# Peer Effects in Microenvironments: The Benefits of Homogeneous Classroom Groups

Fangwen Lu, *Renmin University of China*

Michael L. Anderson, *University of California, Berkeley and NBER*

Many believe that classroom interactions play an important role in students' academic achievement, but there is little evidence on peer effects within subclassroom groups. We exploit random seat assignment in a Chinese middle school to estimate how the gender of neighboring students affects a student's academic achievement. We find that being surrounded by five females rather than five males increases a female's test scores by 0.2–0.3 standard deviations but has no significant effects on a male's test scores. These results suggest a low-cost way to potentially improve performance within the world's largest school system.

## I. Introduction

Social interactions are believed to play an important role in students' academic achievement. Most peer effects studies in primary and secondary education define peers at the classroom or school level and test whether

This study received support from the Fundamental Research Funds for the Central Universities and the Research Funds of Renmin University of China (12XNF018). We thank Jeff Perloff and the editor for insightful comments. Any mistakes are the authors' own. Contact the corresponding author, Fangwen Lu, at lufangwen@ruc.edu.cn. Information concerning access to the data used in this article is available as supplementary material online in a zip file.

[*Journal of Labor Economics,* 2015, vol. 33, no. 1]

Submitted October 31, 2012; Accepted August 6, 2013; Electronically published November 10, 2014

students are influenced by classroom- or school-level averages. However, recent work by Carrell, Sacerdote, and West (2013) finds that students may form subgroups within larger peer groups, implying that peer effect analyses at the classroom or school level may miss important interactions within subclassroom groups.

This article examines peer effects among subgroups of grade 7 students by exploiting an experiment with random seat assignment in a Chinese middle school. As is common in most Chinese schools, students in this school stay at a fixed seat in a fixed classroom for most classes, while teachers rotate through the classrooms. In this experiment, students were assigned to blocks of rows on the basis of height and then randomly assigned to seats within blocks. This within-block randomization controls for nonrandom sorting of students into groups and allows an exploration of peer effects in a microenvironment (i.e., a subclassroom group).

We find that the gender of nearby students influences a student's performance, but the effects vary according to the student's gender. For a female student, being surrounded by five females rather than by five males increases her test scores by 0.2–0.3 standard deviations. For a male student, being surrounded by five males instead of five females does not decrease his test scores and may increase them by up to 0.1–0.3 standard deviations. These effects suggest welfare gains from rearranging students within classrooms. In comparison, there is little evidence that baseline test scores of nearby students affect academic performance.

We consider these results in the context of a large set of peer effects models proposed in the literature. We identify one model emphasizing gains from peer group homogeneity—the "boutique" model—that can plausibly generate our results. We also consider the potential mechanisms underlying our results. We reject the hypothesis that girls improve performance of other girls by reducing disruptions and instead conclude that cooperative learning behavior is the most likely mechanism underlying our results.

## II. Literature Review

An extensive theoretical literature explores different models through which academic peer effects may operate (Epple and Romano 2011). More recently, a large set of empirical peer effects studies leverage variation in peer groups at the classroom or school level to estimate peer effects (Hanushek et al. 2003; Angrist and Lang 2004; Arcidiacono and Nicholson 2005; Hoxby and Weingarth 2006; Lyle 2007; Ammermueller and Pischke 2009; Gould, Lavy, and Paserman 2009), and others explore living arrangements among college students (Sacerdote 2001; Zimmerman 2003). The empirical studies generally find evidence of positive spillovers in academic performance. However, to the best of our knowledge, no studies leverage experimental or quasi-experimental variation to estimate peer effects within subclassroom groups.

A related literature explores the effects of student gender on peer outcomes. Morse (1998) and Mael et al. (2005) review observational studies comparing students in single-sex and coeducational classes; some studies suggest that single-sex schooling may be beneficial while others indicate no difference. Hoxby (2002) and Lavy and Schlosser (2011) explore plausibly exogenous variation in the gender composition of coeducational schools and find that the proportion of female students has positive effects on students' cognitive achievements. However, gender composition does not have differential effects on boys and girls. Whitmore (2005) finds that students assigned to classrooms with higher proportions female in the Tennessee Student-Teacher Achievement Ratio experiment do better in kindergarten and second grade, with some evidence of differential effects on boys and girls.

Our study extends the rich academic peer effects literature to subclassroom groups. The results reveal that even within micro-level environments, there can be strong peer effects. This finding has policy relevance because teachers have significant discretion in organizing groups within classrooms. Implementing single-sex groups within classrooms, for example, is less controversial than implementing single-sex classrooms or single-sex schools. Changes to classroom arrangements thus represent a low-cost way to potentially improve academic performance. In addition, our results are consistent with peer effects models that favor "streaming" or homogeneity and inconsistent with peer effects models that emphasize mixing or disruptive students. While our results do not rule out the importance of mixing or disruptive students in other contexts (Figlio 2007), they suggest that disruptive students are not the main source of peer effects in our context.

## III. School Environment

This experiment was implemented in a coeducational public middle school in Jiangsu, China. At the beginning of the school year, students in grade 7—the starting grade—were assigned to a fixed classroom. They stayed in the same classroom for most classes over the semester, while teachers rotated from classroom to classroom. This arrangement is standard for this middle school and most other schools in China. The middle school is not considered an elite school and does not have special entrance requirements.

Desks and benches are provided in classrooms, typically arranged in sets of rows and columns (see fig. 1). Each desk seats two students, and there are four desks per row with aisles between desks. There are six, seven, or eight rows in each classroom depending on the number of stu-

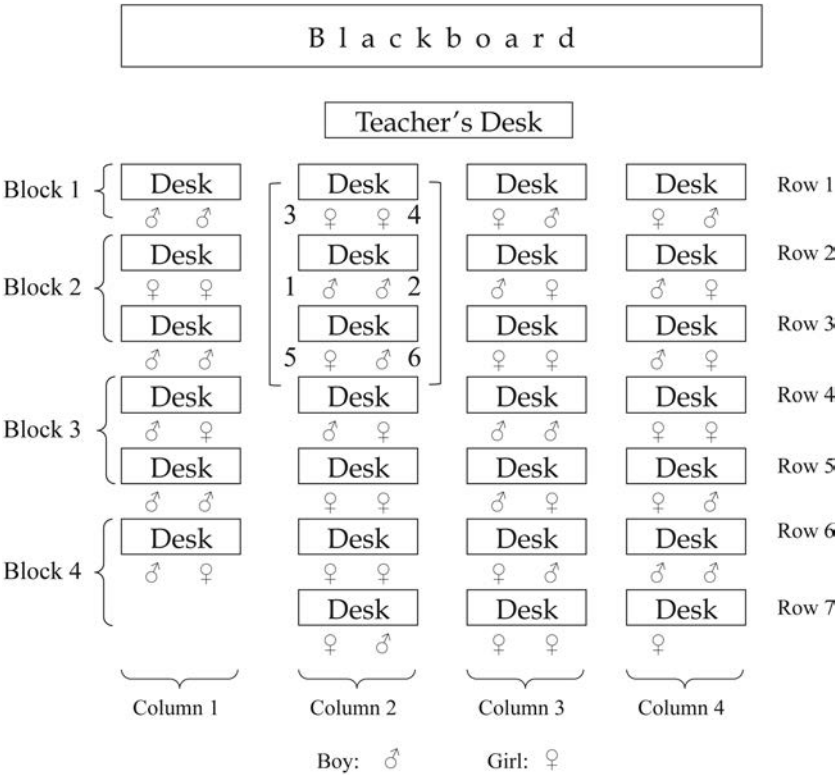

FIG. 1.—Arrangement of a typical classroom

dents. All students are assigned to fixed seats, and they must stay in their assigned seats during class time. The practice of assigning students to fixed seats helps teachers detect absentees and identify students who misbehave during class.

An administrative teacher assigns seats for the classroom.[1] In the assignment of seats, student height is a major consideration. Classrooms are typically crowded, and taller students sitting in the front may block the view of shorter students behind them. In a nonexperimental setting, the administrative teacher may have personal preferences for assigning seats. For example, some administrative teachers like to put students of the same gender together while others tend to mix genders. Seats may also be dynamically adjusted during the school year as administrative teachers learn more about students. In addition, some parents may request to have

[1] An administrative teacher is a regular teacher with additional managerial responsibilities, which include arranging class events, disciplining misbehavior, communicating with parents, and assigning student seats.

their children moved to the front of the classroom or near high-performing students.

A typical day consists of a 30-minute reading session in the early morning, four 45-minute lecture sessions in the morning, three 45-minute lecture or study sessions in the afternoon, and one 40-minute study or physical exercise session in the late afternoon. During most sessions, students must stay in their own seats. In lectures, chatting is generally prohibited. During study sessions, students choose what to study for themselves. Students are typically allowed to talk in a low voice with neighboring students during study sessions. However, seating arrangements remained fixed during study sessions, so in most cases students can communicate only with neighboring students.

Neighboring students have many opportunities to interact with each other, and different peer groups may influence students in different ways. For example, students can talk with their desk mates without moving at all, but they generally have to turn around to talk with students at adjacent desks. Desk mates can always observe each other with ease, but it is difficult to observe details across rows. Though students may interact with other students across aisles, the columns of desks in these classes rotated every several weeks. This rotation is regular school policy, and it ensures that students do not get "stuck" at the edges of classrooms. Students thus had fewer opportunities to form lasting peer groups across aisles.

