Ability peer effects in university: Evidence from a randomized experiment*

Adam S. Booij† Edwin Leuven‡ Hessel Oosterbeek§

Abstract

This paper estimates peer effects originating from the ability composition of tutorial groups for undergraduate students in economics. We manipulated the composition of groups to achieve a wide range of support, and assigned students – conditional on their prior ability – randomly to these groups. The data support a specification in which the impact of group composition on achievement is captured by the mean and standard deviation of peers' prior ability, their interaction, and interactions with students' own prior ability. When we assess the aggregate implications of these peer effects regressions for group assignment, we find that low and medium ability students gain on average 0.19 SD units of achievement from switching from ability mixing to three-way tracking. Their dropout rate is reduced by 12 percentage points (relative to a mean of 0.6). High-ability students are unaffected. Analysis of survey data indicates that in tracked groups, low-ability students have more positive interactions with other students, and are more involved. We find no evidence that teachers adjust their teaching to the composition of groups.

JEL-codes: I22; I28

Keywords: Peer effects; tracking; post-secondary education; field experiment

*We gratefully acknowledge valuable comments from three anonymous referees, Dennis Epple, Erik Plug, Geert Ridder, and from seminar participants in various places.

†University of Amsterdam. Also affiliated with Tinbergen Institute. Email: adam.booij@uva.nl

‡University of Oslo. Also affiliated with CEPR, CESifo, IZA and Statistics Norway. Email: edwin.leuven@econ.uio.no.

§University of Amsterdam. Also affiliated with CESifo and Tinbergen Institute. Email: h.oosterbeek@uva.nl.

1 Introduction

Can we improve student outcomes through ability grouping? The current paper reports the results from a randomized experiment that manipulated the ability composition of tutorial groups for first-year students in economics over a large support. The random assignment of students to groups ensures that we can estimate contextual peer effects even in the presence of endogenous social interactions. The resulting estimates are used to investigate the impact of different grouping scenarios for overall achievement.

A large body of work has documented contextual peer effects in education (see Sacerdote (2014) for a recent review). Identification of peer effects is challenging because reflection and selection typically lead to serious omitted variable bias (Manski, 1993). The main focus of recent studies has therefore been on recovering estimates of contextual peer spillovers based on variation in peer characteristics that is arguably random. There are two broad approaches. The first exploits naturally occurring variation in peer group composition (e.g. Hoxby, 2000; Carrell et al., 2009; Ammermueller and Pischke, 2009; De Giorgi et al., 2012; Feld and Zölitz, 2016) and a second, smaller, and more recent literature uses randomized experiments (Duflo et al., 2011; Carrell et al., 2013). While results are context dependent, the literature generally finds that peer effects are nonlinear and heterogeneous.[1]

The studies that are based on randomized experiments provide credible effect estimates of the particular ability peer configuration they are interested in, but tend to be silent about the effects of alternative peer groupings because discrete treatments limit the scope for extrapolation. More specifically, Carrell et al. (2013) obtain credible estimates of the effects of grouping students from the lowest one-third and from the highest one-third of the prior ability distribution together, relative to ability mixing, while Duflo et al. (2011) estimate the effects of two-way tracking relative to ability mixing. Neither paper is, however, able to identify the impact of the other paper's ability grouping in their specific setting. That is: Carrell et al. (2013) cannot estimate the effect of two-way tracking in their population of cadets of the US Air Force Academy. Nor can Duflo et al. (2011) estimate the effect of grouping students from the lowest and highest

[1] Studies that document nonlinear and/or heterogeneous peer effects include Hoxby (2000), Brodaty and Gurgand (2016), Lavy et al. (2012a), Lavy et al. (2012b), Burke and Sass (2013) and Black et al. (2013).

ability tertiles in their setting of primary school students in Kenya.

