Understanding Peer Effects
– On the Nature, Estimation and Channels of Peer Effects[*]

July, 2014

Jan Feld [a]

Ulf Zölitz [b]

Abstract

This paper provides evidence on ability peer effects in university education. Identification comes from the random assignment of students to sections. We find that students on average benefit from better-ability peers. Low-ability students, however, are harmed by high-ability peers. We introduce a placebo analysis that provides a simple test to quantify the estimation bias driven by the mechanisms described in Angrist (2013). In our setting, the bias is small and does not drive our results. Analyzing students' course evaluations suggests that peer effects are driven by improved student interaction rather than adjustments in teachers' behavior or students' effort.

Keywords: Peer effects, higher education, estimation bias

JEL classification: I21, I24, J24

[*] We thank Sandra Black, Lex Borghans, Andries de Grip, Monique de Haan, Thomas Dohmen, Armin Falk, David Figlio, Bart Golsteyn, Jonathan Guryan, Daniel Hamermesh, Randi Hjalmarsson, Olivier Marie, Nicolás Salamanca, Benedikt Vogt and participants at various seminars and conferences for helpful discussions and comments. We further thank Joël Castermans, Sanne Klasen and Kim Schippers from the SBE Scheduling Department, Sylvie Kersten from the SBE Exams Office, and Jeannette Hommes and Paul Jacobs from the Educational Research and Development Department for providing data and valuable background information.

[a] Department of Economics, Gothenburg University, P.O. Box 640, SE 40530 Gothenburg, Sweden, jan.feld@economics.gu.se.

[b] Department of Economics, Maastricht University, P.O. Box 616, 6200 MD, Maastricht, The Netherlands, u.zoelitz@maastrichtuniversity.nl.

1 Introduction

The promise of the peer effects literature is to provide policy makers with advice that can be used to increase overall student performance simply by reorganizing peer groups. To deliver on this promise, we need to not only show that peer effects exist but also understand their nature. In particular, we need to understand the non-linearities of peer effects because these have to be exploited to improve overall student performance (Hoxby, 2000). At the moment, we are far from delivering this promise. This was recently demonstrated by Carell, Sacerdote and West (2013), who conducted an experiment that aimed at increasing the performance of low-ability students by reorganizing peer groups. Their design was motivated by pre-experimental findings in the same setting that promised a Pareto improvement in outcomes. The reorganization of peers, however, had the opposite effect: low-ability students were actually harmed by the intervention. Carrell et al. (2013) explain this finding with new patterns of social interactions caused by the new peer group assignment. Conversely, Angrist (2013) suggests that their initial findings were a statistical artifact and therefore did not have any predictive value.

The main reason why peer effects are difficult to study is that there are a number of social and statistical forces that lead to similar outcomes between peers, even in the absence of causal peer effects (Manski, 1993; Angrist, 2013). There are three main empirical challenges to estimating peer effects: the selection problem, the reflection problem, and what we will call the "Angrist mechanics." The selection problem states that peer groups are usually formed endogenously and that it is empirically difficult to distinguish peer effects from selection effects. The reflection problem states that it is impossible to distinguish the effect of peers on the individual from the effect of the individual on peers if both are determined simultaneously. The Angrist mechanics state that there is a mechanical relationship between the measures of own ability and peer ability that might lead to biased estimates, even in settings where peers are assigned randomly (Angrist, 2013). A number of recent peer effects studies (Lyle, 2007; Carrell, Fullerton, & West, 2009; Duflo, Dupas, & Kremera, 2011; Carrell et al., 2013) have convincingly addressed the selection and reflection problems by studying peer effects in a setting where students are randomly assigned to peer groups and using pre-treatment characteristics as measures for peer ability. The "Angrist mechanics," however, remain a general threat to the validity of all peer effects studies. In this paper, we exploit the random assignment of students to sections to study ability peer effects at the university level. Our dataset consists of all students enrolled at the School of Business and Economics (SBE) at Maastricht University over a period of three years, which amounts to 7,740 students and 41,608 student grades. Course participants are assigned to sections, groups of 10 to 15 students, which spend most of their contact hours together in one classroom. Our measure of student performance is course grades. Following the standard approach in the literature to avoid the reflection problem, we use a pre-treatment indicator of peer quality: the past GPA of the peers. To identify potential non-linearities in peer effects, we estimate heterogeneous effects in terms of student and peer ability. We address the Angrist critique by redoing our analysis with randomly assigned "placebo sections" that consist exclusively of peers that were not in the actual section. We repeat this procedure 1,000 times in a bootstrap-like procedure to obtain a distribution of placebo-peer-effects estimates. The intuition behind this placebo analysis is that the peer effects estimates obtained from "placebo sections" will only reflect mechanical forces and that their size and sign therefore can inform us about the degree to which our main findings reflect effects of real social interactions. The difference between the average of the peer effects estimates in the placebo setting and the estimates from the actual peer assignment can be interpreted as the causal peer effect corrected for the mechanical bias.

Our results for the linear-in-means specification show that being assigned to a section with on average higher-ability peers increases students' grades in that course by a statistically significant but small amount. A one standard deviation increase in the average peer GPA causes an increase in an individual's grade of 1.15 percent of a standard deviation. These results mask important heterogeneity: low-ability students are actually harmed by high-ability peers. The placebo analysis we develop shows that our results are not purely a result of the "Angrist mechanics." We show that the mechanical estimation bias exists and amounts to 9 percent of the linear-in-means coefficient in our setting. The bias in the non-linear model we estimate varies in size and direction. Analyzing students' individual course evaluations, we find suggestive evidence that the main channel of the observed peer effects is improved group interaction. We find no evidence for an adjustment in teachers' behavior or student effort driven by the classroom peer composition.

This article makes three main contributions. First, we provide clean estimates of peer effects using a large dataset of randomly assigned students. Second, we introduce a placebo analysis that can be used to determine the size and direction of the "Angrist mechanics" in other settings. Third, we are among the first to provide evidence of the underlying channels of peer effects using detailed student level course evaluations.[1]

Peer effects in education have been studied in a vast number of different contexts with various definitions of peer groups. The most convincing studies on peer effects have exploited

[1] See Lavy, Paserman and Schlosser (2012) for evidence on channels of peer effects in middle and high schools using student questionnaires.

random assignment of students to peer groups.[2] Although many studies have considered heterogeneous peer effects, only a few studies have investigated the heterogeneity of peer effects in terms of student and peer ability at the same time. In addition to the study of Carrell et al. (2013) discussed above, Burke and Sass (2013) study peer effects in pre-tertiary education in all public schools in Florida. They find substantial heterogeneity, which could potentially be exploited to increase overall student performance. In their setting, however, students were not randomly assigned to peer groups, and the "Angrist mechanics" were not addressed.

The remainder of the paper is structured as follows. Section 2 describes the institutional environment and the assignment procedure of students to sections. Section 3 discusses the dataset. Section 4 provides evidence that the assignment to sections is random, conditional on scheduling constraints. Section 5 discusses the empirical challenges of estimating peer effects. Section 6 shows the empirical strategy and the baseline results. Section 7 shows the placebo analysis that addresses the "Angrist mechanics." In Section 8, we estimate heterogeneous effects. Section 9 investigates underlying channels using students' course evaluations. Section 10 concludes the article.

[2] This, however, has often only been done in very particular situations or for very particular peer groups. At the university level, studies have exploited (conditionally) random assignments of students to sections (De Giorgi, Pellizzari, & Woolston, 2012), dorm rooms (e.g. Sacerdote, 2001; Zimmerman, 2003; Brunello, De Paola, & Scoppa, 2010) and living communities in military colleges (Lyle, 2007; Carrell et al., 2009; Carrell et al., 2013). At the pre-university level, Duflo, Dupas and Kremer (2011) have randomly assigned students to classes in elementary schools in rural Kenya.

2 Background

2.1 Institutional Environment

The School of Business and Economics (SBE) of Maastricht University is located in Maastricht, a city in the south of the Netherlands.[3] Currently there are approximately 4,200 students at the SBE enrolled in Bachelor's, Master's, and PhD programs. Because of its proximity to Germany, the SBE has a large German student population (53 percent) mixed with students of Dutch (33 percent) and other nationalities. Approximately 37 percent of the students are females. The academic year at the SBE is divided into four regular teaching periods of two months and two skills periods of two weeks. Students usually take two courses at the same time in the regular periods and one course in the skills period. We exclude courses in skills periods from our analysis because these are often not graded and we could not always identify the relevant peer group.[4]

The courses are organized by course coordinators, mostly senior staff, and most of the teachers are PhD students and teaching assistants. Each course is divided into sections with a maximum of 16 students. These sections are the peer group on which we focus. The course size ranges from 1 to 638 students, and there are 1 to 43 sections per course. The sections usually meet in two weekly sessions of two hours each. Most courses also have lectures that are followed by all students in the course and are usually given by senior staff.

The SBE differs from other universities in its focus on Problem Based Learning (PBL).[5] The general PBL setup is that students generate questions about a topic at the end of one session and then try to answer these questions through self-study. In the next session, the findings are

[3] See also Feld, Salamanca and Hamermesh (2013) for a detailed description of the institutional background and examination procedure at the SBE.

[4] In some skills courses, for example, students are scheduled in different sections but end up sitting together in the same room. Further, skills courses have no exam at the end of the skill period, and in many skills course, students do not receive a GPA-relevant grade but only a "pass" or a "fail" grade.

[5] See http://www.umpblprep.nl/ for a more detailed explanation of PBL at Maastricht University.

discussed with other students in the section. In the basic form of PBL, the teacher takes only a guiding role, and most of the studying is done by the students independently. Courses, however, differ in the extent to which they give guidance and structure to the students, which depends on the nature of the subject covered, with more difficult subjects usually requiring more guidance, and the preference of the course coordinator and teacher.

Compared to the traditional lecture system, the PBL system is arguably more group focused because most of the teaching happens in small groups in which group discussions are the central part of the learning process. Much of the students' peer interaction happens with members of their section, either in the sessions, during work for common projects, or in homework and study groups.

2.2 Students' Course Evaluations

Two weeks before the exam, students are invited by email to evaluate the courses they are currently taking in an online questionnaire.[6] Students receive up to three email reminders, and the questionnaire closes before the day of the exam. Students are assured that their individual answers will not be passed on to anyone involved in the course. The teaching staff receives no information about the evaluation before they have submitted the final course grades to the examination office.[7] This "double blind" procedure is implemented to avoid a situation where either of the two parties retaliates with negative feedback in the form of lower grades or evaluations. The exact length and content of the online questionnaires differ by course. The questionnaire typically contains 19-25 closed questions and two open questions. For our analysis,

[6] For more information, see the course evaluation home page: http://iwio-sbe.maastrichtuniversity.nl/default.asp.

