Program.

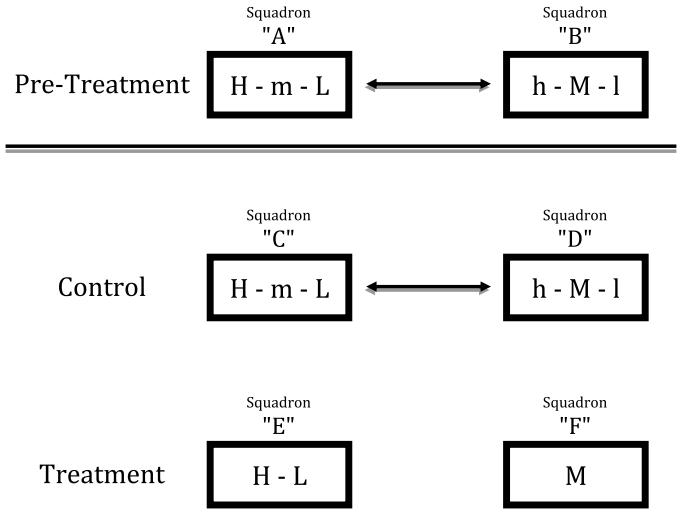

FIGURE 2.—Experimental design. This figure depicts the essence of the experimental design. For instance, H-m-L squadrons are those with a large fraction of High ability students, a small fraction of Medium ability students, and a large fraction of Low ability students. Likewise, h-M-l squadrons are those with a small fraction of High ability students, a large fraction of Medium ability students, and a small fraction of Low ability students.

low ability students in the pre-treatment data, the mean and maximum fraction of peers with high SAT Verbal scores is 0.28 and 0.50. For the treatment squadrons, the mean and maximum fraction is 0.38 and 0.41.[11]

Figure 2 depicts the essence of the experimental design. Pre-treatment, students were randomly assigned to squadrons with respect to academic ability. As a consequence of random sampling, some squadrons were composed of relatively more high and low ability students than medium ability. Let these be called "A." Other squadrons were composed of relatively more medium ability students than high or low, which we refer to as "B." Estimates from equation (2) indicate that low ability students in "A" squadrons benefited from greater exposure to high ability students than the low ability students in "B" squadrons. The composition of control group squadrons is similar to that of pre-treatment squadrons. Squadrons "C" and "D" are analogous to squadrons "A" and "B." In the treatment squadrons, middle ability students are removed from "E" squadrons and placed by themselves in "F" squadrons.

Results from the actual sorting of students in the treatment and control groups are presented graphically in Figure 3, which shows, by treatment and

[11]The main assumption is that the peer effects estimates are linear in the fraction of peers with high SAT Verbal scores for low ability students. For example, the algorithm assumes that going from 30 to 35 percent of your peers with high SAT scores has the same effect as going from 25 to 30 percent of peers with high scores.