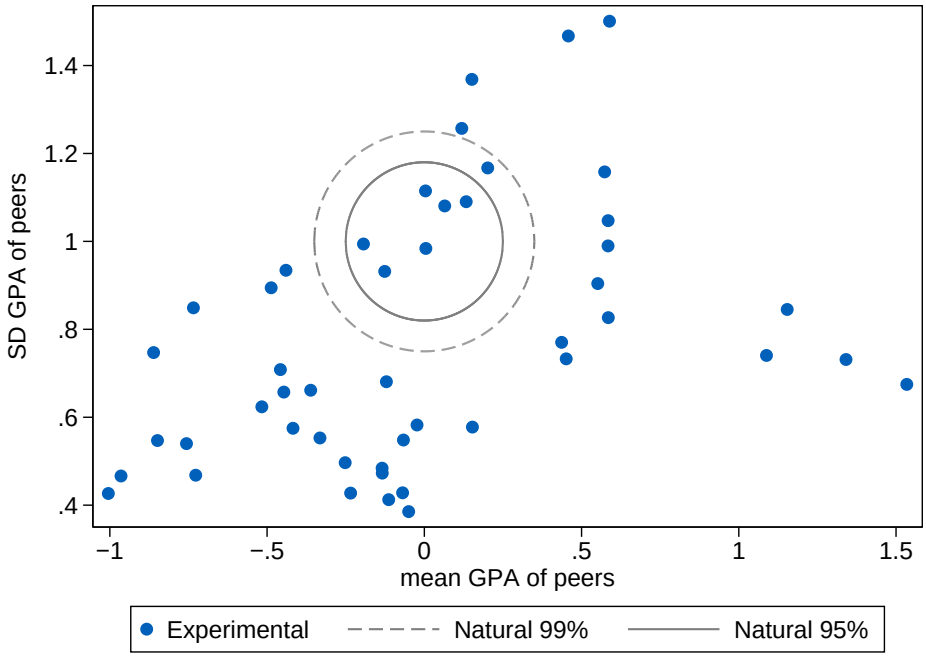
Studies that are based on naturally occurring variation typically exploit continuous variation in peer group composition. Since this allows for a flexible estimation of a peer response function, such estimates are more readily used to predict the effects of alternative peer groupings. Naturally occurring variation is however often limited, and such exercises are then likely to encounter support problems when translating their estimates into policy recommendations: when peer configurations are not covered by the data, extrapolation relies on functional form alone and runs the risk of being invalid.[2]

A compelling illustration of such support issues is provided by Carrell et al. (2013). Results based on naturally occurring variation in peer composition at the US Air Force Academy indicated that freshmen from the lowest one-third of the prior ability distribution would gain from being grouped together with freshmen from the highest one-third of the ability distribution (see also Carrell et al., 2009). The authors then conduct a randomized experiment to test this and find that low-ability students are actually harmed by the policy which was predicted to benefit them.

In order to learn how alternative peer configurations affect outcomes for the same population and context one needs to address the reflection and selection problem – ideally through randomization – while at the same time ensuring consistent estimation of peer response functions over a sufficiently large support. The current study is explicitly designed to address these issues. The context of our experiment is the first-year undergraduate program in economics and business at the University of Amsterdam. The circa 600 students that enter each year are assigned to tutorial groups of around 40 students. The composition of these groups is fixed throughout the entire first year, and more than 60% of all teaching hours take place in these groups. We performed the randomization in the academic years 2009/10, 2010/11, and 2011/12, when we were granted permission to randomly assign incoming students to tutorial groups.

The assignment procedure was designed to achieve large and exogenous variation in the ability peer composition across tutorial groups, where ability is defined by students' grade point average on the nationwide final exams of secondary education (GPA). These nationwide exams

[2]Moreover, Angrist (2014) points out that peer effect studies based on naturally occurring variation may suffer from weak instrument type bias.



Note: Each dot in the graph represents one tutorial group. The dashed (solid) circle represents the area where 99% (95%) of the tutorial groups would be located when the composition of the groups would not be manipulated and when students would be randomly assigned to groups.

Figure 1. Variation in mean and standard deviation of peers' ability

are administered prior to application to university, and GPA is therefore a predetermined measure of ability. Figure 1 shows that our procedure substantially increased the variation in peer group composition relative to the variation that would occur naturally.[3] With random assignment but without manipulation of group composition, mean standardized GPA would for 95% of the groups range from [-0.3, 0.3]. In contrast, with our manipulation the actual range is [-1, 1.6]. Similarly, the heterogeneity of the groups, as measured by the standard deviation (SD) of standardized GPA in a group, increased from [0.8, 1.2] to [0.3, 1.5].