[7] After exam grades are published, teaching staff receive the results of the courses evaluations aggregated at the section level.

we use the 19 core questions that are assessed in most courses. These standard questions ask students to evaluate different course aspects such as teacher performance, group functioning, learning materials, and general course organization and to state hours they spend on self-study.

2.3 Assignment of Students to Sections

The Scheduling Department of the SBE assigns students to sections, teachers to sections, and sections to time slots. Before each period, there is a time frame in which students can register online for the courses they want to take. After the registration deadline, the scheduler is given a list of registered students and allocates the students to sections using a computer program. About ten percent of the slots in each group are initially left empty and are filled with students who register late.[8] This procedure balances the amount of late registration students over the sections. Before the start of the academic year 2010/11, the section assignment for Master's courses and for Bachelor's courses was conducted with the program Syllabus Plus Enterprise Timetable using the allocation option "allocate randomly" (see Figure A1 in the Appendix). Since the academic year 2010/11, all Bachelor's sections have been stratified by nationality with the computer program SPASSAT.[9] Some Bachelor's courses are also stratified by exchange student status. After the assignment of students to sections, the sections are assigned to time slots, and the program Syllabus Plus Enterprise Timetable indicates scheduling conflicts.[10] Scheduling conflicts

[8] About 5.6 percent of students register late. The number of late registrations in the previous year determines the number of slots that are initially left unfilled by the scheduler.

[9] The stratification goes as follows: the scheduler first selects all German students (who are not ordered by any observable characteristic) and then uses the option "Allocate Students set SPREAD," which assigns an equal number of German students to all classes. Then, the scheduler repeats this process with the Dutch students and lastly distributes the students of all other nationalities to the remaining spots.

[10] There are four reasons for scheduling conflicts: (1) the student takes another regular course at the same time; (2) the student takes a language course at the same time; (3) the student is also a teaching assistant and needs to teach at the same time; and (4) the student indicated non-availability for evening education. By default, all students are

arise for approximately 5 percent of the initial assignments. If the computer program indicates a scheduling conflict, the scheduler manually moves students between different sections until all scheduling conflicts are resolved. After all sections have been allocated to time slots, the scheduler assigns teachers to the sections.[11] The section and teacher assignment then is published. After publication, the scheduler receives information on late-registering students and allocates them to the empty slots. The schedulers do not know the students nor do they observe their previous grades.

Only 20-25 students (less than one percent) officially switch sections per period. Switching sections is possible only through a student advisor and is allowed only for medical reasons or due to conflict with sports practice for students who are on a list of top athletes.[12] Students sometimes switch their section unofficially when they have extra appointments. This type of switching is usually limited to one session, and students rarely switch sections permanently.[13]

There are a few exceptions to this general procedure. First, when the number of late registering student exceeds the number of empty spots, the scheduler creates a new section that mainly consists of late registering students. Second, we excluded eight late registration sections

recorded as available for evening sessions. Students can opt out of this default position by indicating this in an online form. Evening sessions are scheduled from 6 p.m. to 8 p.m., and approximately three percent of all sessions in our sample are scheduled for this time slot.

[11] Approximately ten percent of teachers indicate time slots when they are not available for teaching. They do so before they are scheduled, and the signature of the department chair is required.

[12] We do not have a record for these students and therefore cannot exclude them. However, section switching in these rare cases is mostly due to conflicts with medical and sports schedules and therefore unrelated to section peers.

[13] It is difficult to obtain reliable numbers on unofficial switching. From our own experience and consultation with teaching staff, we estimate that session switching happens in less than 1 percent of the sessions and permanent unofficial class switching happens for less than 1 percent of the students.

from the analysis.[14] Third, for some Bachelor's courses, there are special sections consisting mainly of repeating students. Whether a repeater section is created depends on the preference of the course coordinator and the number of repeat students. We excluded 34 repeater sections from the analysis. Fourth, in some Bachelor's courses students who are part of the Maastricht Research Based Learning (MARBLE) program are assigned to separate sections where they often are assigned to more experienced teacher. Students of this program are typically the highest performing students of their cohort. We excluded 15 sections that consist of MARBLE students from the analysis.[15] Fifth, in six courses, the course coordinator or other education staff influenced the section composition.[16] We excluded these courses from our analysis. Sixth, some Master's tracks have part-time students. Part-time students are scheduled mostly in evening classes, and there are special classes with only part-time students. We excluded 95 part-time students from the analysis. Seventh, we excluded the first-year, first-period courses of the two largest Bachelor's programs (International Business and Economics) because, in these courses, only particular students, such as repeating student, have previous grades. Eighth, we excluded sections for which less than five students had a past GPA. For these courses, peer GPA does not reliably capture the peer quality of the students in the section. Ninth, we excluded sections with more than 16 students (two percent) because the official class size limit according to scheduling guidelines is 15 and in special cases 16. Sections with more than 16 students are a result of room availability constraints or special requests from course coordinators. After removing these

[14] Students who register late, for example, generally have a lower GPA and might be particularly busy or stressed during the period in which they registered late, which may also affect their performance. This dynamic might create a spurious relationship between GPA and grades.

[15] We identified pure late registration classes, repeater classes and MARBLE classes from the data. The scheduler confirmed the classes that we identified as repeater classes. The algorithm by which we identified late registration classes and MARBLE classes is available upon request.

[16] The schedulers informed us about these courses.

exceptions, neither students nor teachers, and not even course coordinators, influence the composition of the sections in our estimation sample.

3 Data

We obtained data for all students taking courses at the SBE during the academic years 2009/2010, 2010/2011 and 2011/2012. Scheduling data were provided by the Scheduling Department of the SBE. The scheduling data include information on section assignment, the allocated teaching staff and the day and time the sessions took place, as well as a list of late registrations for our sample period. In total, we have 7,460 students, 430 courses, 3,890 sections and 41,608 grades in our estimation sample. Panel A of Table 1 provides an overview of courses, sections and students in the different years.[17]

The data on student grades and student background, such as gender, age and nationality, were provided by the Examinations Office of the SBE. The Dutch grading scale ranges from 1 to 10, with 5.5 being the lowest passing grade. Data on students' course evaluations at the individual level were provided by the Department of Educational Research and Development of the SBE. The course evaluation data are described in greater detail in Section 9.

[17] We refer to each course-year combination as a separate course, which means that we count a course with the same course code that takes place in three years as three separate courses.

Table 1: Descriptive Statistics

Panel A

Academic year	Number of courses	Number of unique students	Number of sections	Average number of students per section	Number of grades
2009 / 10	114	3,688	1,146	13.18	12,020
2010 / 11	155	3,919	1,437	13.08	14,681
2011 / 12	161	4,064	1,307	14.16	14,907
All years	430	7,460	3,890	13.50	41,608

Panel B

	Mean	S.D.	Min	25p	Median	75p	Max	Obs.
Student level information								
Course dropout	0.083	0.276	0	0	0	0	1	45,373
Grade first attempt	6.572	1.877	1	6	7	8	10	41,608
Final grade	6.793	1.665	1	6	7	8	10	41,608
GPA	6.897	1.120	1	6.25	7	7.64	10	41,608
Section level information								
Number of registered students per section	13.49	1.323	5	13	14	14	16	45,373
Number of students that dropped class	2.326	2.000	0	1	2	3	14	45,373
Peer GPA (based on final grades)	6.767	0.468	4.93	6.45	6.79	7.10	8.50	41,608
Peer GPA (based on first sit grades)	6.541	0.508	3.78	6.21	6.55	6.89	8.48	41,575
Within section SD of peer GPA	1.114	0.363	0.10	0.86	1.09	1.35	2.80	41,608
Student Background information								
Age	20.78	2.154	16.19	19.22	20.48	22.03	41.25	38,650
Female	0.378	0.485	0	0	0	1	1	38,650
Dutch	0.301	0.459	0	0	0	1	1	41,608
German	0.507	0.500	0	0	1	1	1	41,608
Bachelor’s student	0.782	0.413	0	1	1	1	1	41,608
BA International Business	0.403	0.491	0	0	0	1	1	41,608
BA Economics	0.273	0.445	0	0	0	1	1	41,608
Exchange student	0.063	0.243	0	0	0	0	1	41,608

Note: This table shows the descriptive statistics of the estimation sample.

Figure 1: Distribution of Grades after the First Examination

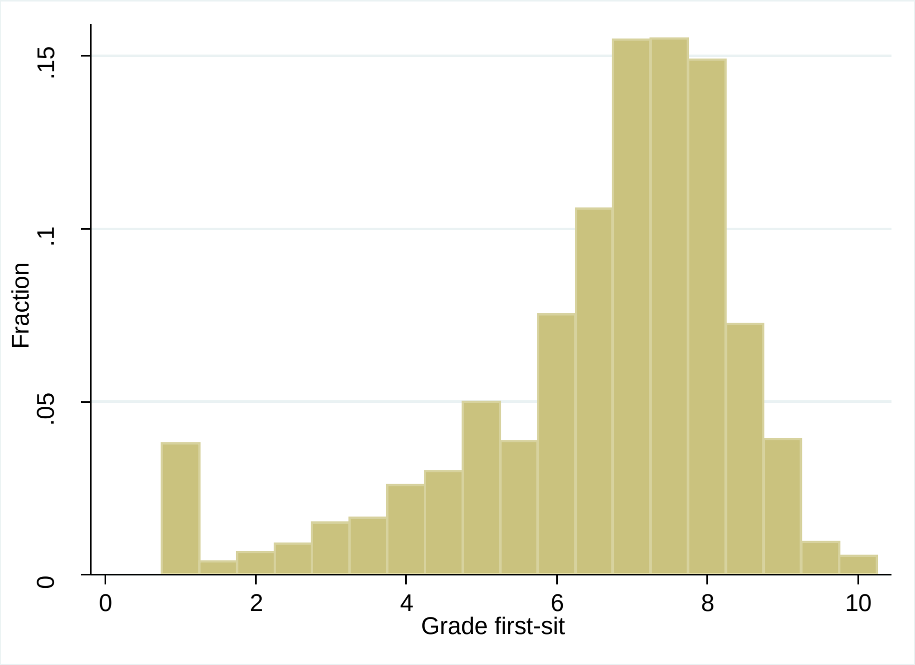


Figure 1 shows the distribution of final grades in our estimation sample. The final course grade is often calculated as the weighted average of multiple graded components such as the final exam grade, participation grade, presentation grade or midterm paper grade. The graded components and their respective weights differ by course, with most courses giving most of the weight to the final exam grade. For some courses, part of the final grade consists of group-graded components such as a group paper or a group presentation, for which all members of the group receive the same grade. The influence of these group grades on the final course grade might be one of the channels through which peers affect grades. Fortunately, our data contains a subset of courses (first year Bachelor's courses) for which the final exam was the only graded component of the final grade. We therefore will be able to rule out that our basic results are driven only by group-graded components. If the final course grade of a student after taking the final exam is lower than 5.5, the student fails the course and has the option of taking a second attempt at the exam. We observe final grades after the first and second attempt separately. For our analysis, we use only the final grade after the first exam attempt as an outcome measure because first- and second-attempt grades are not comparable.[18] For the construction of the student GPA, we use the final grades after the last attempt.[19]