Since each column of desks stayed together during this experiment, we define peer groups within columns of desks. The first peer group is the desk itself: each student has a single desk mate. The second peer group consists of "neighbor 4 students." A student's neighbor 4 peers are the two students sitting at the desk directly in front of her and the two students sitting at the desk directly behind her. Students sitting across aisles are not a relevant peer group since columns are rotated every few weeks. For students sitting in the first and last rows, their neighbor 4 peers consist of fewer than four students. The last peer group, "neighbor 5 students," consolidates the first two groups. A student's neighbor 5 peers are her neighbor 4 peers plus her desk mate.

To see a concrete example of these peer groups, consider student 1 in the second row and second column of figure 1. Student 2 is his desk mate, students 3–6 are his neighbor 4 peers, and students 2–6 are his neighbor 5 peers. For student 3, as no students sit in front of her, her neighbor 4 peers and neighbor 5 peers include only two (1 and 2) and three (1, 2, and 4) students, respectively.

The classroom layout and rotation of teachers through classrooms are typical of Chinese schools. The characteristics of children in our school, however, may not be representative of the average Chinese child. Table 1 presents summary statistics from the 2000 Chinese census comparing households in our study's area to the average Chinese household. The school in our

**Table 1**
**Comparisons between Study Areas and All Areas**

| Variable | All Urban | Study Urban | All Areas | Study Area |
|---|---|---|---|---|
| Years of education | 10.2 | 10.3 | 8.8 | 8.8 |
| | (2.8) | (2.6) | (2.8) | (2.4) |
| Education ≥9 years | .87 | .91* | .72 | .75* |
| | (.33) | (.28) | (.45) | (.44) |
| Household size | 4.0 | 4.0 | 4.2 | 4.3 |
| | (1.5) | (1.4) | (1.5) | (1.4) |
| Running water available | .77 | .75 | .40 | .42 |
| | (.42) | (.44) | (.49) | (.49) |
| Toilet available | .74 | .63 | .70 | .76* |
| | (.44) | (.49) | (.46) | (.43) |
| Households | 16,864 | 51 | 53,300 | 186 |

SOURCE.—Data are from the 0.1% sample of the 2000 Chinese census.
NOTE.—Standard deviations are in parentheses.
* Statistically different from the all urban/all areas average at $p < .05$.

study is located in an urban area, so the first two columns of table 1 compare households living in all Chinese urban areas to households living in our study's urban area. Households in our study's urban area are more educated than households in the average Chinese urban area, but they are less likely to have running water or toilets. However, these differences are modest in magnitude even when statistically significant (e.g., less than 0.25 standard deviations). The last two columns of table 1 compare all Chinese areas to our study's overall area. The differences between the last two columns are even smaller than the differences between the first two columns, perhaps because the sample sizes are larger in the last two columns. For all measures except toilet availability, the urban-rural gap is much larger than the gap between our study area and all Chinese areas. This suggests that the main issue for generalizing our results to other areas in China may be the urban-rural divide rather than the specific area in which we conducted our study.

## IV. Experimental Design

In this experiment, a research group in the local Department of Education randomly assigned students' seats with input from the authors. During the first week of the fall 2009 semester, the Department of Education requested information on students' names, gender, and heights in each classroom. The basic mechanism for assigning seats is as follows. First, students were sorted from shortest to tallest by gender within each classroom. Then, the first eight students were placed in block 1 (corresponding to row 1), the next 16 students were placed in block 2 (rows 2 and 3), and the 16-student blocks continued until all students were assigned to blocks. Students taller than 5 feet, 6.5 inches (169 centimeters), were put in a separate block. Finally, a random sequence was generated, and students were randomly permuted and assigned to seats within each block. The size of the last two blocks varies depending on the number of students and the distribution of students' heights within classrooms. Students in shorter groups always sit in front of students in taller groups, but within a block, taller students may sit in front of shorter students as a result of the randomization. This did not present challenges in the classroom as all students within the same block are of roughly similar height. As a result of the one-child family planning policy and a frequent preference for sons, the ratio of boys to girls was 1.27 in the sample school. As boys and girls were of similar height in grade 7, we placed four boys and four girls in the first block and then nine boys and seven girls (1.28 boys per girl) in subsequent blocks until it became infeasible.[2]

Some students required special seat assignments because of nearsightedness, and in some cases parents lobbied for favorable seat assignments. To increase compliance rates, the researchers allowed administrative teachers to list several student names for favorable seat treatments; students in the favored list account for 9% of all students. Students on this list received a seat assignment in either a front row or a middle column.[3] The remaining students in each block were randomly assigned seats. Normal students are thus randomly assigned with respect to their peers, but "favored" students are not randomly assigned with respect to their peers. In particular, favored students are more likely to sit adjacent to other favored students. We thus drop the outcomes for all favored students from our analysis as these students' seats are not randomly assigned (though our main results are not sensitive to including them). Favored students are still used to construct surrounding peer measures (i.e., the right-hand-side variables); however, excluding the favored students would introduce measurement error in those measures. We can summarize the assignment procedure as first nonrandomly assigning a small number of students and then randomly assigning the remaining students to the remaining seats.[4] As an additional precaution, in all regressions we control for a desk mate's favored student status and the share of neighbor 4 students who have favored status to ensure that gender and baseline test scores do not serve as proxies for

[2] Boys were 0.4 inch taller than girls, on average.

[3] For students on this list, if they were originally assigned to the first four rows, they were moved to the middle columns in the same row. If they were originally in row 5, they were moved to the middle columns in row 4. If they were originally behind row 5, they were moved to row 5.

[4] The only confounding factor that the nonrandom assignment of favored students could introduce would be a correlation between initially sitting near the center of the room and peer characteristics that favored students tend to have. However, we can control for sitting near the center of the room by including column fixed effects. Including these fixed effects does not change our results.

favored student status. Our results are robust to the inclusion or exclusion of these controls.

Administrative teachers were asked to cooperate by adopting the random seat assignments and avoiding seat adjustments over the semester. There were no financial incentives provided to administrative teachers, however, and students were not informed of the research project. It is likely that administrative teachers adjusted seat assignments during the semester so that some students were moved away from their original assignments, but there was no systematic check for compliance in seat arrangement. Strictly speaking, our estimates represent “intent to treat” effects.

## V. Data and Empirical Framework

The data for this study consist of three rounds of test scores and two rounds of surveys for students of grade 7 in fall 2009. We illustrate the data collection time line by week of semester in figure 2. The baseline test and baseline survey were administrated during the first week of the semester before random seat assignment. The random seat assignment was announced during the second week. Students sat according to the random assignment unless the administrative teachers made adjustments. For the midterm and final exams, because of the school’s efforts to prevent cheating, students were seated such that students in the same classroom were generally spread over more than 10 rooms and no student sat immediately adjacent to another student from the same class. Two teachers monitored the exams in each classroom. The postsurvey was administered right after the final exam when students were still seated according to the seat arrangement for the final exam. As students took exams and surveys in seats different from their experimental assignments, any correlations in outcomes among randomly assigned peers are not likely generated by communication among students when taking the exams or surveys.

Grading was rigorously conducted. Teachers in the same subjects allocated exam questions among themselves so that the same question was always graded by the same teacher. In addition, students’ names were hidden during the grading process. In the baseline test, the school tested students on three major subjects: Chinese, English, and math. In the midterm and final, the school tested on seven subjects: Chinese, English, math, politics, history, geography, and biology. Each of the three major subjects accounted for 150 points in the raw scores, and the other four subjects accounted for approximately 50 points each. The exam score represents the sum of all seven scores across both the midterm and the final, standardized to have mean zero and standard deviation one.[5] Figure 3 presents the

[5] All the key results are robust to analyzing the effects on midterm performance and final performance separately.

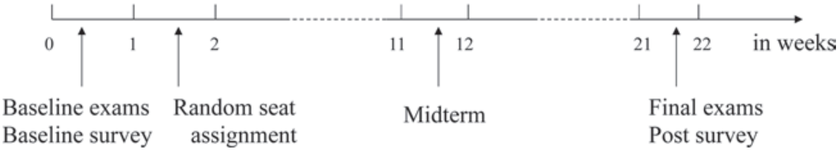

FIG. 2.—Study time line

kernel density of students' baseline scores by gender. The test scores are skewed left, and girls had higher scores overall. The distributions of midterm and final scores (not shown) have a similar shape.

In addition to the administrative data on students' gender, height, and test scores, the surveys provide information on students' family backgrounds and subjective interests. The surveys also report students' evaluations of peer influences. Panel A of table 2 presents baseline summary statistics, while panel B presents postexperiment summary statistics.