The randomization and large support allow us to estimate flexible reduced form models of the relation between student outcomes and the ability composition of tutorial groups. We find evidence that peer effects are nonlinear and heterogenous. Student achievement increases with mean peer GPA and decreases with the SD of peer GPA. The positive impact of mean peer GPA is larger in more heterogeneous groups. The negative impact of the SD of peer GPA is smaller when mean peer GPA is higher. These patterns are less pronounced for students with higher own GPA.

[3]Subsection 2.2 provides details about the assignment procedure.

Although these results are informative about peer effects at the margin, the implications for outcomes of assignment of students to groups at the student population level is less straightforward. This is because overall achievement depends on the trade-offs implied by the nonlinear and heterogenous nature of the peer effects, and on the availability of students of different prior ability levels which puts restrictions on the actual ability groupings that can be made in the population. We therefore use our estimates to contrast the predicted average student outcomes for different peer group configurations. The results indicate that low-GPA and middle-GPA students would gain on average 19% of a standard deviation of realized credits from moving from mixing to three-way tracking. Their dropout rates go down by around 12 percentage points (relative to a mean of 0.60). High-GPA students are unaffected by the GPA composition of their tutorial group. When we use our results to predict student achievement under Carrell et al's configuration where low-GPA and high-GPA students are grouped together, we find that the achievement of low-GPA students goes down by (an insignificant) 2% of a standard deviation and the achievement of middle-GPA students is boosted by 16% of a standard deviation. High-GPA students are again unaffected. These patterns are qualitatively similar to the results that Carrell et al. obtain in their experiment.

A possible explanation for the effect of the ability composition of tutorial groups on student achievement lies in the higher dropout rate among low-ability students. Because of this, the average size of tutorial groups during the year is also affected by the ability composition, and the results may therefore be driven by an effect of average group size on achievement. We assess this explanation by using variation in group size caused by students who were assigned to groups but never showed up, and find that that group size is not an important factor for student performance.

Finally, we also collected survey data to examine other mechanisms underlying the achievement effects. Low-GPA students in tracked groups have more positive interaction with other students and are more involved with their studies than low-GPA students in mixed groups. The survey responses give no support for teachers as a mediating factor; their teaching is not adjusted to the ability composition of tutorial groups.

This paper proceeds as follows. The next section describes the context, the experimental

design, and the data. Section 3 briefly introduces the empirical specifications that we estimate. Section 4 presents and discusses the empirical findings. Section 5 assesses different potential mechanisms explaining our findings. Section 6 summarizes and concludes.

2 Context, design and data

2.1 Context

The experiment was conducted in the academic years starting in September of 2009, 2010, and 2011, among first-year students in the three-year bachelor program in economics and business at the University of Amsterdam.[4] In the first year all students in economics and business follow exactly the same program. Students can thus not substitute easy for difficult courses.

Teaching during the first year takes place in two forms: i) central lectures where all first-year students are grouped together, and ii) tutorial meetings where students are assigned to classes of about 40 students. In these tutorial groups, students typically receive in-depth explanations of the material, ask questions, and practice and discuss exercises and assignments. The teachers of tutorial groups are faculty members and PhD students. A teacher typically teaches three or four tutorial groups in the same subject. Students are assigned to a specific tutorial group before the start of the year and are supposed to stay in the same group for the entire first year. There were 14 tutorial groups in 2009, 17 in 2010, and again 17 in 2011.[5]

Table A1 in the appendix lists the first-year courses together with their scheduling in the year and their study load in terms of total teaching hours, tutorial group hours and credit points. As mentioned above, this shows that just over 60% of total teaching hours take place in tutorial meetings. We do not claim that the tutorial group is the only peer group, or the most relevant one. Students can – and will – also interact with students from other tutorial groups, or even from other studies. Or, in the opposite direction, students can form informal subgroups within their tutorial group of students with whom they interact more frequently.[6] The level of tutorial

[4]Students meeting the admission requirements are automatically accepted for the study without further selection. The main requirement is that students graduated from the academic track in Dutch secondary education.

[5]From the 2009 data we drop two groups with late registrations, and from the 2010 and 2011 data we drop a group of students that pursue the fiscal economics track in the second year. The students in these groups were not randomly assigned and are therefore not part of the experiment.