Panel B of Table 1 shows some descriptive statistics for our estimation sample. Our sample contains 45,373 student course registrations. Out of these, 3,765 (8 percent) dropped out of the course during the course period. We therefore observe 41,608 course grades after the first attempt. The average course grade after the first attempt is 6.54. Approximately one fifth of the graded students obtain a course grade lower than 5.5 after the first attempt and therefore fail the course. The average final course grade (including grades from second and third attempts) is 6.80, and the average GPA is 6.90. Figure 2 shows the distribution of the GPAs based on final grades.

The peer GPA is the section-average GPA excluding the grades of the student of interest.[20] Figure 3 shows the distribution of peer quality, measured as the average past GPA of all other students in the section.

[18] The second-attempt exam usually takes place two months after the first exam.

[19] We decided to use the GPA calculated from final grades because this is closer to the popular understanding of GPA.

[20] For a more detailed explanation, see Section 4, where we describe our empirical strategy.

Figure 2: Distribution of own GPA

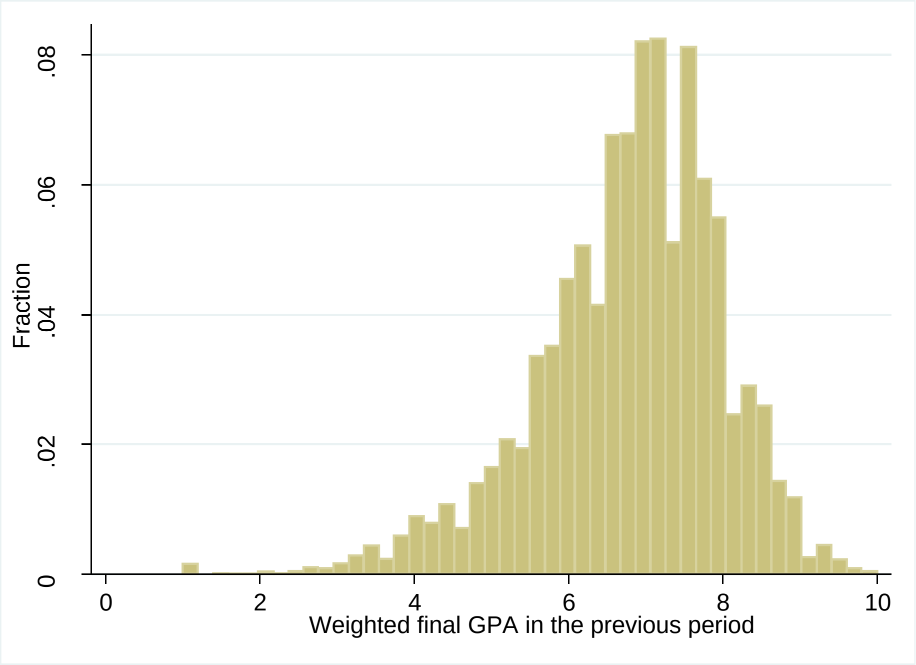
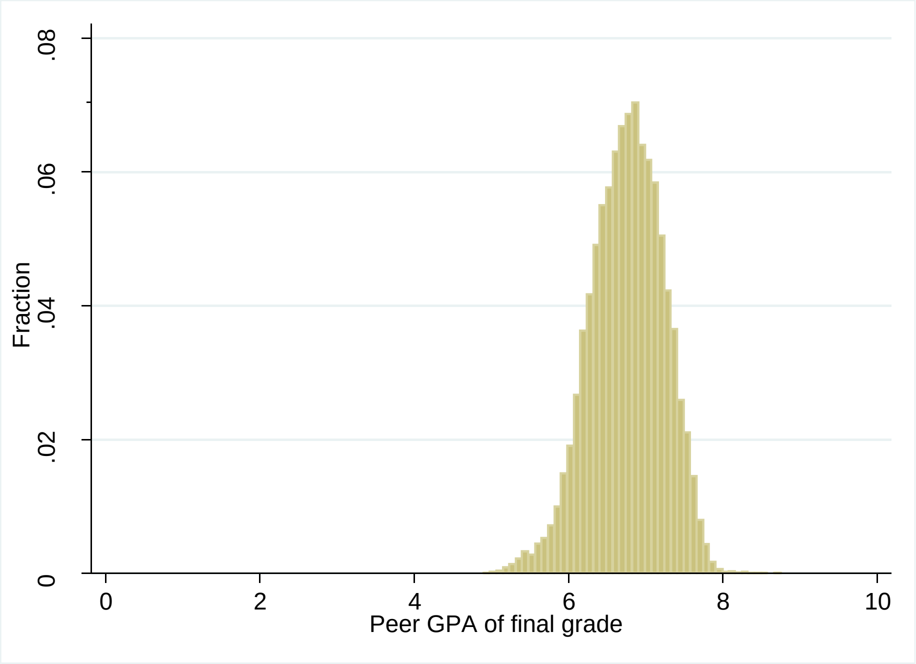


Figure 3: Distribution Peer GPA



4 Test for Random Assignment of Students to Sections

The scheduling procedure we describe in Section 2.2 shows that section assignment is random. Nevertheless, we test whether section assignment has the properties that one would expect under random assignment. In the spirit of standard randomization checks in experiments, we test whether section dummies jointly predict student pre-treatment characteristics when controlling for scheduling and balancing indicators. The pre-treatment characteristics we consider are GPA, age, gender, and student ID rank.[21] More specifically, for each course in our sample, we run a regression of pre-treatment characteristics on section dummies and scheduling and balancing controls and F-test for joint significance of the section dummies. Thus, for each pre-treatment characteristic, we run approximately 430 regressions. Under conditional random assignment, the p-values of the F-tests of these regressions should be uniformly distributed with a mean of 0.5 (Murdoch, Tsai, & Adcock, 2008). Furthermore, if students are randomly assigned to sections within each course, the F-test should reject the null hypothesis of no relation between section assignment and students' pre-treatment characteristics at the 5 percent, 1 percent and 0.1 percent significance level in close to 5 percent, 1 percent and 0.1 percent of the cases, respectively.

The results of these randomization tests confirm that the section assignment is random (Section A2 in the Appendix provides a more detailed description on our randomization check). The average of the p-values of the F-tests is close to 0.5 (see Table A1 in the Appendix), and the p-values are roughly uniformly distributed (see Figure A2 in the Appendix). Table 2 shows in how many cases the F-test actually rejected the null hypothesis at the respective levels. Column

[21] For approximately 9 percent of our sample, mostly exchange students, we do not know the age, gender and nationality. In Maastricht University, ID numbers are increasing in tenure at the university. ID rank is the rank of the ID number. We use ID rank instead of actual ID because the SBE recently added a new digit to the ID numbers, which creates a discrete jump in the series.

(1) shows the total number of courses for each pre-treatment characteristics. Column (3) shows that the actual rejection rates at the 5 percent level are close to the expected rejection rates under random assignment. The F-tests for the regressions with the dependent variables GPA and age are rejected slightly more often than 5 percent, and the rejection rates for the dependent variable gender and ID rank are slightly less than 5 percent. Columns (5) and (7) show the actual rejection rates at the 1 percent and 0.1 percent levels. Additionally, these rejection rates as a whole are close to the expected rates under random assignment, with the exception of age, where the rejection rates is only slightly higher than we expected. All together, we present strong evidence that section assignment in our estimation sample is random, conditional on scheduling and balancing indicators.

Table 2: Randomization Check of Section Assignment

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Dependent variable	Total number of courses	Number significant	Percent significant	Number significant	Percent significant	Number significant	Percent significant
Joint F-test significant:		...at the 5 percent level		...at the 1 percent level		...at the 0.1 percent level	
GPA	430	25	5.81%	7	1.62%	1	0.23%
Age	425	26	6.11%	11	2.56%	4	0.94%
Gender	422	17	4.03%	3	0.71%	0	0.00%
ID rank	430	21	4.89%	8	1.86%	2	0.46%

Note: This table is based on separate OLS regressions with past GPA, age, gender and ID rank as dependent variables. The explanatory variables are a set of section dummies, dummies the for other parallel course taken at the same time, and dummies for day and time of the sessions, German, Dutch, exchange student status and late registration status. Column (1) shows the total number of separate regressions. Columns (2), (4) and (6) show in how many regressions the F-test rejected the null hypothesis at the 5 percent, 1 percent and 0.1 percent level, respectively. Columns (3), (5) and (7) show for what percentage of the regressions the F-test rejected the null hypothesis at the respective levels. Differences in number of courses are due to missing observations for some of the dependent variables.

5 Estimating Peer Effects

There are three empirical challenges for clean identification of peer effects: selection into peer groups, the reflection problem, and what we will call “Angrist-mechanics” (Angrist, 2013).

The first challenge, selection into peer groups, is a general concern for all peer effects studies that arises because the reason for having particular peers in schools, classrooms, living communities or neighborhoods is likely to be correlated with unobserved characteristics. If these unobserved characteristics are correlated with student outcomes, peer effects estimates will be biased. In our study, selection bias is not a concern because we utilize data from a unique environment where assignment to peer groups is random, as shown in Section 4.