We use three types of peer groups in this study: desk mates, neighbor 4 peers, and neighbor 5 peers. For each type of peer group, we construct

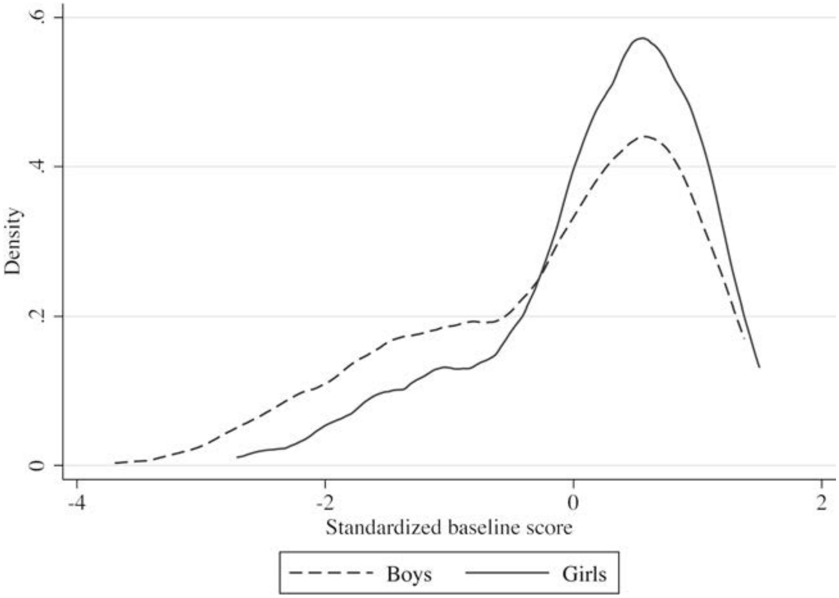

FIG. 3.—Distributions of baseline test scores by gender. The solid (dashed) line represents a kernel density plot of standardized baseline test scores for girls (boys) in our sample. Test scores are normalized to have an overall mean of zero and a standard deviation of one.

**Table 2**
**Summary Statistics**

| Variable | Observations | Mean | SD | Min | Max |
|---|---|---|---|---|---|
| A. Baseline characteristics: | | | | | |
| Female | 682 | .43 | .5 | 0 | 1 |
| Baseline test score | 680 | .00 | 1.00 | −3.69 | 1.50 |
| Favored student | 682 | .09 | .29 | 0 | 1 |
| Body height (cm) | 682 | 156.43 | 6.71 | 135 | 180 |
| Age (years) | 655 | 12.47 | .55 | 10.17 | 14.75 |
| Birth order | 680 | 1.42 | .75 | 1 | 7 |
| Father's education | 650 | 11.65 | 3.04 | 3 | 19 |
| Mother's education | 647 | 10.84 | 3.07 | 3 | 19 |
| Interest in Chinese (pre) | 643 | 4.02 | .81 | 1 | 5 |
| Interest in English (pre) | 638 | 3.93 | 1.05 | 1 | 5 |
| Interest in math (pre) | 643 | 4.05 | .85 | 1 | 5 |
| B. Postexperiment characteristics: | | | | | |
| Midterm score | 675 | .00 | 1.01 | −3.12 | 1.57 |
| Final score | 677 | .01 | 1.01 | −2.83 | 1.66 |
| Desire to change seats | 669 | 3.34 | 1.34 | 1 | 5 |
| C. Characteristics of peers: | | | | | |
| Female desk mate | 672 | .44 | .5 | 0 | 1 |
| % females in surrounding 4 students | 682 | .44 | .27 | 0 | 1 |
| % females in surrounding 5 students | 682 | .44 | .24 | 0 | 1 |
| Baseline score of desk mate | 670 | .001 | 1 | −3.69 | 1.5 |
| Average baseline score of surrounding 4 students | 682 | −.001 | .57 | −2.33 | 1.31 |
| Average baseline score of surrounding 5 students | 682 | 0 | .5 | −1.97 | 1.31 |

two measures of the peer characteristics: gender composition (whether the desk mate is female or the proportion of females among neighbor 4 and neighbor 5 peers) and baseline total test score. Panel C of table 2 presents summary statistics of peer characteristics.

Manski (1993) classifies three types of effects that can generate clustering in peer outcomes: correlated effects, endogenous effects, and exogenous effects. Correlated effects arise when similar individuals self-select into the same group. The randomized seat assignments eliminate correlated effects. Exogenous effects occur when an individual's predetermined characteristics affect the outcomes of other individuals in the same group. In this research we estimate exogenous effects of gender and baseline test scores. Endogenous effects occur when an individual's outcome affects the outcomes of other individuals in the same group. Endogenous effects are inapplicable to gender (as it is typically an immutable characteristic), but they could arise in the context of test scores. While our focus is on exogenous effects, we test for endogenous effects as well but find no significant evidence of them.[6]

The traditional method for verifying random assignment is to regress baseline characteristics of student $i$ on student $i$'s peers' characteristics. If the randomization is valid, the coefficients in these regressions should be insignificant. However, since sampling is performed without replacement, a simple bivariate regression may generate mechanical negative correlations. If student $i$ is male, for example, then it is more likely that student $i$'s peers will be female because the pool of potential peers contains one fewer female (student $i$). We address this issue by controlling for the average value of the peer characteristic among all students who were eligible to be student $i$'s peer (Guryan, Kroft, and Notowidigdo 2009). For example, when testing whether desk mate gender is correlated with student $i$'s own gender, we control for the share of females among other students in the same randomization block as student $i$. This should eliminate the mechanical negative correlation. We thus estimate regressions of the form

$$X_{i,cb} = \alpha_1 \mathrm{Peer}_{i,cb} + \alpha_2 \overline{\mathrm{Peer}}_{-i,cb} + \lambda_{cb} + u_{i,cb}. \tag{1}$$

The variable $X_{i,cb}$ represents a baseline characteristic for student $i$ in block $b$ of class $c$. The regressor of interest, $\mathrm{Peer}_{i,cb}$, represents the gender or baseline test score of student $i$'s desk mate, neighbor 4, or neighbor 5 students. The regressor $\overline{\mathrm{Peer}}_{-i,cb}$ controls for the average value of the peer characteristic among other students in the same randomization block as student $i$. The term $\lambda_{cb}$ contains block fixed effects; these fixed effects are important for identification since randomization occurs within blocks.

Statistical inference in equation (1) (and all other regressions in our article) is complicated by the clustered nature of the data. Outcomes are likely correlated within classrooms, and neighbor 4 peer measures are correlated within classrooms by construction: each student belongs to more than one neighbor 4 peer group. One solution is to cluster by classroom, but with only 12 classrooms, there are very few clusters, potentially biasing the clustered standard errors. Instead, we use our knowledge of the randomization procedure to perform exact permutation tests. These tests are derived solely from the actual randomization and thus have the appropriate size regardless of the dependence structure of the data (Rosenbaum 2007). In essence, we rerun the experiment 10,000 times and compute the resulting distribution of $t$-statistics. Under the sharp null hypothesis of no treatment effect, this distribution can be used for statistical inference. Unlike cluster bootstrap-based techniques (e.g., Cameron, Gelbach, and Miller 2008), these

[6] Endogenous effects give rise to what Manski terms the "reflection problem," or the difficulty in discerning whether student $i$'s outcome affects student $j$'s outcome or vice versa. Since we focus on exogenous effects, the reflection problem does not arise in our context.

tests remain valid even for small numbers of clusters since they are derived from the randomization procedure itself.

To implement the exact permutation tests, we randomly permute the seat assignments according to the original assignment procedure. For each permutation, we calculate $\text{Peer}_{i,cb}$ on the basis of the permuted seat assignments and estimate equation (1). We then collect the $t$-statistics from 10,000 permutations and compute the distribution of these $t$-statistics. We compare the actual $t$-statistic for a given regression to the distribution of $t$-statistics from the 10,000 random permutations. The $p$-value, reported in brackets in all tables, represents the fraction of random permutation $t$-statistics that are larger than the actual $t$-statistic.

Table 3 tests for nonrandom sorting into peer groups. The table presents results from regressions of student *i*'s predetermined characteristics on peer gender composition (or test scores), controlling for block fixed effects and the average value of gender (or test scores) among other students in the same randomization block as student *i*. Each cell represents a separate regression, and permutation-based $p$-values appear in brackets.

The subsequent rows of table 3 test for correlations between student *i*'s characteristics and peer gender or peer baseline test scores. Characteristics include height, age, birth order, mother's and father's education, and interest in Chinese, math, and English. The results are consistent with null effects for all tests. Of the 66 tests in table 3, only two tests, the relationships between age and desk mate gender and birth order and desk mate gender, are statistically significant at the 5% level ($p$ = .04 and .05 for these two tests).[7] Nevertheless, to be conservative we control for baseline characteristics in all subsequent regressions. Our conclusions are unaffected by the inclusion of these controls. If we split the sample by gender and estimate the models in table 3 separately for males and females, we still find insignificant relationships between each of the baseline characteristics and the peer measures (results available on request).

The last row of table 3 tests for correlations between attrition and peer gender or peer baseline test scores. There is a 14% attrition rate in our main regressions, but almost all of this attrition is due to missing values of baseline covariates rather than missing outcomes (the attrition rate for the midterm exam is 1% and the attrition rate for the final exam is 0.6%). We thus expect attrition to occur randomly since there is no way for seating assignment to affect the attrition of baseline covariates. Indeed, the last row confirms that there is no significant relationship between attrition and peer gender or baseline test scores.