[6]Defining the relevant peer group is not obvious. Some studies explore this issue by defining peer groups at

groups is, however, the level at which the university assigns a cohort of incoming students to smaller units, and is therefore the level for which information about the pattern of peer effects can be used to raise achievement.

Whether students pass a course and their grade depends only on their performance on the mid-term and end-term exams of the course. These exams are identical for all students and take place in large rooms fitting all first-year students. The answer sheets of all students are collected in a large pile and not split by tutorial groups. Grading is uniform with many exams consisting of multiple choice questions. The course coordinators are responsible for the grading of exams. It is thus not the case that the grades of students in a tutorial group with many low-GPA peers are inflated to secure a minimum pass rate or average grade within the tutorial group.

Teachers of tutorial groups are not directly rewarded for the performance of the students in their group(s). At the end of the course, teachers are evaluated by their students through a standardized evaluation form. There is no evidence that teachers with more favorable evaluations also realize higher passing rates. The impression is that the evaluations merely reward popular teachers. This is probably best realized by tailoring the instruction to the median student in the group. For tenure and promotion decisions of personnel, student evaluations are taken into account, but the key determinant is research output.

2.2 Design

Assignment. To acquire information about the nature of ability peer effects in tutorial groups, we manipulated the ability composition of first-year tutorial groups, and randomly assigned students to these groups (conditional in their ability). Our measure of a students' pre-treatment ability is their GPA on the final exams in secondary school. This GPA is the average grade over seven (or eight) subjects for which the students write nationwide central exams and which are graded on a scale from 1 to 10, where 6 means a pass. In accordance with the standard procedures of the department of economics and business of the University of Amsterdam, we were required to assign students to tutorial groups before the start of the academic year. At that stage, the university (and therefore we) did not have access to students' exact GPA, but only

different levels. Sacerdote (2001), for example, examines peer effects of roommates as well as of dorm mates. See also Glaeser et al. (2003).

Table 1. Prior GPA distribution of incoming students by cohort and type of math (regular or advanced)

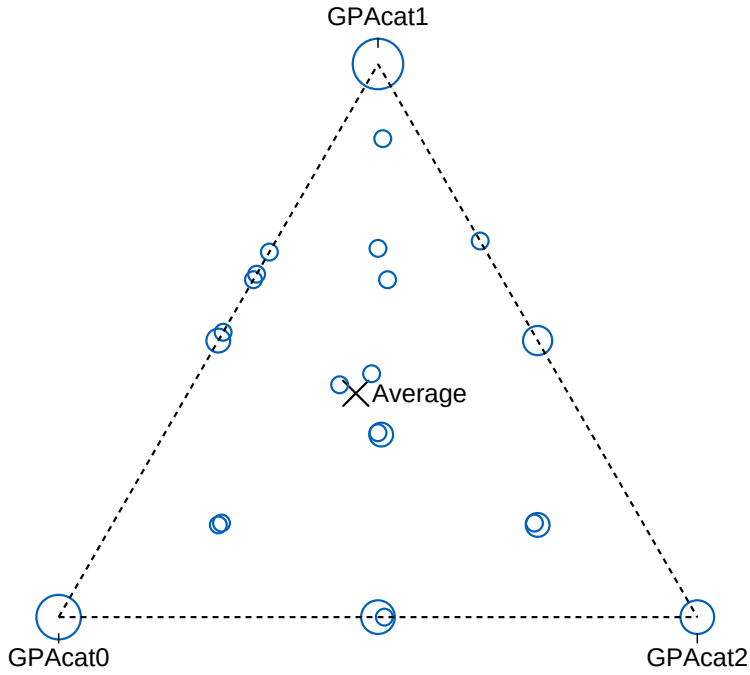
		Cohort					
		2009		2010		2011	
GPA interval	*GPAcat*	Reg.	Adv.	Reg.	Adv.	Reg.	Adv.
$GPA < 6\frac{1}{2}$	0	0.36	0.28	0.30	0.33	0.35	0.34
$6\frac{1}{2} \leq GPA < 7$	1	0.36	0.39	0.45	0.37	0.43	0.39
$GPA \geq 7$	2	0.27	0.33	0.25	0.30	0.22	0.26
Total		1.00	1.00	1.00	1.00	1.00	1.00

Note: The table reports the composition of incoming students in terms of their GPA on the final exams in secondary school, by cohort (2009, 2010 and 2011) and type of mathematics (regular and advanced). At the moment of assignment to tutorial groups, GPA is only known in three categories: less than 6.5, between 6.5 and 7, and 7 or higher.

to a coarse measure of it. This coarse measure reports whether a student's GPA is below 6.5, between 6.5 and 7, or 7 or higher. Table 1 shows the distribution of students across these three GPA categories, by cohort and by the type of math – regular or advanced – the student attended in secondary school.