The second empirical challenge, the reflection problem, consists of the fact that one cannot disentangle the effect of peers on students from the effect of students on peers if student and peer outcomes are determined simultaneously (Manski, 1993). We therefore will follow what has become the standard approach in the recent peer effects literature and estimate peer effects using pre-treatment measures of student and peer quality (e.g., Carrell et al., 2009; Duflo et al., 2011; Carrell et al., 2013). We use measures of student and peer quality that were determined before the actual exposure to a specific peer group to avoid the reflection problem.

The third empirical challenge when estimating peer effects is what we call the “Angrist-mechanics” and lies in the relationship between the coefficients of own and peer ability. Angrist (2013) shows that in a regression of outcomes on pre-treatment average group ability and pretreatment own ability, the average group ability coefficient picks up biases in the own ability measure. Consider the following model:

$$y_i = \beta_1 a_i + \beta_2 \bar{a}_z + \varepsilon_i, \qquad (1)$$

where y is the student grade, a_i is the pre-treatment measure of own ability and $\bar{a}_z$ is the average of the ability measure in group z. Angrist has shown analytically that when we estimate this model with an OLS regression, the peer effects coefficient, $\hat{\beta}_2$, is equal to the difference between the coefficient of ability in an IV regression of grade on ability using group dummies as instruments for ability, γ_{IV}, and the coefficient of an OLS regression of grade and own ability, γ_{OLS}, divided by 1 minus the R-squared associated with the first stage of the IV regression (see Equation (2) and Equation (17) in Angrist (2013) on page 10). As the R-squared of the first stage in the IV regression empirically is often close to zero, $\hat{\beta}_2$ is approximately equal to the difference between the IV and OLS estimator.

$$\hat{\beta}_2 = \frac{\gamma_{IV} - \gamma_{OLS}}{1 - R^2} \approx \gamma_{IV} - \gamma_{OLS} \qquad (2)$$

This means that not only peer effects but all factors that lead to a difference between γ_{IV} and γ_{OLS} will also affect $\hat{\beta}_2$. In the context of estimating peer effects under random assignment, it is in particular measurement error which might bias $\hat{\beta}_2$. If, for example, a_i is measured with error that is averaged out at the group level, γ_{IV} will exceed γ_{OLS}, even in the absence of peer effects.

Our setup, however, differs from the one in Equation (1) because we include additional controls and use as the measure of peer quality the leave-out mean, the mean of all students in a group except student i, instead of the group average. To test to what extent our results are driven by the “Angrist-mechanics,” we redo our analysis in Section 7 using randomly re-assigned “placebo peer groups.” These “placebo peer groups” consist of students who never met in the actual classroom and therefore lack true peer effects but are subject to the same mechanical bias. We show that for our setting the mechanical bias exists and is of a modest size for the linear-in-mean specification.

6 Empirical Strategy and Baseline Results

We use the following model to estimate the effect of peers on grades:

$$Y_{ist} = \alpha + \beta_1 \overline{GPA}_{s-i,t-1} + \beta_2\, GPA_{i,t-1} + \gamma' Z_{ist} + \varepsilon_{ist}. \tag{3}$$

The dependent variable Y_{ist} is the grade of student i, in a course-specific section s, at time t. α is a constant. $\overline{GPA}_{s-i,t-1}$ is the average past GPA of all the students in the section excluding student i. $GPA_{i,t-1}$ is the past GPA of student i. Z_{ist} is a vector of additional controls, and ε_{ist} is an error term with the usual properties. Note that $GPA_{i\,t-1}$ and $\overline{GPA}_{s-i,t-1}$ might measure own and peer ability with some error that might bias our results through the mechanisms described by Angrist (2013).[22] In all specifications, Z_{ist} consists of dummies for day of the week and time of the day of the sessions, German, Dutch, exchange student status, late registration status, and year-course-period fixed effects.[23] The year-course-period fixed effects control for mean differences in outcomes across courses and time. This takes into account different grade levels in different years and courses with differing degrees of difficulty. In other specifications, we also include other-course fixed effects – i.e., fixed effects for the other course taken at the same time – and teacher fixed effects.[24] Conceptually, including scheduling controls and other-course fixed effects should pick up all leftover non-random variation in section assignment that is due to conflicting

[22] Further, note that the precision of own and peer ability estimates increases with tenure when $GPA_{i\,t-1}$ and $\overline{GPA}_{s-i,t-1}$ are calculated with more past grades. This means that we would expect any bias from measurement error to decrease with students' tenure.

[23] For some sections, the time and day of the sessions were missing. We include separate dummies for these missing values.

[24] Other-course fixed effects are dummies for the other course taken in the same period. These are only defined for students who take up to two courses per period. In only 1.5% of the cases, students were scheduled for more than two courses, and these students drop out of our sample when we include other-course fixed effects. Teacher fixed effects are fixed effects of the first teacher assigned to a session.

schedules. Including stratification controls and teacher fixed effects should increase the precision but not affect the size of the estimates. To allow for correlations in the outcomes of students within each course, we cluster the standard errors at the course-year-period level. We standardized $GPA_{i\,t-1}$ and $\overline{GPA}_{s-i,t-1}$ over the estimation sample (0,1) to simplify the interpretation of the coefficients.

Before we consider the effect of peer ability on grades, we check whether peer ability is related to course dropouts. The dropout rate for courses in which students enrolled is only 8 percent at the SBE. OLS regressions, which we omit for brevity, show that neither average peer GPA nor the other peer ability measures we use when estimating heterogeneous effects significantly predict course dropouts. Table 3 shows the results of OLS regressions with the standardized grade as the dependent variable.

The table shows that being assigned to section peers with a higher GPA causes higher course grades. The coefficient of standardized peer GPA is small but statistically significant in all models. The inclusion of teacher fixed effects and other-course fixed effects hardly change the effect size or its standard errors. The reported estimates in the most complete specification in Column (4) shows that being assigned to peers with a one standard deviation higher GPA increases the student's grade by 1.15 percent of a standard deviation. This effect size is approximately 2 percent of the effect of own GPA. In terms of the Dutch grading scale, this estimate means that, for example, an increase of peer GPA from 6.5 to 7.0 is associated with a grade increase from 6.50 to 6.523, a small and economically insignificant effect.[25]

[25] The estimation results are qualitatively similar and statistically significant when we limit our sample to first-year Bachelor's courses whose only graded components was the final exam. We therefore can rule out that our basic results are only driven by group-graded components like group presentations or group papers.

Table 3: Baseline Estimates – Linear-in-means

	(1) Std. Grade	(2) Std. Grade	(3) Std. Grade	(4) Std. Grade
Standardized peer GPA	0.0100*	0.0104*	0.0112**	0.0115**
	(0.005)	(0.006)	(0.005)	(0.006)
Standardized GPA	0.5507***	0.5506***	0.5523***	0.5522***
	(0.016)	(0.016)	(0.016)	(0.016)
Observations	41,608	41,608	41,608	41,608
R-squared	0.432	0.440	0.447	0.455
Course FE	YES	YES	YES	YES
Teacher FE	NO	YES	NO	YES
Other course FE	NO	NO	YES	YES

Note: Robust standard errors clustered at the course-year-period level are in parentheses. The dependent variable is the standardized course grade. All specifications include dummies for day of the session, time of the session, German, Dutch, exchange student status and late registration status. Other-course fixed effects refer to the course that students are taking at the same time. *** p<0.01, ** p<0.05, * p<0.1.

7 Robustness: Estimating Placebo-Peer-Effects

To test whether the mechanism described by Angrist (2013) or any other mechanical bias is driving our results, we re-estimate Equation (3) in a world where true peer interaction is absent. To do so, we need to create placebo peer groups that lack real world social interaction in the classroom. The size of this placebo-peer-effects estimate will reflect only mechanical forces and therefore will be informative about the degree to which our main findings have a causal interpretation of social interaction. To obtain a distribution of placebo-peer-effects point estimates we estimate Equation (3) 1,000 times in a bootstrap-like simulation where we repeatedly reassign students randomly to sections in a way that every student meets a group of completely new peers. For the placebo reassignment, we keep the size and the number of sections identical to the original assignment. Within each course and for every student, we randomly draw with replacement a new group of peers from the pool of possible peers that they were not assigned to under the original section assignment.[26] This “perfect stranger peer reassignment” ensures that no student is matched to one of the true peers with whom he or she actually could have interacted in the classroom. To adjust our estimates for “Angrist mechanics,” we then simply subtract the mean of the coefficients from the placebo analysis from the coefficient obtained under original assignment.

Panel A in Table 4 shows the results of the linear-in-means peer effects estimation under original and placebo section assignment. Panel B shows descriptive statistics for the placebo-peer-effects coefficients that we obtained after 1,000 placebo-peer-effects estimations. The mean of the placebo-peer-effects coefficients is 0.001058, which is 9.2 percent of our original estimate. The point estimate adjusted for “Angrist mechanics” is 0.010396. The placebo-peer-effects estimates exceed our estimate, with actual peer assignment in only 12 out of 1,000 placebo regressions. Figure 4 shows the distribution of the placebo-peer-effects coefficients that we have obtained. The vertical line indicates the size of our estimate with the actual peer group assignment as shown in Column (1) of Panel A in Table 4.

[26] The placebo section peers are drawn in a way that one specific peer never appears more than once in the placebo section of one specific student.