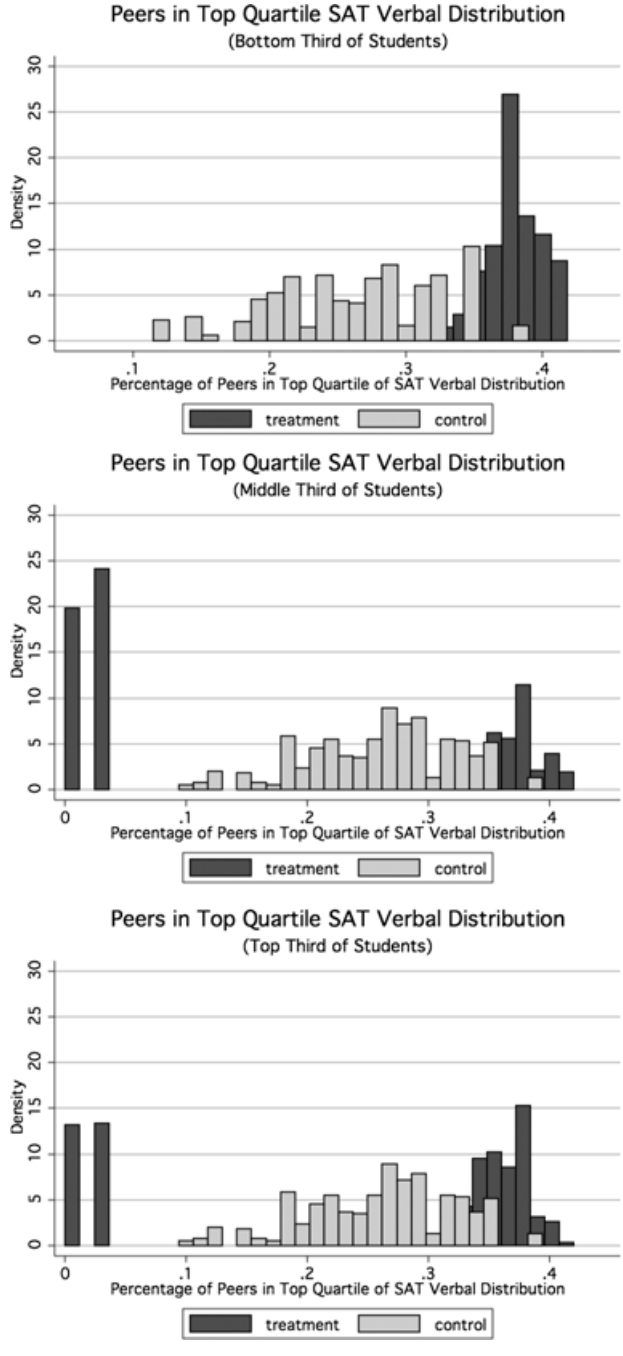

FIGURE 3.—Squadron peer characteristics by student ability.

TABLE IV

PREDICTED TREATMENT EFFECT[a]

Variables	(1) All Students	(2) Bottom $\widehat{GPA}$	(3) Middle $\widehat{GPA}$	(4) Top $\widehat{GPA}$
Student in Treatment Group	2.787	2.390	2.783	3.198
	(0.026)	(0.027)	(0.027)	(0.027)
Student in Control Group	2.772	2.336	2.767	3.195
	(0.026)	(0.027)	(0.027)	(0.026)
Predicted Treatment Effect	0.015	0.053[b]	0.016	0.003
	(0.037)	(0.037)	(0.037)	(0.037)
Observations	2653	881	884	888

[a]We use the regression coefficients in Table II, Column 2 to compute out-of-sample predicted GPAs and forecast standard errors for the students in the treatment and control groups. We test a null hypothesis that predicted grades in the treatment group are not greater than grades in the control group, assuming independence of observations across squadrons.

[b] $p < 0.1$.

control status and by each third of own ability, histograms of the fraction of peers who were in the top quartile of the SAT Verbal distribution. Relative to randomly assigned squadrons, the optimal sorting algorithm created a number of the bimodal squadrons assigning low predicted GPA students in the treatment group to squadrons with a much higher proportion of high SAT Verbal peers. Likewise, the algorithm also created a number of the homogeneous treatment squadrons consisting largely of middle ability students.

Table IV shows predicted GPA and the predicted treatment effect by student ability from the sorting algorithm. We predict GPA for the treatment and control groups using the estimates from Table II, Column 3. Reported standard errors for each group $((low, middle, high) \times (treatment, control))$ are calculated from individual-level forecast standard errors assuming the independence of individual observations across squadrons.[12] For students in the bottom third of incoming academic ability, the estimated gain of the treatment group over the control group is a statistically significant 0.053 grade points. For students in the middle and top third of the academic distribution, the estimated treatment effects are positive, but statistically insignificant. Figure 1, Column 2 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.

[12]We estimate the covariance of grades across students within the same squadron using a random effects model on pre-treatment data.

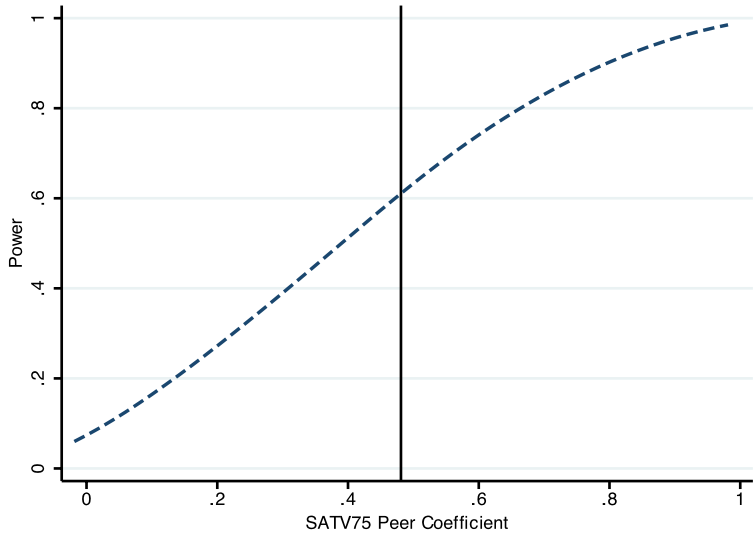

FIGURE 4.—Power of the experiment.

Specifically, we simulated the treatment effect for the bottom one-third of students as being equal to the fitted values from Column 3 in Table II 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 0.464, 609 of 1000 draws were positive and statistically significant at the 0.05 level and 961 were positive.[14]

A sensible specification check is to ask whether the peer effects regressions that we ran in the pre-treatment data (for the classes of 2005–2010) yield similar results when we run those same specifications in the control group (for the classes of 2011–2012). These specification checks are shown in Table V. Results indicate similar peer effects estimates in our randomly assigned control group compared to the pre-treatment data. For Column 4, the coefficient on the fraction of peers in the top quartile of the SAT Verbal distribution is

[13]The estimated variance of the error terms were obtained from the pre-treatment data in predicting student grades using a random effects model.