[7] In contrast to many randomized experiments, we could not mechanically enforce covariate balance by repeating the randomization procedure until all covariate balance tests are statistically insignificant. The reason is that the randomization procedure was performed prior to the processing of the survey data.

**Table 3**
**Relationships between Peer Measures and Baseline Characteristics**

| | Peer Measure (Independent Variable) | | | | | |
|---|---|---|---|---|---|---|
| Baseline Characteristic (Dependent Variable) | Desk Mate Female | Neighbor 4 Female Share | Neighbor 5 Female Share | Desk Mate Score | Neighbor 4 Average Score | Neighbor 5 Average Score |
| Female | .011 | −.103 | −.102 | −.003 | .023 | .023 |
| | (.028) | (.068) | (.085) | (.006) | (.018) | (.021) |
| | [.700] | [.315] | [.263] | [.626] | [.227] | [.306] |
| Baseline test score | .038 | .086 | .166 | .020 | −.118 | −.135 |
| | (.038) | (.095) | (.130) | (.018) | (.084) | (.117) |
| | [.557] | [.391] | [.235] | [.302] | [.185] | [.290] |
| Height (cm) | .166 | .487 | .772 | −.079 | .155 | .119 |
| | (.195) | (.675) | (.744) | (.103) | (.236) | (.276) |
| | [.656] | [.497] | [.328] | [.474] | [.539] | [.679] |
| Age (years) | −.069 | .022 | −.071 | −.018 | .033 | .008 |
| | (.029) | (.098) | (.115) | (.027) | (.041) | (.041) |
| | [.038] | [.824] | [.557] | [.520] | [.448] | [.855] |
| Birth order | −.124 | −.102 | −.258 | .002 | −.123 | −.130 |
| | (.057) | (.139) | (.135) | (.032) | (.063) | (.071) |
| | [.049] | [.486] | [.085] | [.956] | [.074] | [.109] |
| Father's education | .164 | .398 | .743 | −.139 | .179 | .027 |
| | (.321) | (.609) | (.646) | (.144) | (.210) | (.328) |
| | [.799] | [.537] | [.280] | [.351] | [.412] | [.938] |
| Mother's education | −.059 | .658 | .671 | −.040 | −.140 | −.170 |
| | (.206) | (.328) | (.313) | (.085) | (.281) | (.329) |
| | [.778] | [.076] | [.062] | [.652] | [.629] | [.625] |
| Interest in Chinese (pre) | .056 | .022 | .071 | −.060 | .063 | −.013 |
| | (.068) | (.153) | (.131) | (.032) | (.046) | (.047) |
| | [.461] | [.889] | [.616] | [.089] | [.209] | [.788] |
| Interest in English (pre) | .126 | −.133 | .008 | −.046 | −.035 | −.095 |
| | (.068) | (.152) | (.160) | (.033) | (.062) | (.065) |
| | [.113] | [.407] | [.956] | [.199] | [.611] | [.171] |
| Interest in math (pre) | −.065 | −.075 | −.162 | −.007 | .011 | −.005 |
| | (.045) | (.119) | (.125) | (.021) | (.093) | (.104) |
| | [.162] | [.544] | [.225] | [.747] | [.916] | [.964] |
| Missing any covariate values | .009 | −.023 | −.020 | .011 | −.016 | −.002 |
| | (.035) | (.050) | (.072) | (.016) | (.044) | (.040) |
| | [.867] | [.655] | [.790] | [.505] | [.722] | [.966] |

NOTE.—Each cell represents a separate regression. The cell in row *i* and column *j* reports the results from a regression of the dependent variable in row *i* on the peer measure in column *j*. Favored students with nonrandom seat assignments are excluded. All regressions control for gender and baseline test score of other students in the same randomization block. Parentheses contain standard errors clustered by classroom. Permutation-based *p*-values are reported in brackets.

## VI. Results

### A. Main Effects of Peers on Academic Performance

Given the within-block randomization of seats, we estimate the main effects of peers using the following equation:

$$Y_{i,cb} = \beta_1 \mathrm{Peer}_{i,cb} + \gamma X_{i,cb} + \lambda_{cb} + e_{i,cb}. \tag{2}$$

The outcome $Y_{i,cb}$ represents the mean standardized midterm and final scores for student $i$ in block $b$ in class $c$. The regressor of interest, $\mathrm{Peer}_{i,cb}$, represents the gender or baseline test score of student $i$'s desk mate or neighbor 4 students. The term $X_{i,cb}$ includes all of student $i$'s baseline characteristics (gender, baseline test score, height, age, birth order, mother's and father's education, and interest in Chinese, math, and English), an indicator for whether student $i$'s desk mate is a favored student, and the share of favored students among student $i$'s neighbor 4 peers.[8] The term $\lambda_{cb}$ contains block fixed effects, which are necessary since randomization occurs within blocks. Statistical inference is performed using exact permutation tests, as described in Section V.

Table 4 presents results from estimating equation (2). Each column represents a separate regression. Column 1 reports results from regressing exam scores on a female desk mate indicator (and the controls listed above). A female desk mate increases a student's test scores by 0.07 standard deviations (standard error [SE] = 0.03). Column 2 reports results from regressing a student's exam scores on his desk mate's baseline score (plus controls). A desk mate's baseline score does not affect student $i$'s exam scores (0.02 standard deviations, SE = 0.02). Column 3 adds neighbor 4 female share to the specification in column 1, and column 4 adds neighbor 4 baseline scores to the specification in column 2. Adding these measures does not affect the coefficients on desk mate gender or desk mate baseline test scores, which is not surprising since desk mates and neighbor 4 students should be uncorrelated under random seat assignment. The coefficients on the proportion of females among neighbor 4 students and the baseline test scores of neighbor 4 students are positive (0.03 and 0.04, respectively) but statistically insignificant.

Column 5 presents a regression that includes all four peer measures simultaneously: desk mate gender, desk mate baseline score, neighbor 4 gender, and neighbor 4 baseline score. Only the coefficient on desk mate gender is statistically significant, and its magnitude is unaffected by the inclusion of desk mate or neighbor 4 baseline scores. We can rule out large effects of sitting near students with high baseline scores. For example, we

[8] Excluding all baseline characteristics except gender and baseline test score (which are necessary to control for mechanical negative correlation) does not change our conclusions.

**Table 4**
**Effects of Peer Gender and Baseline Score on Test Scores**

| | Dependent Variable: Exam Score (Midterm + Final) | | | | |
|---|---|---|---|---|---|
| Peer Measure<br>(Independent Variable) | (1) | (2) | (3) | (4) | (5) |
| Female desk mate | .070<br>(.027)<br>[.024] | | .071<br>(.027)<br>[.023] | | .065<br>(.028)<br>[.039] |
| Baseline score of desk mate | | .018<br>(.018)<br>[.344] | | .019<br>(.019)<br>[.343] | .012<br>(.019)<br>[.548] |
| Share female in neighbor 4 | | | .034<br>(.085)<br>[.702] | | .018<br>(.082)<br>[.840] |
| Average baseline score of neighbor 4 | | | | .039<br>(.030)<br>[.219] | .034<br>(.031)<br>[.299] |

NOTE.—$N$ = number of observations = 532. Each column represents a separate regression of the dependent variable on one or more of the listed peer measures. Favored students with nonrandom seat assignments are excluded from the sample. All regressions control for gender, baseline test score, proximity of favored students, height, age, birth order, parental education, and baseline interest in Chinese, English, and math. Parentheses contain standard errors clustered by classroom. Permutation-based $p$-values are reported in brackets.

can reject the hypothesis that having a desk mate who scores one standard deviation better on baseline tests increases student *i*'s performance by over 0.05 standard deviations. For neighbor 4 peers, we can reject the hypothesis that having four neighbor 4 peers who all score one standard deviation better on baseline tests increases student *i*'s performance by over 0.10 standard deviations.

## B. Heterogeneous Treatment Effects

If peer gender has similar effects on test scores for all students, then it is difficult to achieve net improvements on test scores by rearranging students. However, heterogeneous treatment effects are of policy interest because they may provide opportunities for improving aggregate achievement by rearranging students. Columns 1 and 2 of panel A in table 5 present regressions in which we estimate the effects of peer gender composition and baseline scores separately for females (col. 1) and males (col. 2). The coefficients on the female desk mate indicator are positive but insignificant for both sexes (0.03 for females and 0.06 for males). For girls, the coefficient on the share of female neighbor 4 students is positive and statistically significant (0.18, SE = 0.07). For boys, the coefficient on the share of female neighbor 4 students is negative and insignificant (−0.15, SE = 0.11). We can rule out large positive effects of neighbor 4 girls on boys' exam performance. For example, we can reject the hypothesis that moving a boy from