Table 4: Baseline Estimates (Linear-in-means) Original vs. Placebo Section Assignment

Panel A: Original and Placebo Regressions

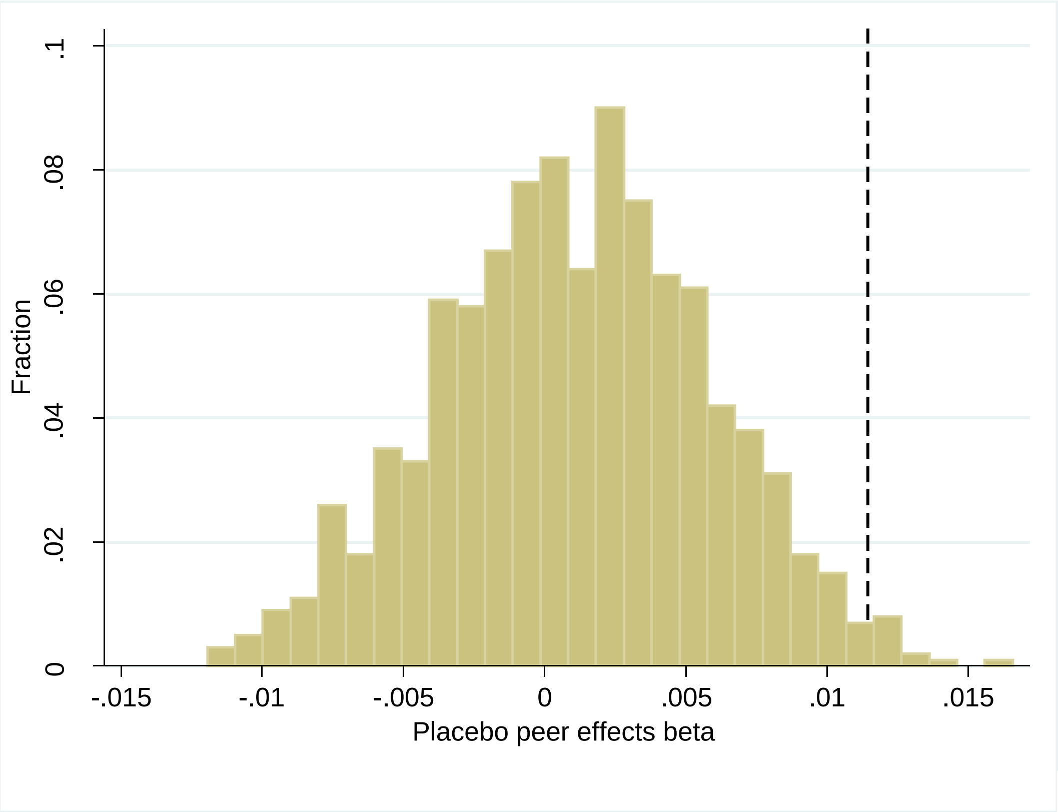
	Original peers	Placebo peers
	(1)	(2)
	Std. Grade	Std. Grade
Standardized peer GPA	0.011454	0.001058
	(0.005757)	[0.005031]
Standardized GPA	0.552160	0.514205
	(0.016389)	[0.015235]
Observations	41,608	41,608
R-squared	0.455	0.455

Panel B: Descriptive Statistics for Placebo Peer GPA Beta

Mean placebo beta	0.001058
Average standard error	0.005031
SD	0.004755
Min	-0.011945
Max	0.016574
Average placebo beta as percentage of true peer effect beta	9.20%
Actual peer beta – mean of placebo peer beta	0.010442
Number of placebo estimations	1,000
Number of placebo betas > original beta	12

Note: Robust standard errors clustered at the course-year-period level are in parentheses. The average of the standard errors of the placebo estimations clustered at the course-year-period level are reported in brackets. The dependent variable is the standardized course grade. All specifications include dummies for day of the session, time of the session, German, Dutch, exchange student status and late registration status, as well as fixed effects for courses, fixed effects for other courses taken at the same time and teacher fixed effects. The R-squared in column (2), Panel A represents the average R-squared of all 1,000 placebo estimations.

Figure 4: Distribution of Placebo Peer Effects



Note: This distribution is based on 1,000 draws. The dashed reference line represents the original point estimate from Table 4, Column (4), which we estimated with the actual sections.

Taken together, the results from our placebo peer-group simulations provide evidence for a moderate bias driven by the mechanics described in Angrist (2013). For all models that we estimate in the remainder of this paper, we report the results of the respective placebo estimations. We believe that the placebo simulation we developed to quantify the size of the mechanical bias could easily be applied to data from other settings (e.g.: Carrell et al., 2009; Duflo et al., 2011; Burke & Sass, 2013; Carrell et al., 2013).

8 Heterogeneous Effects

The specification in Table 3 is a linear-in-mean specification, which implicitly assumes that all students are linearly affected by the mean ability of their peers. However, previous studies have shown that peer effects are likely heterogeneous with respect to both student and peer ability (Burke & Sass, 2013; Carrell et al., 2013). We test for these two sources of heterogeneity simultaneously by estimating a two-way interaction model similar to those of Carrell et al. (2013) and Burke and Sass (2013). To do this, we classify students as high, middle and low GPA based on whether their GPA is in the top, middle or bottom third of the course GPA distribution, respectively. We then calculate for each section the fraction of peers with high and low GPA and include interactions of students' own type (high, middle and low) with the fraction of high- and low-GPA peers in the model we estimate.[27] Table 5 shows the coefficients of these six interactions. The first coefficient "High GPA * Fraction of High GPA peers," for example, can be interpreted as showing how high-GPA students are affected by increasing the fraction of high-GPA peers in the section while keeping the fraction of low-GPA peers constant. Put differently, the coefficient shows how high-GPA students are affected if middle GPA peers (the reference group) are replaced with high-GPA peers. The reference group is the fraction of middle-GPA students.

The estimation results for high- and middle-GPA students are in line with the linear-in-mean model: high- and middle-GPA students are positively affected by high-GPA peers and negatively affected by low-GPA peers. The results for low-GPA students, however, are substantially different. The point estimates suggest that low-GPA students are *negatively* affected by high-GPA peers. They are also negatively affected by peers from their own ability group –

[27] Interactions with fraction of middle-GPA peers are excluded because of collinearity.

low-GPA peers. The effect of increasing the fraction of high-GPA peers is significantly different for low-GPA students compared to high- and middle-GPA students.

Table 5: Heterogeneous Effects

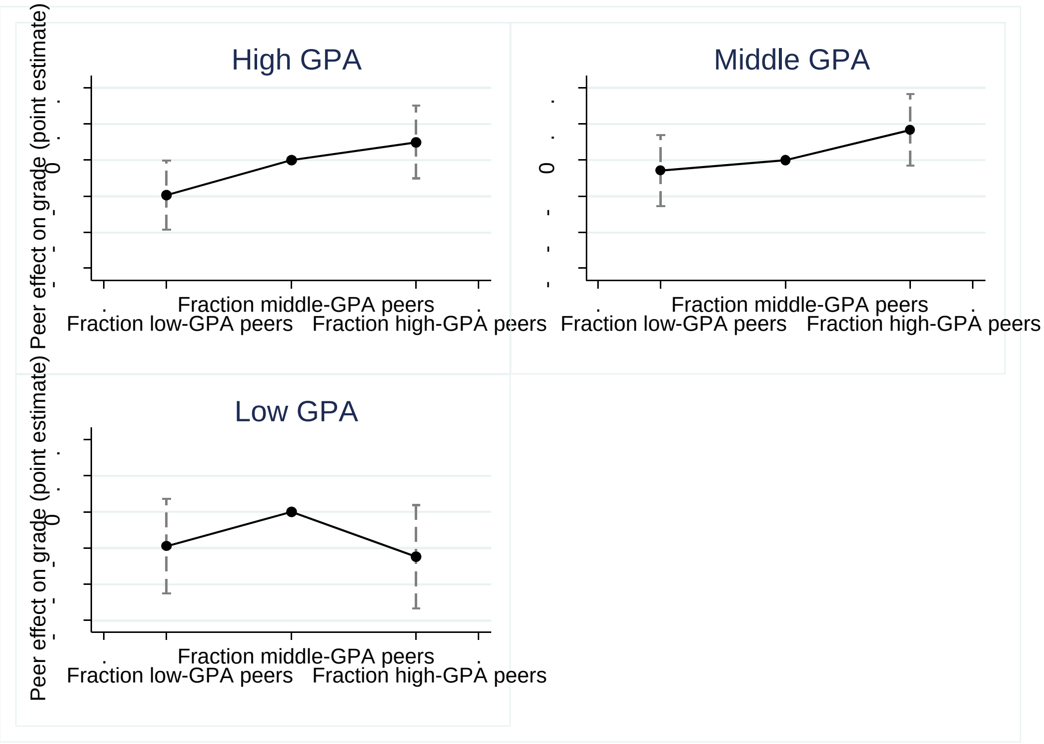
	(1) Std. Grade
High GPA * Fraction of high-GPA peers	0.0405
	(0.051)
High GPA * Fraction of low-GPA peers	-0.0999**
	(0.048)
Middle GPA * Fraction of high-GPA peers	0.0827
	(0.051)
Middle GPA * Fraction of low-GPA peers	-0.0293
	(0.050)
Low GPA * Fraction of high-GPA peers	-0.1224*
	(0.073)
Low GPA * Fraction of low-GPA peers	-0.0913
	(0.067)
Observations	41,608
R-squared	0.459
F fraction of high peers [middle vs. low]	4.83**
p-value	0.0285
F fraction of high peers [high vs. low]	3.51*
p-value	0.0616
F fraction of high peers [high vs. middle]	0.34
p-value	0.5573
F fraction of low peers [middle vs. low]	0.56
p-value	0.4562
F fraction of low peers [high vs. low]	0.01
p-value	0.9164
F fraction of low peers [high vs. middle]	1.16
p-value	0.2827

Note: Robust standard errors clustered at the course level are in parentheses. The dependent variable is the standardized course grade. Additional controls include Std. GPA and dummies for high GPA, low GPA, course, teacher, other course, day of the session, time of the session, German, Dutch, exchange student status and late registration status. *** p<0.01, ** p<0.05, * p<0.1.

To visualize this relationship, we plot the coefficients of the interactions in Table 5 in Figure 5. Although peer effects seem to increase linearly with peer ability for high- and middle-

GPA students, the effect first increases and then decreases for low-ability students. The results indicate that lower performing students do not benefit from the same peers that increase the performance of middle- and higher-performing students. Low-ability students appear to perform better when they are in one classroom with more middle-ability peers than high-ability peers. Overall, these effects are small in magnitude: for example, the coefficient "High GPA * Fraction of low-GPA peers" suggests that an increase of 20 percent in low-GPA peers, which is equivalent to replacing three out of 15 middle-GPA peers with low-GPA peers, decreases the grade of a high-GPA students by 1.9 percent of a standard deviation.

Figure 5: The Effect of Peer Fractions for Students with High, Middle and Low GPAs



Note: The data points in this figure are taken from Table 5 using the fraction of middle-GPA peers as a reference category.

Table 6 compares the estimates from Table 5 to the average of the placebo estimates that we obtained after 1,000 placebo peer group estimations. We apply the same procedure that we used for the linear-in-means placebo simulation described in Section 7. The table shows that our findings are not driven by a mechanical bias. The average of the placebo coefficients is small in magnitude, and the direction of the bias varies. Column (3) shows our estimates adjusted for "Angrist mechanics" – the difference between the original coefficient and the average of the placebo coefficients. Column (4) shows the size of the bias as the percentage of the coefficients with actual data.