[14]We also perform power calculations in the more traditional way; we allow the size of the treatment effect (as opposed to the size of the underlying peer coefficient) to vary and we use simulation to ask how frequently we observe a statistically significant treatment effect. Results show that at the expected treatment effect of 0.053, we would observe a statistically significant coefficient on the treatment regression roughly 58 percent of the time. Results are available upon request.

TABLE V

PEER EFFECTS IN THE CONTROL GROUP[a]

Variables	(1) GPA	(2) GPA	(3) GPA	(4) GPA
Fraction of High SAT-V Peers	0.279	0.326		
	(0.223)	(0.235)		
Fraction of Low SAT-V Peers	−0.101	−0.011		
	(0.280)	(0.280)		
Fraction of High SAT-V Peers × High $\widehat{GPA}$			0.256	0.480
			(0.328)	(0.347)
Fraction of High SAT-V Peers × Middle $\widehat{GPA}$			−0.125	−0.102
			(0.446)	(0.442)
Fraction of High SAT-V Peers × Low $\widehat{GPA}$			0.769[b]	0.594
			(0.426)	(0.440)
Fraction of Low SAT-V Peers × High $\widehat{GPA}$			0.202	0.322
			(0.462)	(0.422)
Fraction of Low SAT-V Peers × Middle $\widehat{GPA}$			−0.404	−0.347
			(0.419)	(0.476)
Fraction of Low SAT-V Peers × Low $\widehat{GPA}$			−0.131	−0.069
			(0.359)	(0.381)
Observations	2423	2423	2423	2423
R^2	0.345	0.339	0.348	0.340
F All Peer Variables	2.702		3.134	
p-value All	0.027		0.001	
F Peer SAT Verbal	1.060	1.279	1.012	0.934
p-value SAT Verbal	0.356	0.290	0.432	0.482
F Peer Effect High v Middle			0.586	1.215
p-value T v M			0.448	0.277
F Peer Effect High v Low			0.927	0.046
p-value T v B			0.342	0.832
F Peer Effect Low v Middle			1.671	1.029
p-value B v M			0.204	0.317

[a]We run our baseline peer effects specifications in the pre-treatment and control groups. 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. Odd specifications additionally control for peer SAT Math and peer academic composite variables. Low, Middle, and Top groups are based on the distribution of predicted GPA using own pre-treatment characteristics. Robust standard errors in parentheses are clustered by class by squadron.

[b]$p < 0.1$.

positive for both low (0.594) and high (0.480) ability students and negative for middle ability students (−0.102).

In Table A.I of the Supplemental Material (Carrell, Sacerdote, and West (2013)), we combine the pre-treatment and control data. We run our baseline peer effects specification allowing for the peer effects to vary by pre-treatment versus control. The coefficient on fraction of High SAT Verbal peers is 0.456 for low ability students in the pre-treatment group and 0.855 for low ability stu-

TABLE VI

OBSERVED TREATMENT EFFECTS[a]

Variables	(1) All Students	(2) Low $\widehat{GPA}$	(3) Middle $\widehat{GPA}$	(4) High $\widehat{GPA}$
Student in Treatment Group	0.001	−0.061[c]	0.082[b]	−0.012
	(0.022)	(0.031)	(0.039)	(0.036)
Observations	4834	1571	1626	1637
R^2	0.357	0.136	0.067	0.151

[a]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 and control for own SAT Verbal Score, SAT Math Score, HS Academic Composite, HS Fitness Score, and HS Leadership Score. Standard errors are clustered at the Class by Squadron level. Robust standard errors are in parentheses.

[b] $p < 0.05$,

[c] $p < 0.1$.

dents in the control group. The difference is not statistically significant. Overall, we do not find any evidence of a change in the nature of peer effects between the pre-treatment and control groups.

3. EXPERIMENTAL RESULTS

Actual results of the experiment are shown in Table VI and in Figure 1, Column 3. 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 (−0.061) indicates that the treatment was of the approximate magnitude predicted but the opposite sign, meaning that low ability students in the treatment group performed significantly worse than those in the control group.[15]

The second striking finding is the positive and statistically significant (0.082) treatment effect for students in the middle third of the predicted GPA distribution.[16]

3.1. *Is There a Simple Explanation for the Unexpected Results?*

One possible simple explanation is that the negative treatment effect is simply due to sampling variation; meaning that a positive treatment effect exists,

[15]In Table A.X of the Supplemental Material, we estimate the treatment effects by gender and race. Results show that the negative treatment effect is primarily driven by male students who have a significant negative treatment effect of −0.094, while female students have an insignificant positive treatment effect.

[16]Further results in Table A.X show positive effects across most subgroups.

but that it was not observed 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 1000 was the treatment effect negative and significant at the 0.10-level and in only 39 of 1000 draws did the treatment effect have a negative sign. In addition to being significantly different than zero, the negative treatment effect is significantly different from the predicted treatment effect.[17] The likelihood of the distribution of grades being as we hypothesized at the outset of our experiment is only 1 percent. Hence, we reject sampling variation as a simple explanation for the results we observed.