We refer to the different categories, which contain roughly 32% (GPA below 6.5), 40% (GPA at least 6.5 but below 7), and 28% (GPA at least 7) of the students, as *GPAcat* 0, 1, and 2. Using this division we manipulated the shares of each GPA-category in each tutorial group, by setting different assignment probabilities for each tutorial group conditional on students' GPA-category. The triangle in Figure 2 shows the support spanned by the three GPA category shares. The fraction of each GPA category in a group positions it on the triangle. The vertices of the triangles correspond to groups that consist of only one category, such as *GPAcat*1 in case of the top one. Coordinates on the edges are group assignments that combine the GPA categories of the corresponding vertices. A point on the left edge for example corresponds to a tutorial group assignment that combines only students from *GPAcat*0 and *GPAcat*1. Interior points combine students from all three ability categories. The dots in Figure 2 shows the tutorial group assignments resulting from our randomization, and illustrates the large variation in the ability composition of the tutorial groups that cover the full prior ability support.

For the 2010 and 2011 cohorts, the conditional random assignment was conducted just before the start of the academic year in September. For these cohorts we could use the prior distribution of students over GPA-categories from Table 1 to set the assignment probabilities for the GPA-categories to different tutorial groups such that the groups are of equal size. For



Note: Circles in the triangle represent the assignment of tutorial groups, larger circles indicate more groups. The location of a circle in the triangle corresponds to the composition of the tutorial group in terms of the shares of students from GPA categories 0, 1 and 2. Circles in the three corners resemble tutorial groups that consist of students from one category only. Circles on the line segments resemble tutorial groups that consist of two categories of students, and circles in the triangle resemble tutorial groups that consist of three categories of students. The cross near the center of the triangle corresponds to the composition of all incoming students.

Figure 2. GPA category composition of assigned tutorial groups

the 2009 cohort, we were required to assign students to tutorial groups *at the moment of application*, which could be anytime between June to September. Because there is a correlation between the date of application and the GPA of new candidates, the higher ability groups were filled more quickly in this procedure, and therefore closed sooner. As this may generate a correlation between moment of registration – which might reflect motivation – and peer ability in the assigned group for this cohort, we include a measure of the application order as control variable in all our regressions.[7] Finally, students with advanced math were grouped together. Since the random assignment of these groups was stratified, we add an advanced math status dummy variable to the interaction terms in the regressions. These groups therefore contribute equally to the identification of peer effects. The complete list of assignment probabilities is given in Table A2 in the appendix.

The assignment of teachers to tutorial groups is done for each course by the coordinator

[7] Another reason why some groups filled up more rapidly, is that we could not use the 2009 prior distribution for setting the probabilities, but had to use the 2008 distribution as a proxy because the 2009 distribution was known only after all new entrants had registered.

of the course. Our design would be contaminated if these coordinators base the assignment of teachers to tutorial groups on the GPA composition of groups. Since only a few people in the faculty were informed about the experiment, we are confident that this did not happen. Unfortunately, we only have data from a limited number of courses to corroborate that. The reason is that the allocation of teachers to groups is not centrally registered and that most course coordinators keep a poor record of the teacher allocation. For our experimental cohorts we managed to obtain the complete allocations for the Math I, Academic skills I, and the Micro course (cf. Table A1 in the appendix). For the Organization course we only obtained the 2011 allocation, the other years were lost in the records of a teacher who left. Regressions of measures of the GPA composition of tutorial groups on the seniority (PhD, Assistant-, or Full-Professor) and gender of the teacher does not show any significant relationships ($N = 3*48+17 = 161$, p-value=0.45). We have therefore no reason to believe that teacher assignment to tutorial groups is related to the GPA composition of the groups.