Table 6: Heterogeneous Effects –Original vs. Placebo-peer-effects Estimations

	(1)	(2)	(3)	(4)
	Original peers	Placebo peers	Difference (1) - (2)	Bias direction and size as percentage of original beta
	Std. Grade	Std. Grade		
High GPA * Fraction of high-GPA peers	0.0405	0.0375	0.0030	+ 92.59 %
	(0.051)	[0.0436]		
High GPA * Fraction of low-GPA peers	-0.0999**	-0.0012	-0.0987	+ 1.20 %
	(0.048)	[0.0426]		
Middle GPA * Fraction of high-GPA peers	0.0827	-0.0095	0.0922	- 11.49 %
	(0.051)	[0.0497]		
Middle GPA * Fraction of low-GPA peers	-0.0293	0.0074	-0.0367	- 25.26 %
	(0.050)	[0.0494]		
Low GPA * Fraction of high-GPA peers	-0.1224*	0.0081	-0.1305	- 6.62 %
	(0.073)	[0.0623]		
Low GPA * Fraction of low-GPA peers	-0.0913	0.0065	-0.0978	- 7.12 %
	(0.067)	[0.0624]		
Observations	41,608	41,608		
R-squared	0.4587	0.4586		

Note: Robust standard errors clustered at the course level are in parentheses. Average standard errors from placebo estimations are included in brackets. The dependent variable in (1) and (2) is the standardized course grade. Additional controls include Std. GPA and dummies for high GPA, low GPA, course, teacher, other course, day of the session, time of the session, German, Dutch, exchange student status and late registration status. The placebo estimates reported in Column (2) show the average coefficients obtained from 1,000 reassignment draws and estimations of placebo peer effects. *** p<0.01, ** p<0.05, * p<0.1.

9 Evidence on the Channels of Peer Effects: Students' Course Evaluation

Peer effects could work through a number of different channels like, for example, better group functioning or an increase in student or teacher effort. The channels of peer effects have important implications for the costs and benefits of peer group assignment. Although better group functioning implies greater efficiency in the educational production, higher teacher or student effort suggests that the estimated peer benefits come at some costs.

To explore some potential channels of the peer effects in our setting, we investigate how peer ability affects students' answers in the course evaluation questionnaire. Table 7 shows the full list of standard questions and their respective answering scales. In our estimation sample, 39 percent of the students start filling out the questionnaire. The last column in Table 7 shows that once students started the questionnaire, they answered almost all of the standard questions.[28]

Answering the course evaluation questionnaire is selective. We observe, for example, that students with higher GPAs are more likely to take part in the evaluation, and our results should be interpreted in light of this finding. We have nevertheless chosen to analyze students' answers for two reasons. First, the survey response is not significantly related to peer quality as measured by mean peer GPA or to the other peer variables used in Section 8 (see Table A2 in the Appendix). Second, the student evaluation data gives us a unique insight into potential mechanisms in a way that is not available to many other studies. All results, however, should be interpreted with caution, and we interpret them as providing suggestive evidence.

[28] Conditional on having started the questionnaire, the response rate for each of our core questions ranges from 87.2 to 99.4 percent. The relatively low response rate for question 13, which evaluates the lecture, is likely driven by students who did not follow the lectures because their attendance is usually voluntary.

Table 7: Evaluation Questions, Answering Scales and Response Rates

Nr	Question domain	Question wording	Answer scale	Response rate conditional on survey participation
1	Teacher evaluation	Evaluate the overall functioning of your tutor in this course with a grade	1 - 10	98.5%
2	Teacher evaluation	The tutor sufficiently mastered the course content.	1 - 5	99.5%
3	Teacher evaluation	The tutor stimulated the transfer of what I learned in this course to other contexts.	1 - 5	99.3%
4	Teacher evaluation	The tutor encouraged all students to participate in the (tutorial) group discussions.	1 - 5	98.9%
5	Teacher evaluation	The tutor was enthusiastic in guiding our group.	1 - 5	99.4%
6	Teacher evaluation	The tutor initiated evaluation of the group functioning.	1 - 5	97.2%
7	Group interaction	My tutorial group has functioned well.	1 - 5	99.1%
8	Group interaction	Working in tutorial groups with my fellow-students helped me to better understand the subject matters of this course.	1 - 5	99.4%
9	Learning materials	The learning materials stimulated me to start and keep on studying.	1 - 5	98.5%
10	Learning materials	The learning materials stimulated discussion with my fellow students.	1 - 5	98.7%
11	Learning materials	The learning materials were related to real life situations.	1 - 5	98.7%
12	Learning materials	The textbook, the reader and/or electronic resources helped me studying the subject matters of this course.	1 - 5	95.7%
13	Learning materials	The lectures contributed to a better understanding of the subject matter of this course.	1 - 5	87.2%
14	Learning materials	In this course ELEUM has helped me in my learning.	1 - 5	91.8%
15	General evaluation	Please give an overall grade for the quality of this course	1 - 10	98.6%
16	General evaluation	The course fits well in the educational program.	1 - 5	97.2%
17	General evaluation	The course objectives made me clear what and how I had to study.	1 - 5	98.7%
18	General evaluation	The time scheduled for this course was not sufficient to reach the block objectives.	1 - 5	98.7%
19	Self-study hours	How many hours per week on the average (excluding contact hours) did you spend on self-study (presentations, cases, assignments, studying literature, etc.)?	0 - 80	92.6%

Note: At Maastricht University, the teaching staff member present in the classroom is referred to as "tutor." Sections are commonly called "tutorial groups." ELEUM is the name of a web-based learning platform that offers course materials and other course-related information.

For our analysis, we group all questions into four different evaluation domains: teacher evaluation, group interaction, learning materials and general evaluation. The question that assessed the number of self-study hours is included as a separate item. In our analysis, we include all students that have answered at least one item in each of these four domains and the question on self-study hours. We impute missing values using the sample average values. Most missing values are missing because the particular question was not asked in the course-specific questionnaire. Only a small share of missing values is due to non-response of the student.[29] We aggregate items in each of the four domains by first summing all the standardized item answers and then standardizing the sum of these values in each domain. Answers to the question about self-study hours are left in their natural units.

Table 8 shows how peers affect the evaluation of different course characteristics.[30] Panel A shows that the average peer GPA affects the evaluation of the group interaction positively and the learning material negatively. [31] The effect of peers on the evaluation of the group interaction suggests that this is an important channel through which peers affect student grades. It appears surprising that students assigned to peers with on average higher GPAs evaluate the course material less positively because the objective quality of the course material was the same for all sections and students independent of the assigned peers. Differences in material evaluation could reflect differences in usefulness or perceived quality of the material. We interpret this finding as suggesting that peer quality and learning material are substitutes: students who have good peers may use the learning material less and therefore evaluate it less positively. The teacher evaluation, general course evaluation and reported self-study hours are not affected by average peer GPA.

Panel B shows the results using the same specification for identifying peer effect heterogeneity as in Section 8 (i.e., ability and peer ability tertiles). This model allows us to investigate how the interaction between own and peer ability affects students' course evaluations. When comparing the different evaluation domains, we see that the peer variables are only jointly

[29] We obtain qualitatively similar results when we drop all students that answered only a subset of the 19 core questions.
[30] Table A3 in the Appendix shows the placebo estimations for the results shown in Table 8.
[31] Table A3 in the Appendix shows the linear-in-mean coefficients for each question separately.

significant in explaining the evaluation of the group interaction and not jointly significant in explaining the evaluation of the teacher, learning material, the course in general or the self-study hours.

Table 8: The Effect of Peer Composition on Student Evaluations

	(1)	(2)	(3)	(4)	(5)
Panel A	Std. Teacher evaluation	Std. Group interaction	Std. Learning materials	Std. General course evaluation	Hours worked
Standardized peer GPA	-0.0019	0.0564***	-0.0262**	-0.0151	0.0535
	(0.014)	(0.015)	(0.011)	(0.012)	(0.092)
Observations	14,982	14,982	14,982	14,982	14,982
R-squared	0.382	0.227	0.284	0.238	0.278

Panel B

	(1)	(2)	(3)	(4)	(5)
Highest tertile * fraction of peers in highest tertile	0.0847	0.5281***	-0.0505	0.0747	-0.4580
	(0.103)	(0.129)	(0.117)	(0.109)	(0.869)
Highest tertile * fraction of peers in lowest tertile	-0.0455	-0.1083	0.0163	0.1037	-0.8965
	(0.096)	(0.119)	(0.113)	(0.110)	(0.780)
Middle tertile * fraction of peers in highest tertile	-0.0260	0.2487*	-0.0176	0.0956	-0.2056
	(0.124)	(0.138)	(0.126)	(0.125)	(0.985)
Middle tertile * fraction of peers in lowest tertile	0.0349	0.0866	0.1876	0.2149*	1.8675**
	(0.114)	(0.117)	(0.118)	(0.115)	(0.917)
Lowest tertile * fraction of peers in highest tertile	0.0589	0.2750**	-0.1457	0.0592	-0.5887
	(0.118)	(0.124)	(0.126)	(0.136)	(0.935)
Lowest tertile * fraction of peers in lowest tertile	-0.0906	-0.0084	-0.0179	0.1285	-1.7458*
	(0.116)	(0.126)	(0.119)	(0.136)	(0.922)
F joined significance of peer variables	0.68	6.11	0.94	0.84	1.69
Prob > F =	0.6644	0.0000***	0.4678	0.5415	0.1210
Observations	14,982	14,982	14,982	14,982	14,982
R-squared	0.382	0.229	0.285	0.238	0.279

Note: All regressions include fixed effects for the course, fixed effects for the other course taken at the same time and teacher fixed effects. For a list of the exact question wording, see table 2 in section 3.2. Robust standard errors clustered at the course-year-period level are in parentheses. All specifications include dummies for day of the session, time of the session, German, Dutch, exchange student status and late registration status. *** p<0.01, ** p<0.05, * p<0.1.

Column (2) in Panel B shows that an increase in the fraction of high-GPA peers leads to a more positive evaluation of the group functioning. This effect is significantly larger and approximately twice as large for high-GPA students compared to medium-GPA students (p-value: 0.076).[32]

The positive effect of high-GPA peers on low-GPA students' evaluations of the group functioning is unexpected given the negative effect on students' grades that we have shown in Section 8. Paradoxically, lower-ability students appear to feel that group functioning and interaction has increased, but they are not able to benefit from the interaction with high-ability peers in terms of grades.[33]

10 Conclusion

We investigate peer effects in a large sample of university students where assignment to sections within a course is random. Consistent with previous research, we find effects of average peer quality on student grades that are small in size but statistically significant. These average effects hide some heterogeneity. Although the high- and middle-ability students benefit from better peers, the low-ability students are negatively affected by their high-ability peers.