A second possible simple explanation is that the peer effects estimates used to motivate the experiment are not robust. To test their robustness, Table A.II of the Supplemental Material shows results in the pre-treatment 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) as well as our measure for leadership ability (*HS Leadership Composite*) and physical fitness (*HS Fitness Score*), 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 10 possible effects to vary by own predicted GPA (three groups), yielding a total of 30 peer coefficients.[18] 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 from the restricted model of equation (2) reported in Table II. Additionally, none of the other peer effects coefficients (SAT Math, Academic Composite, Leadership Composite, or HS Fitness Score) are individually significant, while the coefficients for the peer SAT Verbal variables are jointly significant at the 0.05-level ($p = 0.034$). These results provide evidence that SAT Verbal is the key peer attribute, as it is quite robust to the inclusion of multicollinear peer variables.

As a second robustness test, Table A.III of the Supplemental Material 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. Additionally, the magnitude of the effects is statistically indistinguishable across the two sets of years.[19] Hence, we conclude that the

[17] $t = \frac{-0.061-0.053}{\sqrt{0.031^2+0.037^2}} = -2.36$, assuming the predicted and observed treatment effects are independent.

[18] For brevity, we only show results for the 18 academic peer effects variables. All coefficients are individually and jointly insignificant for the leadership and fitness peer variables.

[19] We also note that Lyle (2009) found that cadets at the U.S. Military Academy benefit from having peers with high SAT Math scores. Specifically, he regressed outcomes on the 75th percentile of math scores within one's company. The fact that Lyle found similar results in a very similar context increases our confidence that the original results from the pre-treatment data were not spurious.

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.

Although the peer effects in the pre-treatment data appear to be robust, a final possible simple explanation 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 (e.g., an increased use of texting or Facebook to sort into friendship groups or interact with peers). To consider this hypothesis, we refer back to the earlier results from Tables III and A.I in which 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 pre-treatment data. Recall that the peer effects specifications yielded similar results in both the pre-treatment and control groups. Furthermore, in Table A.XII, we show that our predicted treatment effect for low ability students is a significant 0.067 when using the estimated peer effects from the control group sample. This estimate is larger and more significant than the estimated treatment effect of 0.053 from Table IV, and is not consistent with a hypothesis of class-wide changes in peer interactions.

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 the difficulty of interpreting these estimates as true peer effects as opposed to common shocks (Manski (1993)), 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 A.IV of the Supplemental Material show large positive and statistically significant endogenous effects for both the pre-treatment and control groups. However, the effects are smaller and less statistically significant in the treatment group, particularly for the low ability students.

These results do not provide any evidence of changing peer interactions between the randomly assigned pre-treatment and control squadrons. However, the results suggest that something very different may have occurred in the treatment squadrons. We explore the role of endogenous peer group formation in the next section.

4. PEER DYNAMICS AND ENDOGENOUS PEER GROUP FORMATION

When designing the experimental squadrons, the model implicitly assumed that the peer dynamics within the treatment squadrons would remain similar to those observed in the randomly assigned pre-treatment squadrons. However,

we present evidence that students in the treatment squadrons endogenously sorted into subgroups in ways that were not observed in the pre-treatment data.

As shown in Figure 3, the sorting algorithm created rather different squadrons than those previously observed under random assignment. Figure 5 provides more detail by showing the distribution of low predicted GPA peers in the pre-treatment, treatment, and control groups. While low ability students in the treatment group were assigned an unusually large number of peers with high SAT Verbal scores (Figure 3), they were also assigned an unusually large number of low ability peers (Figure 5). This was achieved by removing many of the middle ability peers and placing them in squadrons of primarily middle ability peers. In other words, the sorting procedure led to a combination of (1) heterogeneous squadrons with many low ability students grouped together with students with high SAT Verbal scores and (2) homogeneous squadrons consisting of middle and high ability students that earned lower SAT Verbal scores.

Although the bimodality of ability in the treatment squadrons was not commonly present in the pre-treatment data, some bimodality did occasionally occur as a result of random sampling variation. In Table VII, we test whether various indicators of heterogeneity had any effect on the academic achievement of low predicted GPA students in the pre-treatment data. We report results for five different ways of defining a squadron as being bimodal. These indicators include (1) having fewer than 6 middle predicted GPA students in the squadron, (2) having the proportions of bottom and top peers (as defined by predicted GPA) both exceed 40 percent, (3) having the proportions of peers in the top quartile of SAT Verbal scores and the bottom quartile of SAT Verbal scores both exceed 35 percent, (4) having 15 or more low predicted GPA students in the squadron, and (5) having the fraction of peers in the first quartile of predicted GPA and fourth quartile of SAT Verbal scores each exceed 40 percent.