No-shows and experimental variation. The aim of the assignment procedure is to create large variation in the ability composition of peers across tutorial groups. After students were assigned to groups and the academic year started, we obtained their exact GPA from the student registry. At that stage we were also informed about the students who were assigned to tutorial groups but never showed up and never wrote any exam (no-shows). Since the decision of these students to not show up cannot have been affected by the composition of their assigned tutorial group, we drop them from our data. For the empirical analyses we construct measures of peer quality and peer heterogeneity in a tutorial group based on the actual (continuous) GPA of the students who started their study.[8]

Figure 1 (in the Introduction) is based on this information, and each dot represents one tutorial group. The solid (dashed) circle represents the area where 95% (99%) of the groups would be located when the composition of the tutorial groups would not have been manipulated and students would simply have been randomly assigned to groups. The figure shows that with unconditional random assignment of students to groups peer quality, mean standardized GPA

[8] In subsection 5.1 we also present results from a specification where the GPA composition at the start of the year (excluding no-shows) is instrumented with the GPA composition before the start of the year (including no-shows). The implied peer effects are virtually identical.

per group would, for 95% of the groups, vary between -0.3 and 0.3, while peer heterogeneity, the standard deviation of standardized GPA in a group would, for 95% of the groups, vary between 0.8 and 1.2. The figure therefore clearly illustrates how the experimental design increased the support of the the GPA composition of the tutorial groups relative to naturally occurring variation.

2.3 Data

Our main data come from the student administration of the department of economics and business of the University of Amsterdam.[9] This source contains information on students' gender, birth date, grades on the final exams in secondary education, the assigned tutorial group, and study performance and study status during the first year. Table 2 reports summary statistics, separately for the three cohorts. Panel A shows that almost three quarters of the students is male and that the average age at entrance is somewhat above 19 years old. Students who enroll without any delay, would on average enter at the age of 18.5. The high school GPA of students entering the department of economics and business of the University of Amsterdam equals on average about 6.7 with a standard deviation close to 0.5.[10] Students can also enroll in university after studying in a professional college. The last row of panel A shows that the fraction of students coming through this route is small. All these statistics do not vary much across the three cohorts.

Panel B reports summary statistics of the variables on which the randomization is conditioned. The share of students with a GPA below 6.5 increases somewhat over time. The share of students who took advanced math in high school is around 0.30. Application order captures the moment of registration, where the first applicant is assigned the value 0, and the last the value one 1.

Panel C reports summary statistics of the treatment variables. We summarize the ability composition of the peers who are assigned to the same tutorial group in terms of the mean and standard deviation of the standardized value (by entering cohort) of their GPA in secondary

[9] We also collected additional data through a survey among students. We describe and report about this data source in Section 5.

[10] We take the high school GPA over all courses, as we cannot, at the individual level, separate elective courses from those part of the high school passing criterion.

Table 2. Summary statistics – Mean (SD)

	Cohort		
	2009	2010	2011
A. Background Characteristics			
Male	0.73	0.73	0.74
Age	19.4 (1.6)	19.4 (1.6)	19.4 (1.5)
Professional college	0.05	0.04	0.04
High school GPA: GPA_i	6.7 (0.5)	6.7 (0.5)	6.6 (0.5)
B. Randomization controls			
High school GPA category			
- $GPAcat0$: $GPA < 6.5$	0.26	0.31	0.35
- $GPAcat1$: $6.5 \leq GPA < 7$	0.44	0.42	0.42
- $GPAcat2$: $GPA \geq 7$	0.30	0.27	0.23
Advanced math	0.26	0.37	0.37
Application order (percentile)	0.47 (0.29)	0.50 (0.29)	0.50 (0.29)
C. Peer characteristics			
Mean GPA peers: $\overline{GPA}_{-i}$	0.00 (0.57)	-0.01 (0.54)	0.01 (0.63)
SD GPA peers: $SD(GPA_{-i})$	0.81 (0.28)	0.80 (0.28)	0.74 (0.30)
D. Outcomes			
Credits (raw)	32.1 (22.1)	34.7 (23.3)	32.3 (24.8)
Grades (raw)	5.3 (1.4)	5.6 (1.4)	5.3 (1.7)
Dropout	0.55	0.46	0.46
Nr. of tutorial groups	14	17	17
Nr. of students	606	668	602

school. For each student these values are calculated on the basis of all students assigned to the same group excluding the student's own GPA, hence we use "leave-out" means and standard deviations.