These non-linear effects are in line with the results of the intervention of Carrell et al. (2013), who find that low-ability students are harmed when put in a group with a large share of high-ability peers. The non-linear effects we find suggest that it would be possible to increase

[32] The effect of increasing the fraction of high-GPA students is not significantly different between high- and low-GPA students (p-value: 0.174). The effect is not statistically different for medium- and low-GPA students (p-value: 0.890). The effect of increasing the fraction of low-GPA students is not significant for high-, middle- or low-GPA students.

[33] Although high-ability peers may improve the group functioning, they also can increase the difficulty of topics discussed in the classroom by, for example, asking more challenging questions. Although all students may have a more pleasant classroom experience, the grades of low-ability students may suffer from having group discussions beyond their capabilities.

overall student performance by reorganizing peer groups. The optimal allocation of peer groups, however, also depends on the objective of the social planner. A social planner who cares more about the welfare of low-ability students, for example, would want to allocate them to more medium-ability peers and fewer high-ability peers. Such an intervention, however, would harm medium-ability students.

We show some evidence from students' course evaluations that suggests that peer effects are driven mainly by changes in student interaction and not by adjustments in teachers' behavior or students' effort. This finding has important implications for potential welfare gains from the reorganization of peer groups. In principle, better student outcomes can be a result of an increase in efficiency, an increase in students' or teachers' effort or both. When student or teacher effort is driving peer effects on grades, the welfare implications of a reorganization aimed at increasing student performance become less straightforward. If, however, as our results suggest, the increase in students' performance is driven by better student interaction, reorganization of peer groups can lead to higher efficiency and therefore promises welfare gains.

Finally, we want to point out that it is not clear whether the effects we estimated by exploiting relatively small natural variations in peer quality have a predictive value for interventions that lead to large changes in peer quality. The results from Carell et al. (2013) suggest that this is not the case. Currently, we do not know whether this result is due to the generally low predictive power of studies exploiting small peer variation or due to "Angrist mechanics." The placebo analysis we have introduced in this paper provides a tool for past and future peer effects papers to assess the size and the direction of the bias caused by "Angrist mechanics."

References

Angrist, J. (2013). *The Perils of Peer Effects*. NBER Working Paper. National Bureau of Economic Research. Cambridge, MA.

Brunello, G., De Paola, M., & Scoppa, V. (2010). Peer Effects in Higher Education: Does the Field of Study Matter? *Economic Inquiry, 48*(3), 621-634.

Burke, M. A., & Sass, T. R. (2013). Classroom Peer Effects and Student Achievement. *Journal of Labor Economics, 31*(1), 51-82.

Carrell, S. E., Fullerton, R. L., & West, J. E. (2009). Does Your Cohort Matter? Measuring Peer Effects in College Achievement. *Journal of Labor Economics, 27*(3), 439-464.

Carrell, S. E., Sacerdote, B. I., & West, J. E. (2013). From Natural Variation to Optimal Policy? The Importance of Endogenous Peer Group Formation. *Econometrica, 81*(3), 855-882.

De Giorgi, G., Pellizzari, M., & Woolston, W. G. (2012). Class Size and Class Heterogeneity. *Journal of the European Economic Association, 10*(4), 795-830.

Duflo, E., Dupas, P., & Kremera, M. (2011). Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya. *The American Economic Review, 101*(5), 1739-1774.

Feld, J., Salamanca, N., & Hamermesh, D. S. (2013). *Endophilia or Exophobia: Beyond Discrimination*. NBER Working Paper. National Bureau of Economic Research. Cambridge, MA.

Hoxby, C. (2000). *Peer Effects in the classroom: Learning from gender and race variation*. NBER Working Paper. National Bureau of Economic Research. Cambridge, MA.

Lavy, V., Paserman, M. D., & Schlosser, A. (2012). Inside the Black Box of Ability Peer Effects: Evidence from Variation in the Proportion of Low Achievers in the Classroom. *The Economic Journal, 122*(559), 208-237.

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

Manski, C. F. (1993). Identification of Endogenous Social Effects: The Reflection Problem. *The Review of Economic Studies, 60*(3), 531-542.

Murdoch, D. J., Tsai, Y.-L., & Adcock, J. (2008). P-values are Random Variables. *The American Statistician, 62*(3), 242-245.

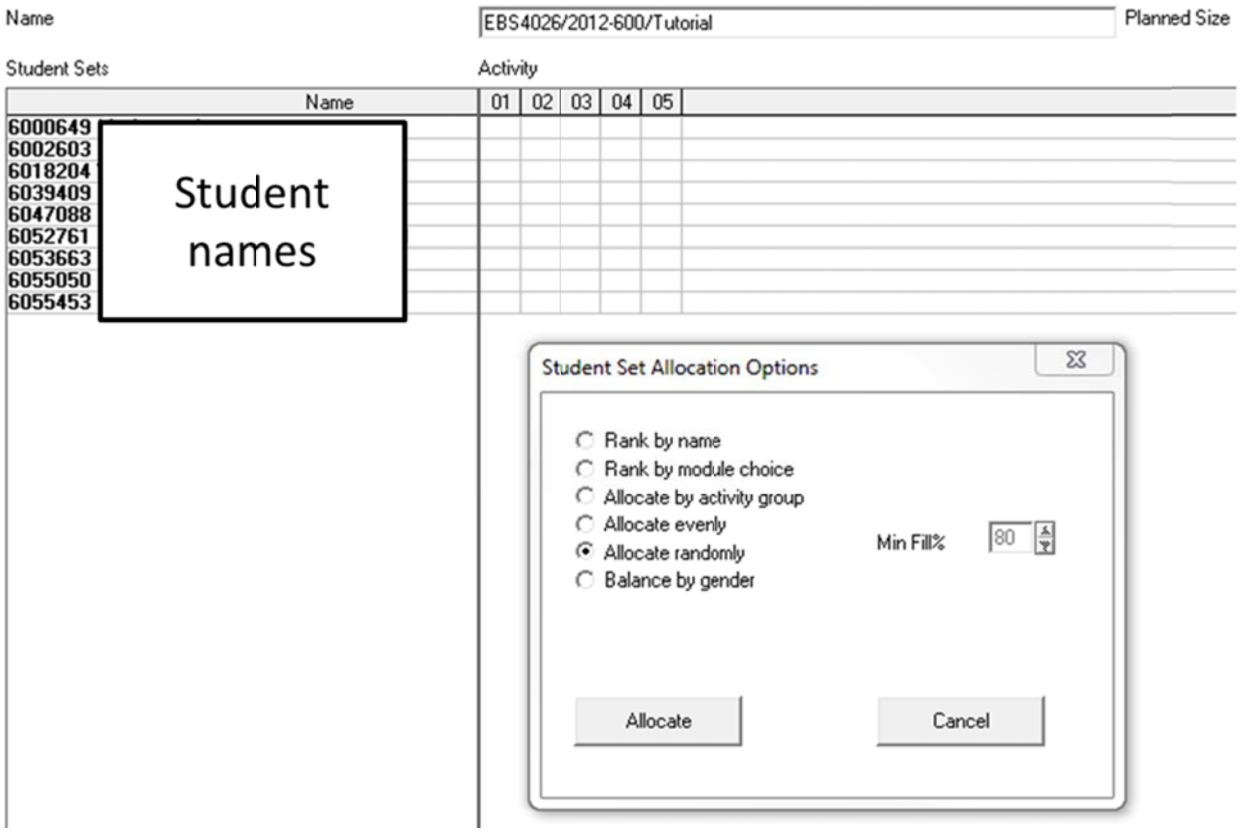
Sacerdote, B. (2001). Peer Effects with Random Assignment: Results for Dartmouth Roommates. *The Quarterly Journal of Economics, 116*(2), 681-704.

Zimmerman, D. J. (2003). Peer Effects in Academic Outcomes: Evidence from a Natural Experiment. *Review of Economics and Statistics, 85*(1), 9-23.

APPENDIX

A1 Additional Figure

Figure A1: Screenshot of the Scheduling Program Used by the SBE Scheduling Department



Name: EBS4026/2012-600/Tutorial

Planned Size

Student Sets

Activity

Name	01	02	03	04	05
6000649					
6002603					
6018204					
6039409					
6047088					
6052761					
6053663					
6055050					
6055453					

Student names

Student Set Allocation Options

○ Rank by name
○ Rank by module choice
○ Allocate by activity group
○ Allocate evenly
● Allocate randomly
○ Balance by gender

Min Fill% 80

Note: This screenshot shows the scheduling program Plus Enterprise Timetable©.

A2 Randomization Check

We use the following empirical specification for our tests. Take y_i as a *1xN$_i$* vector of the pre-treatment characteristics of students in course *i*. The pre-treatment characteristics we consider are GPA, age, gender, and student ID rank. $\boldsymbol{T} = (t_{1,} \dots t_n)$ is a matrix of section dummies. $\mathbf{Z}$ is a matrix that includes dummies for other course taken at the same time, day and time of the sessions, German, Dutch, exchange student status and late registration status and ε_i a vector of zero-mean independent error terms.