Across four of five indicators of heterogeneity, low predicted GPA students in more heterogeneous squadrons performed *better* than average. Hence, we find no evidence that the unexpected negative treatment effect could have been predicted, ex ante, in our pre-treatment data, and the dynamic that occurred within the treatment squadrons does not appear to be detectable in the pre-treatment data.

However, the true peer group for a student may not be his entire squadron, but rather a smaller and endogenously chosen group of students within the squadron. To illustrate, suppose that when student i, who is of relative low academic ability, arrives at USAFA, he is randomly assigned to a squadron with 31 other first year students. His optimal size study group or close friendship group is likely less than the full 31 other freshmen members of his squadron. Hence he uses a combination of closeness of physical distance, closeness of ability and background, and common interests to endogenously sort into a smaller peer

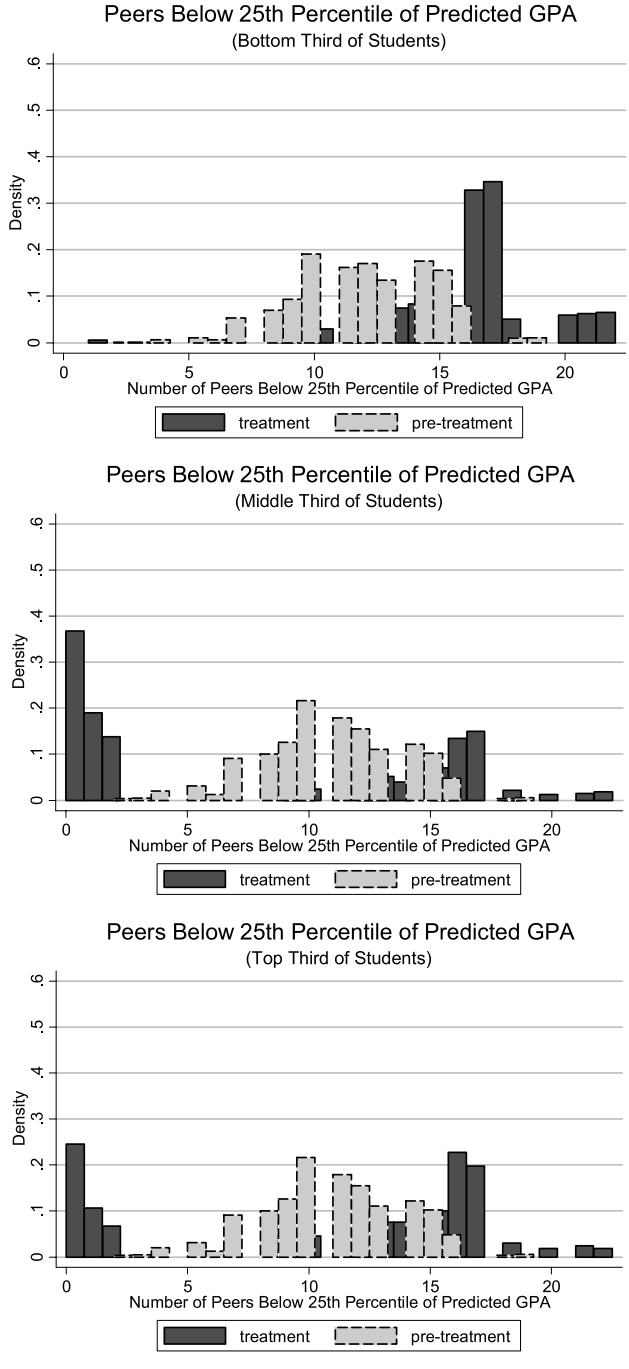

FIGURE 5.—Distribution of low ability peers.

TABLE VII

EFFECTS FROM BEING IN A BIMODAL SQUADRON IN THE PRE-TREATMENT GROUP[a]

Variables	(1) GPA	(2) GPA	(3) GPA	(4) GPA	(5) GPA
Fewer than 6 Middle $\widehat{GPA}$ Students in Squadron	0.074				
	(0.050)				
Fraction of Low $\widehat{GPA}$ Peers and High $\widehat{GPA}$ Peers > 0.40		0.046			
		(0.043)			
Fraction of High SAT-V Peers and Low SAT-V Peers > 0.35			−0.002		
			(0.033)		
Greater than 15 Low $\widehat{GPA}$ Students in Squadron				0.036	
				(0.030)	
Fraction of Low $\widehat{GPA}$ Peers and High SAT-V Peers in 4th Quartile					0.054
					(0.051)
Observations	4638	4638	4638	4638	4638
R^2	0.096	0.096	0.095	0.096	0.096
Number Observations Who Meet This Definition of Bimodality	213	95	219	662	242

[a]Sample includes only students in lowest third of predicted GPA. We regress own GPA on indicators for various measures of squadron heterogeneity for students with low predicted GPA in the pre-treatment group. 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. Robust standard errors in parentheses are clustered by class by squadron.

group within the squadron.[20] In the pre-treatment data (and the control group data), randomization creates a good mixing of all student types and abilities into a squadron, and this mixing can limit the degree to which student *i* will form a study (friendship) group that is homogeneous in terms of race, gender, or academic ability. However, the optimally designed peer groups in the treatment may have unintentionally provided student *i* with more opportunities to form a homogeneous study (friendship) group within the squadron since the treatment combined groups of 15 or so of the lowest ability students with 15 or so of the highest ability students.