**Table 5**
**Effects of Peer Gender and Baseline Score on Test Scores for Females and Males**

| | Dependent Variable: Exam Score (Midterm + Final) | | |
|---|---|---|---|
| Peer Measure<br>(Independent Variable) | Female<br>Sample<br>(1) | Male<br>Sample<br>(2) | Female/Male<br>Coefficient<br>Difference<br>(3) |
| | A. Regressions with Desk Mate and Neighbor 4 Measures | | |
| Female desk mate | .029 | .061 | −.032 |
| | (.042) | (.053) | (.059) |
| | [.514] | [.286] | [.651] |
| Baseline score of desk mate | .034 | .005 | .029 |
| | (.016) | (.035) | (.040) |
| | [.057] | [.889] | [.460] |
| Share female in neighbor 4 | .182 | −.152 | .334 |
| | (.073) | (.114) | (.128) |
| | [.033] | [.221] | [.027] |
| Average baseline score of neighbor 4 | −.023 | .089 | −.112 |
| | (.061) | (.063) | (.105) |
| | [.711] | [.203] | [.228] |
| | B. Regressions with Neighbor 5 Measures | | |
| Share female in neighbor 5 | .211 | −.121 | .332 |
| | (.076) | (.139) | (.132) |
| | [.021] | [.408] | [.057] |
| Average baseline score of neighbor 5 | .012 | .101 | −.089 |
| | (.060) | (.084) | (.120) |
| | [.838] | [.260] | [.406] |
| Observations | 245 | 287 | |

NOTE.—Within each panel, each column represents a separate regression of the dependent variable on the peer measures listed in that panel. Favored students with nonrandom seat assignments are excluded from all samples. All regressions control for baseline test score, proximity of favored students, height, age, birth order, parental education, and baseline interest in Chinese, English, and math. Parentheses contain standard errors clustered by classroom. Standard errors for the difference in female and male coefficients come from a pooled regression in which every regressor is interacted with a female indicator. Permutation-based *p*-values are reported in brackets.

an all-male neighbor 4 environment to an all-female neighbor 4 environment would increase his test scores by more than 0.07 standard deviations. The coefficients on desk mate baseline score and neighbor 4 baseline scores are small and statistically insignificant for both genders.

Column 3 in table 5 reports differences in the coefficient estimates for females and males. Standard errors for these differences come from a pooled regression in which all regressors are interacted with a female indicator. The difference in the effect of female neighbors on girls and boys is large and statistically significant (0.33 standard deviations, SE = 0.13). No other differences are statistically significant.

When comparing the effects of desk mates and neighbor 4 students, we cannot reject the hypothesis that both peer types have similar effects. If a neighbor 4 peer exerts the same influence as a desk mate, then the coefficient on the share of female neighbor 4 peers should be four times larger than the female desk mate coefficient. In column 1 the coefficient on share female in neighbor 4 is 6.3 times larger than the female desk mate coefficient. This implies that, for girls, a neighbor 4 student's gender has approximately 1.5 times the impact of her desk mate's gender. We cannot rule out the possibility that this ratio equals unity. For boys, the coefficients on the share of female neighbor 4 peers and the female desk mate indicator have the opposite sign in column 2. However, both are statistically insignificant, and we cannot reject the hypothesis that the former is equal to four times the latter.

To evaluate the combined effects of desk mates and neighbor 4 students, panel B estimates regressions that replace the separate desk mate and neighbor 4 measures with a combined neighbor 5 peer measure (recall that we define student *i*'s neighbor 5 peers as her desk mate plus her neighbor 4 peers). Column 1 of panel B demonstrates that female students have positive and statistically significant effects on neighboring female students (0.21 standard deviations, SE = 0.08). This estimate implies that moving a female student from an all-boy microenvironment to an all-girl microenvironment increases her test scores by approximately 0.2 standard deviations. However, in column 2 of panel B, females have no significant effects on neighboring male students, and the point estimate is negative (−0.12, SE = 0.14). The average baseline score of neighbor 5 students also has no significant effect on exam scores for girls or boys, though we cannot rule out effects as large as 0.1 standard deviations for girls and 0.2 standard deviations for boys.

In column 3 of panel B, the difference in the effect of female neighbors on girls and boys is 0.33 standard deviations (SE = 0.13). To interpret this difference, consider a case in which girl *i* is surrounded by boys and boy *j* is surrounded by girls. Swapping the seats of girl *i* and boy *j* increases average achievement for these two students by 0.17 standard deviations (i.e., half of 0.33). The coefficients in columns 1 and 2 imply that the effects are stronger for the girl than for the boy, but we cannot reject the hypothesis that both genders benefit equally from being surrounded by students of the same gender.

Table 6 presents regressions in which we estimate peer effects separately for students with high and low baseline scores. Column 1 of panel A presents results from a sample containing only students scoring above the median on the baseline test. All the peer measures in this sample have coefficients that are close to zero and statistically insignificant. Column 2 of panel A presents results from a sample containing only students scoring below the median

**Table 6**
**Effects of Peer Gender and Baseline Score on Test Scores for High- and Low-Scoring Students**

| | Dependent Variable: Exam Score (Midterm + Final) | | |
|---|---|---|---|
| Peer Measure (Independent Variable) | High Baseline Score Sample (1) | Low Baseline Score Sample (2) | High/Low Coefficient Difference (3) |
| | A. Regressions with Desk Mate and Neighbor 4 Measures | | |
| Female desk mate | .025 | .104 | −.079 |
| | (.043) | (.048) | (.074) |
| | [.578] | [.057] | [.245] |
| Baseline score of desk mate | .001 | .016 | −.015 |
| | (.022) | (.025) | (.027) |
| | [.965] | [.542] | [.660] |
| Share female in neighbor 4 | .031 | −.114 | .145 |
| | (.053) | (.141) | (.128) |
| | [.576] | [.457] | [.362] |
| Average baseline score of neighbor 4 | .020 | .065 | −.045 |
| | (.032) | (.063) | (.066) |
| | [.548] | [.337] | [.540] |
| | B. Regressions with Neighbor 5 Measures | | |
| Share female in neighbor 5 | .045 | −.026 | .071 |
| | (.082) | (.152) | (.144) |
| | [.604] | [.870] | [.691] |
| Average baseline score of neighbor 5 | .032 | .071 | −.039 |
| | (.042) | (.072) | (.079) |
| | [.470] | [.358] | [.653] |
| Observations | 267 | 265 | |

NOTE.—Within each panel, each column represents a separate regression of the dependent variable on the peer measures listed in that panel. Favored students with nonrandom seat assignments are excluded from all samples. All regressions control for gender, baseline test score, proximity of favored students, height, age, birth order, parental education, and baseline interest in Chinese, English, and math. Parentheses contain standard errors clustered by classroom. Standard errors for the difference in low- and high-score student coefficients come from a pooled regression in which every regressor is interacted with a high baseline score indicator. Permutation-based *p*-values are reported in brackets.

on the baseline test. In this sample the effect of a female desk mate is positive and marginally significant (0.10 standard deviations, SE = 0.05), but none of the other peer measures are statistically significant. Column 3 reports differences in the coefficient estimates for students with high and low baseline scores. None of these differences are statistically significant.[9]

[9] We reach the same conclusions if we instead conduct these tests by interacting the continuous baseline score measure with the peer gender and peer baseline score measures.

Panel B presents regressions that evaluate the overall effects of neighbor 5 students separately for students with high and low baseline scores. In both subsamples there are no significant effects of either neighbor gender or neighbor baseline scores. Column 3 reports differences in the neighbor 5 coefficient estimates for students with high and low baseline scores. The differences are small and statistically insignificant. We can reject the hypothesis that the neighbor 5 baseline score coefficient for students with high baseline scores is at least 0.1 standard deviations above the same coefficient for students with low baseline scores. This implies that, in contrast with gender, there is little or no gain to segregating students by aptitude in this context. High-scoring students do not appear to help other high-scoring students more than they help low-scoring students (at least within the variation observed in our data).

Table 7 tests the robustness of our results to dropping the first and last rows from the analysis. Recall that students in the first (last) row have smaller neighbor 4 and neighbor 5 groups because there are no students sitting directly behind (ahead of) them. This fact may attenuate our estimates if the effect increases in the number of surrounding students. Dropping the first and last rows from our analysis (but not from the constructed peer measures) increases the effect sizes for both females and males. For females, the effect of moving from an all-boy microenvironment to an all-girl microenvironment—measured by the coefficient on share female of neighbor 5 students in column 1 of panel B—is now 0.29 standard deviations (SE = 0.09). For males, the effect of moving from an all-boy microenvironment to an all-girl microenvironment, reported in column 2 of panel B, is now −0.32 standard deviations (SE = 0.17). This coefficient is marginally significant, suggesting that males may also benefit from gender homogeneous microenvironments.

Table 8 presents regressions in which we separately estimate the effects of “front” and “rear” peer characteristics. We define a student’s front (rear) peers to be the two students sitting in the desk directly ahead of (behind) her.[10] Column 1 presents a regression in which we estimate the effects of front and rear peer gender composition for females. For girls, the coefficient on the share of female front peers is positive and statistically significant (0.16, SE = 0.06), and the coefficient on the share of female rear peers is positive but insignificant (0.07, SE = 0.06). Column 2 presents a regression in which we estimate the effects of front and rear peer gender composition for males. For boys, the coefficient on the share of female front peers is negative but insignificant (−0.11, SE = 0.08), and the coefficient on the share of female rear peers is negative and statistically

[10] In fig. 1, student 1’s front peers are students 3 and 4, and student 1’s rear peers are students 5 and 6.