Finally, panel D reports summary statistics of our measures of student performance, which are the outcome variables in our analyses. The first and main performance measure is the number of credit points that students collect in the first year. The maximum number of credit points that students can collect in the first year is 60. This requires them to pass the exams of all 13 first-year courses. The share that manages to do so is low (21%), and the average number of collected credit points is slightly above 30. This shows that there is quite some scope for improvement; we will not fail to find peer effects on the number of credit points because of ceiling effects. The second performance measure is the average grade on the exams taken. While grade point average is a common performance measure, its informativeness is less when students do not take all exams, which may be selective. Only 46% of the students take all first-year exams during the first year. The other 54% of the students skip at least one exam, and on average they skip about 6 exams (out of 13). There is no obvious way to correct students' average grades for this missing information, which is why this is not our main outcome measure. The final performance measure is a dummy variable that equals one for students who collected less than 45 credit points during the first year. Students who fail to collect at least 45 credit points are not allowed to continue studying in the second year, and we refer to this variable as "Drop-out". Drop-out is an important outcome from the perspective of the University of Amsterdam, as it has stated that one of its main goals for the next couple of years is to reduce the share of students that fail to pass the threshold of 45 credit points.[11]

To assess whether the randomization is valid we examine if background characteristics are balanced across tutorial groups with different ability compositions. Columns (1) and (2) of Table 3 show results from regressions of the treatment measures on background characteristics, conditional on students' own GPA-category and application order. This shows as expected no systematic patterns (p-value = 0.94). At the same time, columns (3) to (5) show that the

[11] In the further analyses both the number of credit points and the average grade in the estimation sample are standardized (by cohort) to mean 0 and standard deviation 1. Effect estimates can therefore be interpreted in terms of standard deviation units. The variable Dropout is a dummy, so that effect estimates can be interpreted in terms of percentage point changes.

Table 3. Balancing checks

	Peer characteristic		Outcomes		
	$\overline{GPA}_{-i}$	$SD(\overline{GPA}_{-i})$	Credits	Grade	Dropout
	(1)	(2)	(3)	(4)	(5)
Male	-0.009	0.003	-0.177***	-0.096**	0.106***
	(0.023)	(0.013)	(0.037)	(0.042)	(0.020)
Age = youngest third	0.023	-0.012	0.113**	0.125***	-0.033
	(0.024)	(0.014)	(0.045)	(0.035)	(0.022)
Age = oldest third	0.003	-0.016	-0.118**	-0.044	0.044*
	(0.024)	(0.015)	(0.046)	(0.046)	(0.022)
Professional college	-0.007	0.012	0.050	0.117	-0.037
	(0.043)	(0.029)	(0.115)	(0.117)	(0.059)
GPA	-0.006	0.003	0.314***	0.457***	-0.150***
	(0.019)	(0.011)	(0.033)	(0.031)	(0.018)
Randomization controls	✓	✓	✓	✓	✓
$\bar{y}$	-0.004	0.785	0.000	0.000	0.488
$SD(y)$	0.580	0.289	1.000	1.000	0.500
p-value	0.940		0.000	0.000	0.000
R^2	0.508	0.279	0.266	0.348	0.235
N	1876	1876	1876	1753	1876

Note: Each column reports the results from a different OLS regression. Dependent variable indicated in the column entry. Randomization controls are a saturated set of own GPA category, advanced math and cohort-dummies, interacted with application order. Robust standard errors are in parentheses in columns (1) and (2). Standard errors clustered at tutorial group level are in parentheses in columns (3) to (5). p-value in columns (1) and (2) is based on χ^2-test, p-values in column (3) to (5) on F-tests. */**/*** denote significance at a 10/5/1% confidence level.

background variables are relevant predictors of the outcomes. Male students have worse performance on all three outcomes than female students. Students from the youngest one third of the age distribution collect more credits and get higher grades than others while students from the oldest one third of the age distribution collect fewer credits than others and are more likely to drop out.

3 Empirical specification

Previous experimental studies of ability peer effects in education have examined the effect of one specific peer configuration as compared to another (cf. Carrell et al., 2013; Duflo et al., 2011). This allows for estimation of the effect of interest through a regression of the outcome on a binary treatment indicator (and control variables). The experimental design in the current paper generates large variation in peer GPA configurations, and encompasses the treatments

considered in the previous experimental studies.