Hence, the unexpected results from the experiment could stem from endogenous sorting of students into peer groups in a way that is not fully captured in our model estimated with pre-treatment data. We note that the treatment could affect peer choice in at least two ways. First, the treatment changed the availability of certain types of peers. For example, low ability treatment students saw marked increases in the availability of high and low ability peers, to the exclusion of middle ability peers. Additionally, the binding constraints (e.g., minimum number of athletes, minorities, etc.) of our sorting algorithm also al-

[20]See Marmaros and Sacerdote (2006) and Mayer and Puller (2008) for evidence on both the size and determinants of friendship groups.

tered the peer ability/demographic mix within the squadron. Low (high) ability athletes are more likely to be with other low (high) ability athletes in treatment squadrons.[21] Second, *conditional* on the availability of peers, the treatment may have also altered the degree to which low ability students are attracted to one another (homophily).

To measure the degree to which the treatment altered endogenously chosen peers, we surveyed students about their study partners and friendships within the squadron, and also collected administrative data on roommate choices. Using these data, we first calculate the overall effect on peer choice by regressing peer characteristics (e.g., fraction of friends who are low ability) on treatment status. We then calculate the pure availability (compositional) effect of the treatment by running a simulation in which we impose the null hypothesis that treatment and control students choose peers at random from the available set of students within the squadron. Thus, we ask how much treatment students' choices differ from control students' choices purely because the composition of potential peers is different. We label as homophily the difference between the overall treatment effect on peer choice and the availability effect.[22]

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 were friends (i.e., spent free time with) as a freshman. We received usable responses from approximately 25 percent of the experimental subjects.[23] Table VIII examines how the treatment affected study partner and friend choices of low predicted GPA students. For each respondent, we calculate the fraction of study partners that were low, middle, high predicted GPA or high SAT Verbal. The first column within the first row indicates that low ability students in the treatment squadrons were 17.1 percentage points more likely to have low predicted GPA study partners than low ability students in the control squadrons. The second column of the first row indicates that 12.5 percentage points can be attributed to the different composition of low, medium, and high ability students in treatment versus control squadrons. The difference of 4.6 percentage points reported in Column 3 is attributed to additional homophily in the treatment squadrons. The empirical p-value reported below indicates that in only 1.1 percent of resampled draws was the treatment minus control

[21] In the treatment squadrons, among athletes and minorities, own SAT Math scores are positively correlated in the treatment squadrons, while they are uncorrelated in control squadrons.

[22] An alternative approach which yields similar results is to calculate homophily directly by running regressions of peer choice on treatment dummies while controlling for various measures of squadron composition.

[23] We ran our treatment effect regressions on the subset of students who responded to the survey and found very similar results. Among respondents, low ability students had a treatment effect of -0.055 and middle ability students had a treatment effect of 0.123. Additionally, we found no differences in response rates between the treatment and control; however, response rate was positively correlated with academic ability.

TABLE VIII

LOW PREDICTED GPA STUDENTS: TREATMENT EFFECTS ON STUDY PARTNER AND FRIEND CHOICES[a]

	(1)	(2)	(3)	(4)	(5)	(6)
	Study Partners			Friends		
Treatment Effect On ...	Actual Peer Choices (sd)	If Peers Chosen Randomly (sd)	Actual Minus Random $P(A < R)$	Actual Peer Choices (sd)	If Peers Chosen Randomly (sd)	Actual Minus Random $P(A < R)$
Fraction Low $\widehat{GPA}$	0.171[b]	0.125[b]	0.046[c]	0.201[b]	0.135[b]	0.071[b]
	(0.061)	(0.020)	0.011	(0.050)	(0.019)	0.000
Fraction Middle $\widehat{GPA}$	−0.214[b]	−0.105[b]	−0.109[b]	−0.105[b]	−0.114[b]	0.008
	(0.046)	(0.015)	1.000	(0.038)	(0.014)	0.272
Fraction High $\widehat{GPA}$	0.042	−0.020	0.062[b]	−0.095[c]	−0.022	−0.074[b]
	(0.060)	(0.016)	0.000	(0.048)	(0.015)	1.000
Fraction High SAT-V	0.064	0.103[b]	−0.039[c]	0.004	0.112[b]	−0.107[b]
	(0.052)	(0.015)	0.994	(0.046)	(0.014)	1.000
Fraction Low $\widehat{GPA} > 0.50$	0.269[b]	0.170[b]	0.100[c]	0.236[c]	0.170[b]	0.066
	(0.097)	(0.042)	0.008	(0.092)	(0.041)	0.057
Observations	494	10,000	10,000	543	10,000	10,000