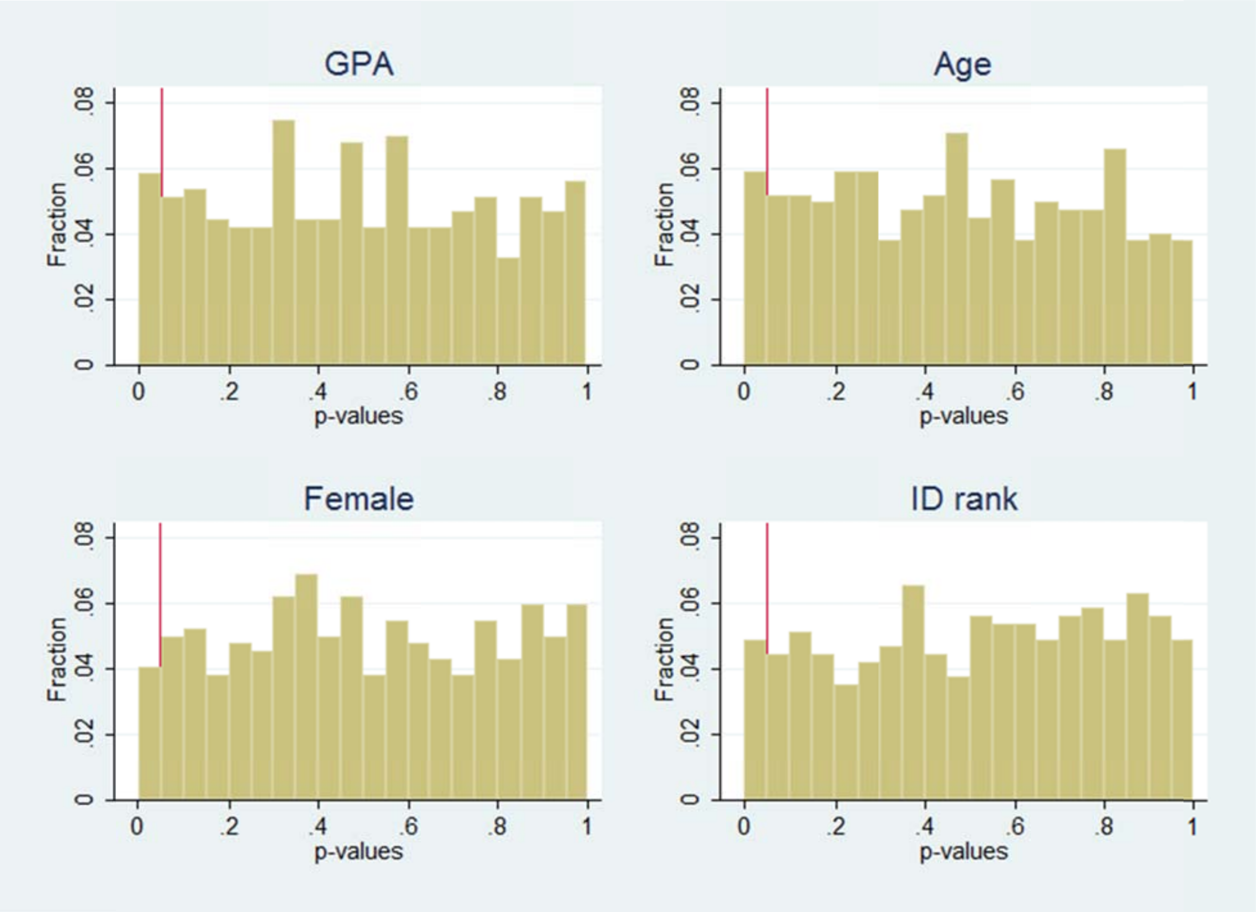
Our randomization tests consist of running, for each course, the following regression:

$$y_i = \alpha + \boldsymbol{T\beta} + \boldsymbol{Z\gamma} + \varepsilon_i \qquad \text{(A1)}$$

Under the null-hypothesis of (conditionally) random assignment to sections within each course, **β**= **0**, which means that the section assignment does not systematically relate to students' pre-treatment characteristics, holding constant scheduling and stratification indicators. Therefore, we expect the F-test to be significant at the 5 percent level in approximately 5 percent of the cases, at the 1 percent level in approximately 1 percent of the cases, and at the 0.1 percent level in approximately 0.1 percent of the cases. Table 2 in Section 4 shows that the actual rejection rates are close to the rejection rates expected under random assignment.

To investigate this issue more closely, we also consider the distribution of p-values. Under the null hypothesis of conditionally random assignment, we would expect the p-values of all the regressions to closely fit a U[0,1] uniform distribution with a mean of 0.5 (Murdoch et al., 2008). Figure A2 shows histograms of the p-values of all four specifications, all of which are roughly uniformly distributed. Column (2) of Table A1 shows the mean of the p-values over all regressions reported in Table 2. The mean of the p-values ranges from 0.48 to 0.52.

Figure A2: Distribution of F-test p-values of β from Equation (A1) as Reported in Table A1



Note: These are histograms with p-values from all the regressions reported in Table 2. The vertical line in each histogram shows the 0.05 significance level.

Table A1: Randomization Check: Mean p-values

	(1)	(2)
Dependent variable:	Total number of courses	Mean of p-value
GPA	430	0.49
Age	425	0.48
Gender	422	0.51
ID rank	430	0.52

Note: This table is based on the regressions reported in Table 2. Column (2) shows the means of the p-values.

Table A2: Determinants of Survey Response (OLS)

	(1) Response	(2) Response
Standardized peer GPA	-0.0015	
	(0.004)	
Highest tertile * fraction of peers in highest tertile		-0.0033
		(0.037)
Highest tertile * fraction of peers in lowest tertile		-0.0034
		(0.034)
Middle tertile * fraction of peers in highest tertile		-0.0212
		(0.034)
Middle tertile * fraction of peers in lowest tertile		0.0292
		(0.037)
Lowest tertile * fraction of peers in highest tertile		-0.0181
		(0.027)
Lowest tertile * fraction of peers in lowest tertile		-0.0248
		(0.029)
Standardized GPA	.0722***	.0607***
	(.0038)	(.0056)
F joined significance of peer variables		0.45
Prob > F =		0.8456
Observations	45,332	45,332
R-squared	0.104	0.104

Note: Robust standard errors clustered at the course-year-period level are in parentheses. Both regressions include fixed effects for the course, fixed effects for the other courses taken at the same time and teacher fixed effects. All specifications include dummies for day of the session, time of the session, German, Dutch, exchange student status and late registration status. *** p<0.01, ** p<0.05, * p<0.1.

Table A3: Placebo-peer-effects Estimations for Table 8 (Effects on Students' Evaluations)

	(1)	(2)	(3)	(4)	(5)
Panel A	Tutor evaluation	Group interaction	Learning materials	General course evaluation	Hours worked
Standardized peer GPA	0.0002	-0.0003	0.0013	0.0003	0.0073
	[0.0127]	[0.0143]	[0.0138]	[0.0142]	[0.1111]
Observations	14,982	14,982	14,982	14,982	14,982

Panel B					
	(6)	(7)	(8)	(9)	(10)
Highest tertile * fraction of peers in highest tertile	-0.0087	0.0023	0.0185	-0.0214	0.0008
	[0.1091]	[0.1236]	[0.1186]	[0.1234]	[0.9916]
Highest tertile * fraction of peers in lowest tertile	-0.0011	0.0017	0.0005	0.0143	0.0713
	[0.1078]	[0.1219]	[0.1169]	[0.1215]	[0.982]
Middle tertile * fraction of peers in highest tertile	0.0025	-0.0147	-0.0079	0.0072	-0.1546
	[0.1025]	[0.1145]	[0.111]	[0.1131]	[0.9469]
Middle tertile * fraction of peers in lowest tertile	-0.0153	-0.0164	-0.0209	-0.0406	0.1516
	[0.1005]	[0.1117]	[0.1094]	[0.1119]	[0.9353]
Lowest tertile * fraction of peers in highest tertile	-0.0051	-0.0039	-0.0129	0.0023	0.163
	[0.1021]	[0.1155]	[0.1116]	[0.1151]	[0.9225]
Lowest tertile * fraction of peers in lowest tertile	0.0071	0.0045	0.0116	0.0193	-0.1565
	[0.1008]	[0.1132]	[0.11]	[0.114]	[0.9183]
Observations	14,982	14,982	14,982	14,982	14,982

Note: Robust standard errors clustered at the course level are in parentheses. Average standard errors from placebo estimations are included in brackets. The dependent variable in (1) and (2) is the standardized course grade. Additional controls include Std. GPA and dummies for high GPA, low GPA, course, teacher, other course, day of the session, time of the session, German, Dutch, exchange student status and late registration status. The placebo estimates reported in Column (2) show the average coefficients obtained from 1,000 reassignment draws and estimations of placebo peer effects. *** p<0.01, ** p<0.05, * p<0.1.

Table A4: Separate Regressions for Each Course Evaluation Question

Nr.	Question domain	Dependent variable	Coefficient of std. peer GPA	SE std. peer GPA	R-squared
1	Teacher evaluation	Evaluate the overall functioning of your tutor in this course with a grade	0.0034	(0.007)	0.844
2	Teacher evaluation	The tutor sufficiently mastered the course content.	-0.0148**	(0.007)	0.749
3	Teacher evaluation	The tutor stimulated the transfer of what I learned in this course to other contexts.	0.0046	(0.007)	0.778
4	Teacher evaluation	The tutor encouraged all students to participate in the (tutorial) group discussions.	-0.0064	(0.009)	0.695
5	Teacher evaluation	The tutor was enthusiastic in guiding our group.	0.0092	(0.007)	0.795
6	Teacher evaluation	The tutor initiated evaluation of the group functioning.	0.0040	(0.009)	0.655
7	Group interaction	My tutorial group has functioned well.	0.0822***	(0.014)	0.400
8	Group interaction	Working in tutorial groups with my fellow-students helped me to better understand the subject matters of this course.	0.0198	(0.012)	0.348
9	Learning materials	The learning materials stimulated me to start and keep on studying.	-0.0267**	(0.011)	0.287
10	Learning materials	The learning materials stimulated discussion with my fellow students.	-0.0166	(0.012)	0.285
11	Learning materials	The learning materials were related to real life situations.	-0.0147	(0.012)	0.254
12	Learning materials	The textbook, the reader and/or electronic resources helped me studying the subject matters of this course.	-0.0197	(0.012)	0.277
13	Learning materials	The lectures contributed to a better understanding of the subject matter of this course.	-0.0093	(0.010)	0.357
14	Learning materials	In this course ELEUM has helped me in my learning.	-0.0287**	(0.012)	0.186
15	General evaluation	Please give an overall grade for the quality of this course	-0.0143	(0.011)	0.412
16	General evaluation	The course fits well in the educational program.	-0.0163	(0.011)	0.297
17	General evaluation	The course objectives made me clear what and how I had to study.	-0.0034	(0.013)	0.262
18	General evaluation	The time scheduled for this course was not sufficient to reach the block objectives.	0.0102	(0.013)	0.150
19	Self-study hours	How many hours per week on the average (excluding contact hours) did you spend on self-study (presentations, cases, assignments, studying literature, etc.)?	0.0067	(0.011)	0.279
		Number of observations in each regression	14,982		

Note: Robust standard errors clustered at the course-year-period level are in parentheses. All 19 regressions include fixed effects for the course, fixed effects for the other course taken at the same time and teacher fixed effects. All specifications include dummies for day of the session, time of the session, German, Dutch, exchange student status and late registration status. *** p<0.01, ** p<0.05, * p<0.1.