Table 7
**Effects of Peer Gender and Baseline Score on Test Scores for Females and Males (Front and Rear Rows Dropped)**

| | Dependent Variable: Exam Score (Midterm + Final) | | |
|---|---|---|---|
| Peer Measure (Independent Variable) | Female Sample (1) | Male Sample (2) | Female/Male Coefficient Difference (3) |
| | A. Regressions with Desk Mate and Neighbor 4 Measures | | |
| Female desk mate | .040 | .031 | .009 |
| | (.067) | (.067) | (.083) |
| | [.573] | [.665] | [.927] |
| Baseline score of desk mate | .028 | .004 | .024 |
| | (.027) | (.037) | (.048) |
| | [.328] | [.916] | [.608] |
| Share female in neighbor 4 | .257 | −.320 | .577 |
| | (.075) | (.134) | (.113) |
| | [.005] | [.043] | [.002] |
| Average baseline score of neighbor 4 | −.063 | .088 | −.151 |
| | (.066) | (.072) | (.113) |
| | [.370] | [.262] | [.146] |
| | B. Regressions with Neighbor 5 Measures | | |
| Share female in neighbor 5 | .286 | −.315 | .601 |
| | (.090) | (.166) | (.150) |
| | [.010] | [.093] | [.007] |
| Average baseline score of neighbor 5 | −.017 | .097 | −.114 |
| | (.072) | (.083) | (.119) |
| | [.822] | [.278] | [.312] |
| Observations | 171 | 205 | |

NOTE.—Within each panel, each column represents a separate regression of the dependent variable on the peer measures listed in that panel. Students in front and rear rows and favored students with non-random seat assignments are excluded from all samples. All regressions control for baseline test score, proximity of favored students, height, age, birth order, parental education, and baseline interest in Chinese, English, and math. Parentheses contain standard errors clustered by classroom. Standard errors for the difference in female and male coefficients come from a pooled regression in which every regressor is interacted with a female indicator. Permutation-based *p*-values are reported in brackets.

significant (−0.19, SE = 0.06). For both genders, the coefficients on front and rear peer baseline scores are small and statistically insignificant.

Column 3 in table 8 reports differences in the coefficient estimates for females and males. The differences by gender in the effects of female front and rear peers are large and statistically significant. The effect of share female front peers is 0.27 points larger for girls than for boys (SE = 0.07), and the effect of share female rear peers is 0.26 points larger for girls than for boys (SE = 0.07). Overall, the results in table 8 suggest that girls

**Table 8**
**Effects of Front and Rear Peers on Test Scores for Females and Males**

| | Dependent Variable: Exam Score (Midterm + Final) | | |
|---|---|---|---|
| Peer Measure (Independent Variable) | Female Sample (1) | Male Sample (2) | Female/Male Coefficient Difference (3) |
| Share female in front peers | .162 | −.112 | .274 |
| | (.058) | (.077) | (.068) |
| | [.019] | [.190] | [.001] |
| Average baseline score of front peers | −.066 | .041 | −.107 |
| | (.043) | (.037) | (.063) |
| | [.154] | [.307] | [.114] |
| Share female in rear peers | .066 | −.191 | .257 |
| | (.059) | (.058) | (.070) |
| | [.290] | [.009] | [.001] |
| Average baseline score of rear peers | −.007 | .053 | −.059 |
| | (.049) | (.049) | (.076) |
| | [.893] | [.316] | [.452] |
| Observations | 180 | 215 | |

NOTE.—Each column represents a separate regression of the dependent variable on the peer measures listed in that panel. Student *i*'s front (rear) peers are the two students seated in front of (behind) student *i*. Favored students with nonrandom seat assignments are excluded from all samples. All regressions control for baseline test score, proximity of favored students, desk mate gender and baseline score, height, age, birth order, parental education, and baseline interest in Chinese, English, and math. Parentheses contain standard errors clustered by classroom. Standard errors for the difference in female and male coefficients come from a pooled regression in which every regressor is interacted with a female indicator. Permutation-based *p*-values are reported in brackets.

particularly benefit from having other girls seated in front of them, and boys particularly benefit from having other boys seated behind them.

## VII. Discussion

Our experimental results demonstrate that gender homogeneous environments can improve academic outcomes, particularly for girls. However, these "reduced-form" estimates do not reveal the structure of peer effects or the mechanisms through which they may operate. While it is impossible to identify a single model explaining our results to the exclusion of all other models, we identify one model—the "boutique" model—that can plausibly generate our results out of a larger set of models proposed in the literature. We also reject a model—the "disruptive student" model—that prima facie appears likely as an explanation for our results.

### A. Models of Peer Effects

Hoxby and Weingarth (2006) discuss a series of informal peer effects models in the context of student assignments across schools. We summarize these models here using Hoxby and Weingarth's terminology and, where possible, link them to formal models in the literature.

The first set of models specifies achievement as a function of mean peer characteristics. The simplest of these models is the "linear-in-means" model, in which student *i*'s achievement is a linear function of his or her peers' mean characteristics. The linear-in-means model is a special case of the peer effects model developed in Arnott and Rowse (1987); the more general Arnott and Rowse model assumes that peer effects may be a nonlinear function of mean peer characteristics.[11] Because it assumes that peer effects are homogeneous and operate as a linear function of mean peer characteristics, the linear-in-means model has the strong implication that organization of peer groups does not affect aggregate achievement. The more general Arnott and Rowse model implies that mixing (segregation) may improve aggregate achievement, depending on the concavity (convexity) of the function relating mean peer characteristics to student *i*'s achievement.

The second set of models relaxes the assumption that mean peer characteristics are a sufficient statistic for peer influences. These models assume that a single bad (good) student can affect her peers' achievement regardless of how many good (bad) students may offset her in the peer group. Two examples are the "bad apple" and "shining light" models. The former assumes that a single bad student disrupts learning for all his peers, while the latter assumes that a single good student inspires learning for all her peers. Lazear (2001) formalizes the bad apple model by assuming that each student has some probability $p$ of disrupting his entire group. Lazear's model implies that smaller groups should learn better and that if $p$ differs across students, total learning is maximized when students are segregated according to $p$.

A third set of models focuses on within-group heterogeneity. The boutique model assumes that students perform better when surrounded by similar peers. This occurs either because similarity allows students to more easily help each other or because teachers can tailor lessons and materials specifically for students in each group. The "focus" model posits that even a student in peer group $j$ who is different from the predominant type in peer group $j$ may benefit from homogeneity because it generates a more cohesive learning environment. The "single-crossing" model likewise implies that within-group homogeneity increases aggregate achievement, but only because high-skill students help other high-skill students more than they help low-skill students. Thus, while the gains in high-skill groups outweigh

[11] Another special case of these models is the "invidious comparison" model. In this model, student *i*'s achievement is negatively related to his peers' average achievement because higher-performing peers lower student *i*'s self-esteem. We find no evidence supporting the invidious comparison model in our results.

the losses in low-skill groups, segregation is not Pareto improving. In contrast to these models, the "rainbow" model assumes that students do better when placed in diverse groups, perhaps because they are exposed to a variety of viewpoints. Epple, Romano, and Sieg (2002, 2003) develop a formal model with similar implications. In their model, students are more likely to succeed in the workplace if their peers are representative of society at large. However, in the Epple et al. model, the benefits of heterogeneous peer groups do not become apparent until students enter the labor market.

## B. Empirical Support for Peer Effects Models

We now review which models are consistent with our empirical results. In the context of our results, we primarily consider gender as the relevant peer characteristic rather than baseline test score. Our primary specifications implement a linear-in-means model; student *i*'s achievement is a function of the proportion female and mean baseline scores among her peers. In contrast to the basic linear-in-means model, however, we allow for heterogeneous responses by gender or baseline score. Thus our specification does not restrict the organization of peer groups to have zero net effect on aggregate achievement. Our main result that females benefit from sitting near other females while males do not strongly rejects the basic linear-in-means model. In appendix table A1, we consider whether the gender-specific linear specification is sufficient or whether the effects may be nonlinear. Overall, we find no evidence of strong nonlinearities in the relationships between a female student's performance and the female share of her neighbor 5 peers or between a male student's performance and the female share of his neighbor 5 peers.[12]

Given the well-known behavioral differences between boys and girls, it is tempting to interpret our results using the bad apple model or Lazear's disruptive student framework; perhaps being surrounded by girls improves performance because girls are less likely to misbehave. However, our results are inconsistent with these models. If boys were detrimental to learning because they were more likely to be disruptive, then we would expect student *i* to perform worse when surrounded by boys regardless of

[12] Appendix table A1 explores whether the effect of neighboring female students is linear in female share. We modify eq. (1) to include dummy variables for four categories: female share of neighbor 5 students is 20%–40%, female share of neighbor 5 students is 40%–60%, female share of neighbor 5 students is 60%–80%, and female share of neighbor 5 students is 80%–100%. The omitted category is a female share of neighbor 5 students of 0%–20%. The dummy variable coefficients are estimated with limited precision, but for female students, there appears to be a general upward trend in the female share coefficients. Overall, there is no evidence of strong nonlinearities in the relationship between a female student's performance and the female share of her neighbor 5 peers. There is also no evidence of a relationship between a male student's performance and the female share of his neighbor 5 peers.

whether student *i* is male or female. Our results, however, suggest that, if anything, males perform better when surrounded by other males. For similar reasons the shining light model is also inconsistent with our data: if girls were more "inspirational" to their fellow students than boys, we would expect performance for both genders to increase when surrounded by females. We do not find this.