[a]Data on study partners and friends 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 friends during their freshman year. Response rate was approximately 25 percent. For each study partner and friend dependent variable, estimated coefficients represent the difference between the treatment and control groups. Columns 1 and 4 report estimated coefficients using actual (endogenous) study partner choices. Columns 2 and 5 report estimated coefficients using 10,000 iterations of resampled study partner or friend assignments within each treatment and control squadron. These coefficients represent the purely compositional treatment effect on study partner availability. Standard errors are in parentheses under each estimated coefficient. Columns 3 and 6 report the difference between actual choices and random choices. Below these differences, we report empirical p-values, which are the proportion of random draws less than the actual choices observed. All specifications include year fixed effects and individual-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.

[b] $p < 0.01$,

[c] $p < 0.05$.

difference in resampled choices observed to be greater than actual choices. This is evidence of homophily; low predicted GPA students in the treatment squadrons actively sought out other low predicted GPA students as their study partners (relative to control squadrons) beyond what would have occurred as a result of random selection.

We note, in particular, changes in the fraction of high SAT Verbal study partners. Although low ability students in treatment were placed in squadrons composed of more students with high SAT Verbal scores, results in Row 4, Column 3 indicate that low predicted GPA students in the treatment squadrons purposefully selected high SAT Verbal study partners with less frequency (−3.9 percentage points) than in the control squadrons over and above composi-

tional effects. The empirical p-value indicates that 99.4 percent of resampled (random) choices exceed actual choices, showing active selection away from high SAT Verbal study partners. We believe this is a critical piece of evidence in understanding the outcome of our experiment. The final row of Table VIII presents differences in treatment versus control squadrons in whether the fraction of low predicted GPA study partners exceeds 50 percent. Column 3 indicates that 10 percentage points of this difference is over and above compositional effects. We note that there is no reason to think that GPA is linear in the proportion of one's study partners who are low ability. It is entirely possible that once the fraction of study partners who are low ability reaches a certain threshold, a student's outcomes suffer a great deal. Thus, the treatment could have pushed certain treatment students past this threshold and harmed rather than helped their GPA.

Friendship choices, as represented by Columns 4 through 6, also show evidence of endogenous selection and homophily. Low predicted GPA students in the treatment squadrons are significantly more likely to choose fellow low predicted GPA students as friends relative to control squadrons, and less likely to choose high predicted GPA and high SAT Verbal friends relative to control squadrons.

The treatment effects on study groups for middle and high predicted GPA students are also of interest. These results appear in Tables A.XVII and A.XVIII of the Supplemental Material, respectively. Middle ability students in treatment squadrons select other middle ability students as both friends and study partners, while avoiding both high predicted GPA and high SAT Verbal study partners relative to control squadrons. We consider this to be further evidence of homophily and endogenous selection within treatment squadrons relative to control squadrons. High predicted GPA students in treatment squadrons avoid low predicted GPA students both as study partners and as friends relative to control squadrons.

As an alternative to relying on survey evidence, for the treatment and control groups we examine roommate choices within the squadron. To do so, we obtained a partial list of roommate pairs for the class of 2012.[24] To test for endogenous sorting into roommate pairs, we regress second semester roommate characteristics on treatment status and own characteristics.[25] These results are shown in Table A.XIX of the Supplemental Material, which provides some evidence of homophily. For instance, consider the degree to which low ability students select high predicted GPA roommates (Row 3 and Columns 1, 2, and 3). Due to compositional or availability effects, members of the treatment

[24] In their first semester, students at USAFA are not permitted to choose their own roommates. However, in the second semester, this prohibition is relaxed.

[25] In results not reported here, we find that, due to randomization, first semester roommate backgrounds are uncorrelated in both the treatment and control groups.

group would be 4.7 percentage points less likely to select a high ability roommate. However, low ability members of the treatment group chose high ability roommates 9.5 percentage points less often.

Overall, the results on study partner choices, friendships, and roommate pairs show different patterns of endogenous peer group formation between the treatment and control groups. Although our experiment provided low ability members of the treatment group with larger proportions of more beneficial potential study partners from which to choose, they were not chosen in favor of other low ability study partners. We cannot know with certainty the exact cause of our unexpected treatment effects. But evidence we have presented is consistent with endogenous responses in how peer groups formed sufficiently large to reverse the predicted treatment effect.

On the basis of our experience, we would like to be able to suggest a methodology in which a structural model of peer effects, such as our equation (1), could lead to a reduced form peer effects model, such as our equation (2), which could be used for successful policy predictions. Regrettably, we do not now see such a methodology. Acemoglu (2010), in discussing the importance of external validity in policy analysis, noted that "when political economy factors are important, evidence on the economic effects of large-scale policy changes under a given set of political conditions is not sufficient to forecast their effect on the economy and society." Likewise, endogenous policy responses in peer group formation appear to make the effect of large policy changes very difficult to forecast.