Another implication of Lazear's disruptive student framework is that what affects student *i*'s performance is not the share of potentially disruptive students nearby but the number of potentially disruptive students nearby. To test this prediction we estimate two specifications. In the first specification, student *i*'s performance is a function of the share of males among neighbor 5 students (similar to cols. 1 and 2 of table 5, panel B). In the second specification, student *i*'s performance is a function of the number of males among neighbor 5 students. If the disruptive student framework applies, then we expect the second specification to have better explanatory power than the first. In actuality, the reverse is true. Among boys, the first specification generates a partial $R^2$ of .005 while the second specification generates a partial $R^2$ of .004. Among girls, the first specification generates a partial $R^2$ of .022 while the second specification generates a partial $R^2$ of .012. We thus find no evidence to support the disruptive student framework.[13]

Models focusing on within-group heterogeneity are most relevant to our results. Our results are clearly inconsistent with the rainbow model since gender heterogeneity decreases performance. Girls do significantly better when seated near other girls, and boys do weakly better when seated near other boys. The single-crossing model is also inconsistent with our results. In the single-crossing model, sitting next to girls is beneficial for both girls and boys, but it is more beneficial for girls than for boys. Segregation should thus increase total achievement and achievement among girls, but it should decrease achievement among boys. However, we find no evidence that segregation decreases performance among boys; if anything, it appears to increase performance.

The boutique and focus models are the most consistent with our results. Both models predict gains to segregation for girls and boys. Discriminating between these two models is difficult, however, because they share

[13] Of course, we cannot rule out alternative explanations for these results. The only reason the number of male neighbor 5 peers is not a linear transformation of the share male of neighbor 5 peers is that students at the front and back of the classroom have only three peers instead of five. However, students in front (back) rows are also shorter (taller) than students in middle rows. It is thus possible that the model with number of males is in fact the correct model but that the peer effects in this case are modified by height (or some characteristic correlated with height) in such a way that makes the model with share male fit better. While we view this explanation as somewhat unlikely, we cannot eliminate it.

similar predictions. The key distinction between the two is that the focus model predicts that even a student in peer group *j* who is different from the predominant type in peer group *j* may benefit from homogeneity because the learning environment is more cohesive. To differentiate between the two models, we therefore estimate two specifications. In the first specification, student *i*'s performance is a function of the share of females among neighbor 5 students (similar to cols. 1 and 2 of table 5, panel B). In the second specification, student *i*'s performance is a function of the standard deviation of the female indicator in a group composed of the neighbor 5 students plus student *i*. The idea is that if homogeneity is beneficial because it generates a more cohesive learning environment regardless of student *i*'s gender, then the second specification should have more explanatory power than the first. For example, the second specification allows a female student to perform better when 100% of neighboring students are male than when 50% of neighboring students are male. Under the focus model, this occurs because the former environment is more homogeneous than the latter. In actuality, however, the second specification does not fit the data well. Among boys, the first specification generates a partial $R^2$ of .005 while the second specification generates a partial $R^2$ of .000. Among girls, the first specification generates a partial $R^2$ of .022 while the second specification generates a partial $R^2$ of .007. We thus conclude that only the boutique model, among the models we consider, is consistent with our data.

While our results suggest test score gains from segregation, the Epple et al. model implies that these gains may come at some future cost. Peer effects in the Epple et al. model appear only after students enter the labor market. At that point, workplace success is more likely when students have experience working with peers who are representative of society at large. It is thus possible that segregation inhibits development of some noncognitive skills even as it increases cognitive skills. While we have no empirical evidence regarding this hypothesis, it is worth noting as a limitation when considering the overall costs and benefits of segregating classroom groups.

### C. Mechanisms Underlying Peer Effects

Our results appear most consistent with the boutique model of peer effects. However, the model itself describes a data-generating process, and there are several mechanisms that may underlie the data-generating process. Three mechanisms proposed in the classroom context are the opportunity for teachers to tailor lessons and materials toward the specific student type (in this case male or female), a decrease in disruptive behavior or confusion within homogeneous groups, and an increase in cooperative learning behavior—or positive interactions—within homogeneous groups.

In contrast to previous studies that define peer groups at the class level, our focus on peer microenvironments rules out the first mechanism, tailoring of lesson plans (a mechanism suggested by the name "boutique" itself). It is unlikely that teachers can tailor lesson plans to individual clusters since there are roughly 10 nonoverlapping clusters of students in each classroom.[14] We also find little evidence supporting the second mechanism, a reduction in confusion within homogeneous groups. This mechanism is more consistent with the focus model, which our results in the previous section suggest does not fit our data. Cooperative learning behavior among students of the same gender is thus the most plausible mechanism underlying our results.

For further evidence of cooperative learning behavior in gender homogeneous groups, we consider the endpoint survey questions. The endpoint survey includes several questions on the relationship between the surveyed student and his desk mate. Three questions of relevance are as follows: (1) How frequently does the surveyed student communicate with her desk mate? (2) How strongly does the surveyed student wish to remain in her current seat? (3) How well can the surveyed student concentrate in class? Appendix table A2 presents results from regressing each of these measures on desk mate and neighbor 4 gender and academic background. We summarize four notable findings here.

First, moving a girl from a microenvironment in which she is surrounded by four boys to a microenvironment in which she is surrounded by four girls reduces the reported frequency of communication with her desk mate by 0.40 standard deviations (SE = 0.15). An intuitive explanation for this result is that if a girl communicates more frequently with her neighbor 4 peers because they are female, this may crowd out communication with her desk mate. However, since the survey does not ask about communication with neighbor 4 peers, we cannot directly test this hypothesis. Second, boys communicate less frequently with female desk mates than with male desk mates (−0.21 standard deviation effect, SE = 0.09). This result is consistent with gender homogeneity encouraging communication, though it does not explain why males do not appear to benefit from male desk mates. Third, females express a stronger desire to remain in their current seats when they have female desk mates (0.28 standard deviation effect, SE = 0.13). This suggests that females appreciate being seated next to other females. However, there is no significant effect of neighbor 4 gender on seating satisfaction for females, so the evidence is not uniformly consistent. Finally, neither girls nor boys report better ability to concentrate when sitting next to or nearby girls. This further suggests that the relevant mechanism is not a reduction in disruptive behavior.

[14] Recall also that all specifications include classroom fixed effects, so homogeneity at the cluster level is not confounded with homogeneity at the classroom level.

Overall, the survey results offer suggestive evidence supporting the hypothesis that gender homogeneous groups improve outcomes through cooperative learning behavior. There are important caveats, however: statistical power is limited, the results are not uniformly consistent, and the survey does not ask about interactions with neighbor 4 peers.

These caveats notwithstanding, cooperative learning behavior also provides a possible explanation for potential differences in the effects of desk mates and neighbor 4 peers. The overall pattern in panel A of tables 5 and 7 suggests that girls benefit from neighbor 4 females, boys benefit from neighbor 4 males, but neither gender benefits strongly from having a desk mate of the same gender. Communication with desk mates is difficult to avoid, but communication with neighbor 4 peers is voluntary and thus may be more dependent on the quality of peer relationships. If students of the same gender are more likely to communicate with each other (as suggested by the first two columns of app. table A2), we might expect larger gains from having neighbor 4 peers of the same gender than from having desk mates of the same gender.

A final consideration is that many of the students in our study are going through puberty (their ages range from 10 to 14). They therefore may be developing an interest in the opposite sex at this juncture and may spend time trying to communicate with classmates of the opposite sex regarding nonacademic matters. If so, students in gender heterogeneous environments may not be disruptive per se, but they may spend less time discussing academic topics than students in gender homogeneous environments. This possibility would represent a hybrid of a reduction in "disruptive" behavior and an increase in "cooperative" learning behavior. It is consistent with the evidence in table 8, which suggests that girls benefit from sitting behind other girls and boys benefit from sitting in front of other boys. If boys turn around to engage in conversation with girls behind them, this could reduce the performance of the boys that are turning around and the girls that they engage. In contrast, boys seated behind girls have no opportunity to even make eye contact with the students in front of them. If this mechanism were important, it would suggest that the external validity of our results could be limited when considering students who have not yet entered puberty.

## VIII. Conclusion

We identify peer effects within subgroups inside classrooms by exploiting the random assignment of seats in a Chinese middle school. The results suggest that while having a female desk mate may be beneficial for both boys and girls, having more female neighbors has significant positive effects on girls but potential negative impacts on boys. The differing patterns between the desk mate results and the neighbor 4 results may be due to differences in interactions: interactions between desk mates are easier and to some degree unavoidable, while interactions with neighboring students are voluntary. The most plausible mechanism underlying our findings is the possibility of cooperative learning behavior among students of the same gender.

It is interesting to compare the results of this study to the results of classroom-level and school-level studies on gender peer effects. Whitmore (2005) finds that increasing the classroom female share by 20 percentage points increases kindergarten test scores by approximately 0.1 standard deviations. In comparison, we find that increasing the female share of neighboring students by 20 percentage points increases female test scores by approximately 0.04 standard deviations. Whitmore finds no difference in effects for males and females in kindergarten but a large difference in effects for males and females in third grade: increasing the female share by 20 percentage points increases female test scores by 0.13 standard deviations but decreases male test scores by 0.16 standard deviations. The implied benefits of gender homogeneity in third grade are consistent with our results.

Lavy and Schlosser (2011) use Israeli data to measure the effect of the fraction female within a school grade level on peer test scores. In eighth grade (the grade closest to our study), they find that a 20 percentage point increase in the female share increases female test scores by 0.06–0.08 standard deviations on average. These effects are 50%–100% larger than our equivalent effects for neighboring female students. They find no significant effect of female share on eighth-grade male test scores. However, a 20 percentage point increase in female share raises high school test scores by 0.04–0.05 standard deviations for both males and females. These effects are similar in magnitude to our effects (for females), but the absence of heterogeneous effects by gender is in contrast to our results.

External validity is a key issue in our study as we focus on one middle school in China. Table 1 establishes that our study's urban area resembles the typical Chinese urban area on several dimensions, but it also reveals the substantial urban-rural divide in China. We would caution against drawing specific conclusions regarding effect sizes in rural Chinese areas or at much younger ages. We are likewise hesitant to extrapolate our results to other countries. The Chinese education system is heavily structured around the classroom and includes in-class study sessions. Furthermore, students stay in the same seat throughout the day and do not switch rooms. This stands in sharp contrast to systems in the United States and the United Kingdom, where middle school students change classrooms many times throughout the day. These frequent room changes ensure that a student's classroom peers are constantly changing, likely affecting how peer relationships affect performance.

These caveats notwithstanding, our results demonstrate the potential for net test score gains from improving classroom arrangements within the

Chinese context. This finding suggests a low-cost way to improve test scores within a significant segment of the world's largest education system, and it underscores the potential return to further research on peer effects within subclassroom microenvironments.

## Appendix

**Table A1**
**Nonlinear Effects of Peer Gender and Baseline Score on Test Scores for Females and Males**

| | Dependent Variable: Exam Score (Midterm + Final) | |
|---|---|---|
| Peer Measure (Independent Variable) | Female Sample | Male Sample |
| 20%–40% females in neighbor 5 | .059 | −.002 |
| | (.118) | (.172) |
| | [.643] | [.991] |
| 40%–60% females in neighbor 5 | .021 | −.058 |
| | (.107) | (.163) |
| | [.857] | [.747] |
| 60%–80% females in neighbor 5 | .155 | −.042 |
| | (.086) | (.174) |
| | [.115] | [.828] |
| 80%–100% females in neighbor 5 | .192 | −.127 |
| | (.120) | (.236) |
| | [.168] | [.630] |
| Average baseline score of neighbor 5 | .016 | .100 |
| | (.061) | (.085) |
| | [.794] | [.273] |
| Observations | 245 | 287 |

NOTE.—Each column represents a separate regression of the dependent variable on the listed peer measures. The omitted category is students with 0%–20% females among neighbor 5 peers. Favored students with nonrandom seat assignments are excluded from all samples. All regressions control for baseline test score, proximity of favored students, height, age, birth order, parental education, and baseline interest in Chinese, English, and math. Parentheses contain standard errors clustered by classroom. Permutation-based *p*-values are reported in brackets.

**Table A2**
**Effects on Communication, Seating Satisfaction, and Focus**

| | Dependent Variable | | | | | |
|---|---|---|---|---|---|---|
| | Frequency of Communication with Desk Mate (Standardized) | | Desire to Remain in Current Seat (Standardized) | | Ability to Concentrate in Class (Standardized) | |
| Peer Measure (Independent Variable) | Female Sample | Male Sample | Female Sample | Male Sample | Female Sample | Male Sample |
| Desk mate female | .024 | −.206 | .280 | .057 | −.049 | −.055 |
| | (.153) | (.094) | (.128) | (.219) | (.091) | (.113) |
| | [.880] | [.052] | [.059] | [.803] | [.613] | [.642] |
| Desk mate's baseline score | .036 | .219 | −.015 | .035 | −.022 | .047 |
| | (.057) | (.101) | (.091) | (.063) | (.066) | (.074) |
| | [.556] | [.058] | [.870] | [.595] | [.759] | [.548] |
| Share female in neighbor 4 | −.400 | .009 | −.200 | .039 | .043 | .112 |
| | (.153) | (.236) | (.192) | (.212) | (.239) | (.221) |
| | [.026] | [.971] | [.333] | [.859] | [.866] | [.630] |
| Average baseline score of neighbor 4 | .130 | −.007 | .226 | −.192 | −.024 | .096 |
| | (.149) | (.195) | (.147) | (.125) | (.136) | (.120) |
| | [.411] | [.972] | [.171] | [.160] | [.871] | [.448] |
| Observations | 244 | 281 | 245 | 285 | 241 | 278 |

NOTE.—Each column represents a separate regression of the dependent variable on the listed peer measures. Favored students with nonrandom seat assignments are excluded from all samples. All regressions control for baseline test score, proximity of favored students, height, age, birth order, parental education, and baseline interest in Chinese, English, and math. Parentheses contain standard errors clustered by classroom. Permutation-based $p$-values are reported in brackets.

## References

Ammermueller, Andreas, and Jörn-Steffen Pischke. 2009. Peer effects in European primary schools: Evidence from the Progress in International Reading Literacy Study. *Journal of Labor Economics* 27, no. 3:315–48.

Angrist, Joshua D., and Kevin Lang. 2004. Does school integration generate peer effects? Evidence from Boston's Metco program. *American Economic Review* 94, no. 5:1613–34.

Arcidiacono, Peter, and Sean Nicholson. 2005. Peer effects in medical school. *Journal of Public Economics* 89, nos. 2–3:327–50.

Arnott, Richard, and John Rowse. 1987. Peer group effects and educational attainment. *Journal of Public Economics* 32, no. 3:287–305.

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller. 2008. Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics* 90, no. 3:414–27.

Carrell, Scott E., Bruce I. Sacerdote, and James E. West. 2013. From natural variation to optimal policy? The Lucas critique meets peer effects. *Econometrica* 81, no. 3:855–82.

Epple, Dennis, and Richard Romano. 2011. Peer effects in education: A survey of the theory and evidence. In *Handbook of social economics*, ed. Jess Benhabib, Alberto Bisin, and Matthew O. Jackson. San Diego, CA: Elsevier.

Epple, Dennis, Richard Romano, and Holger Sieg. 2002. On the demographic composition of colleges and universities in market equilibrium. *American Economic Review* 92, no. 2:310–14.

———. 2003. Peer effects, financial aid and selection of students into colleges and universities: An empirical analysis. *Journal of Applied Econometrics* 18, no. 5:501–25.

Figlio, David N. 2007. Boys named Sue: Disruptive children and their peers. *Education Finance and Policy* 2, no. 4:376–94.

Gould, Eric D., Victor Lavy, and M. Daniele Paserman. 2009. Does immigration affect the long-term educational outcomes of natives? Quasi-experimental evidence. *Economic Journal* 119, no. 540:1243–69.

Guryan, Jonathan, Kory Kroft, and Matthew J. Notowidigdo. 2009. Peer effects in the workplace: Evidence from random groupings in professional golf tournaments. *American Economic Journal: Applied Economics* 1, no. 4:34–68.

Hanushek, Eric A., John F. Kain, Jacob M. Markman, and Steven G. Rivkin. 2003. Does peer ability affect student achievement? *Journal of Applied Econometrics* 18, no. 5:527–44.

Hoxby, Caroline M. 2002. How does the makeup of a classroom influence achievement? *Education Next* 2, no. 2:56–63.

Hoxby, Caroline M., and Gretchen Weingarth. 2006. Taking race out of the equation: School reassignment and the structure of peer effects. Unpublished manuscript, Department of Economics, Harvard University.

Lavy, Victor, and Analía Schlosser. 2011. Mechanisms and impacts of gender peer effects at school. *American Economic Journal: Applied Economics* 3, no. 2:1–33.

Lazear, Edward P. 2001. Educational production. *Quarterly Journal of Economics* 116, no. 3:777–803.

Lyle, David. 2007. Estimating and interpreting peer and role model effects from randomly assigned social groups at West Point. *Review of Economics and Statistics* 89, no. 2:289–99.

Mael, Fred, Alex Alonso, Doug Gibson, Kelly Rogers, and Mark Smith. 2005. *Single-sex versus coeducational schooling: A systematic review*. Jessup, MD: US Department of Education.

Manski, Charles F. 1993. Identification of endogenous social effects: The reflection problem. *Review of Economic Studies* 60, no. 3:531–42.

Morse, Susan. 1998. *Separated by sex: A critical look at single-sex education for girls*. Washington, DC: American Association of University Women.

Rosenbaum, Paul R. 2007. Interference between units in randomized experiments. *Journal of the American Statistical Association* 102:191–200.

Sacerdote, Bruce. 2001. Peer effects with random assignment: Results for Dartmouth roommates. *Quarterly Journal of Economics* 116, no. 2:681–704.

Whitmore, Diane. 2005. Resource and peer impacts on girls' academic achievement: Evidence from a randomized experiment. *American Economic Review* 95, no. 2:199–203.

Zimmerman, David J. 2003. Peer effects in academic outcomes: Evidence from a natural experiment. *Review of Economics and Statistics* 85, no. 1:9–23.