5. 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 pre-treatment data. We sorted the entire freshman cohorts in the graduating classes of 2011 and 2012. We randomly split the incoming freshman class into treatment and control groups. Members of the control group were randomly placed into control squadrons, while members of the treatment group were sorted into 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, 0.053 grade points, while higher ability students' 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 −0.061. For the middle ability students, predicted to be unaffected, we observed a positive and statistically significant treatment effect of 0.082.

We find evidence that the endogenous sorting of roommates, study partners, and friends evolved in a different way in the treatment group than in the control group. Low ability students in the treatment group saw significant increases in their number of low ability study partners and low ability friends. This increase appears to be not merely the result of larger numbers of low ability peers from which to choose (a compositional effect), but also the result of choices which reveal a same type attraction (homophily). We believe that endogenous responses to large policy interventions such as the ones we observe are a major obstacle to foreseeing the effects of manipulating peer groups. Despite our emphasis on endogenous sorting, we are unable to reject a related story in which the presence of middle ability students is a crucial part of generating positive peer effects for the lower ability students. We conclude that social processes are so rich and complex that one needs a deep understanding of their formation before one can formulate "optimal policy."

REFERENCES

ACEMOGLU, D. (2010): "Theory, General Equilibrium, and Political Economy in Development Economics," *Journal of Economic Perspectives*, 24 (3), 17–32. [857,880]

BÉNABOU, R. (1996): "Heterogeneity, Stratification, and Growth: Macroeconomic Implications of Community Structure and School Finance," *American Economic Review*, 86 (3), 584–609. [855]

BHATTACHARYA, D. (2009): "Inferring Optimal Peer Assignment From Experimental Data," *Journal of the American Statistical Association*, 104 (486), 486–500. [855,856,863]

CARRELL, S. E., AND J. E. WEST (2010): "Does Professor Quality Matter? Evidence From Random Assignment of Students to Professors," *Journal of Political Economy*, 118 (3), 409–432. [857]

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. [855, 857,859,860]

CARRELL, S. E., F. V. MALMSTROM, AND J. E. WEST (2008): "Peer Effects in Academic Cheating," *Journal of Human Resources*, 43 (1), 173–207. [855]

CARRELL, S. E., B. I. SACERDOTE, AND J. E. WEST (2013): "Supplement to 'From Natural Variation to Optimal Policy? The Importance of Endogenous Peer Group Formation'," *Econometrica Supplemental Material*, 81, http://www.econometricsociety.org/ecta/Supmat/10168_tables.pdf; http://www.econometricsociety.org/ecta/Supmat/10168_data_and_programs.zip. [870]

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. [855]

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. [855,856,863]

LUCAS, R. (1976): "Econometric Policy Evaluation: A Critique," in *The Phillips Curve and Labor Markets*, ed. by K. Brunner and A. H. Melzer. Amsterdam: North-Holland, 19–46. [857]

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. [855,873]

——— (2009): "The Effects of Peer Group Heterogeneity on the Production of Human Capital at West Point," *American Economic Journal: Applied Economics*, 1 (4), 69–84. [860,872]

MANSKI, C. F. (1993): "Identification and Endogenous Social Effects: The Reflection Problem," *Review of Economic Studies*, 60 (3), 531–542. [859,873]

MARMAROS, D., AND B. SACERDOTE (2006): "How Do Friendships Form?" *The Quarterly Journal of Economics*, 121 (1), 79–119. [876]

MAYER, A., AND S. L. PULLER (2008): "The Old Boy (and Girl) Network: Social Network Formation on University Campuses," *Journal of Public Economics*, 92 (1–2), 329–347. [876]

SACERDOTE, B. I. (2001): "Peer Effects With Random Assignment: Results for Dartmouth Roommates," *Quarterly Journal of Economics*, 116 (2), 681–704. [855,873]

STINEBRICKNER, R., AND T. R. STINEBRICKNER (2006): "What Can Be Learned About Peer Effects Using College Roommates? Evidence From New Survey Data and Students Form Disadvantaged Backgrounds," *Journal of Public Economics*, 90 (8–9), 1435–1454. [855]

ZIMMERMAN, D. J. (2003): "Peer Effects in Academic Outcomes: Evidence From a Natural Experiment," *The Review of Economics and Statistics*, 85 (1), 9–23. [855,861]

Dept. of Economics, UC Davis, One Shields Avenue, Davis, CA 95616, U.S.A.; secarrell@ucdavis.edu,

Dept. of Economics, Dartmouth College, 6106 Rockefeller Hall, Hanover, NH 03755, U.S.A.; Bruce.i.sacerdote@dartmouth.edu,

and

Dept. of Economics, Baylor University, One Bear Place #98003, Waco, TX 76798, U.S.A.; J_west@baylor.edu.

Manuscript received July, 2011; final revision received December, 2012.