The point of departure in the peer effects literature is the linear-in-means model:

$$y_i = \alpha_1 \overline{GPA}_{-i} + \delta GPA_i \tag{1}$$

where y_i is the outcome of student i (credits, grade, Dropout), $\overline{GPA}_{-i}$ is the mean of the prior ability of the other students in the group to which student i is assigned, and GPA_i is student i's own ability. This specification reflects the idea that being surrounded by smarter peers is beneficial. Although the linearity is convenient, it is also restrictive. First, it does not allow for heterogeneous peer effects. Second, only the mean of peer quality matters. Consequently, average performance at the aggregate level is in this setup invariant to the particular allocation of individuals to peer groups. Some studies introduce peer effect heterogeneity by interacting (functions of) GPA_i and $\overline{GPA}_{-i}$ (e.g. Carrell et al., 2009; Burke and Sass, 2013; Brodaty and Gurgand, 2016; Feld and Zölitz, 2016). Introducing other moments of the peer distribution than the mean alone is less common. De Giorgi et al. (2012), Hanushek et al. (2003), Hurder (2012) and Lyle (2009) use the standard deviation to account for peer group heterogeneity and Hurder (2012) uses quantiles.

A general reduced form outcome equation that can describe the variation in our data is the following[12]

$$y_i = g(f(GPA_{-i}), GPA_i) \tag{2}$$

where the outcome y_i depends on a person's own ability, GPA_i, and the relevant group characteristics, $f(GPA_{-i})$, through the function $g(\cdot)$. The aim of our analysis is to approximate equation (2) in our data. Even though our design generates substantial variation in peer group composition, the amount of data available is limited which means that we have to impose some structure on equation (2). In our main analysis, we use the mean and the standard deviation of the prior ability of other students in the same tutorial group to summarize group composition

[12] In this exposition we frame this discussion in terms of GPA. We also abstract from other factors that affect outcomes but our estimation below accounts for this.

$f(GPA_{-i})$ as follows

$$f(GPA_{-i}) \approx \alpha_0 + \alpha_1 \overline{GPA}_{-i} + \alpha_2 SD(GPA_{-i}) + \alpha_{12} \overline{GPA}_{-i} \times SD(GPA_{-i}) \tag{3}$$

where $SD(GPA_{-i})$ is the standard deviation of the prior ability of the other students in the group to which student i is assigned. Inclusion of the mean of peers' prior ability concurs with the canonical linear-in-means model. Inclusion of the standard deviation is for example implied by social cognitive learning theory which argues that achievement may benefit from the presence of similar classmates (Bandura, 1986; Schunk, 1991), but can also be rationalized by a model where students care about their rank in the class (Tincani, 2014). Teachers may also be more effective in more homogenous groups. We also include the product of the mean and the standard deviation of peers' prior ability in equation (3). A specification without this interaction term is restrictive as it imposes that the mean and standard deviation of peer GPA are perfect substitutes in the production of student outcomes.

Graham et al. (2010) also highlight the importance of correctly specifying $f(\cdot)$. They non-parametrically estimate the relation between the share of girls in a classroom and student outcomes. These results show that a parametric specification would have needed a third order polynomial in the share of girls to fit their data. This is particularly important when it comes to estimating the impact of student re-allocation. A second order polynomial results in a relationship between share of girls and student outcomes that is either globally convex or globally concave. In the globally convex case, the optimal allocation is complete segregation. In the globally concave case, it is optimal to assign all classes an equal share of girls. In both cases the optimal assignment rule is independent of the availability of girls in the population. A third order polynomial is the simplest specification where the availability of girls matters for the optimal assignment rule. Our specification in equation (3) is akin to a third order polynomial in shares.[13]

Equation (3), and a first-order approximation of equation (2) can now motivate the following

[13] Appendix B presents results from specifications where the composition of tutorial groups is measured in terms of the shares of students from the top and bottom thirds of the overall GPA distribution. The findings regarding the effects of different ability groupings are virtually identical to the findings reported in Section 4.

estimating equation

$$y_i = \alpha_0 + \alpha_1 \overline{GPA}_{-i} + \alpha_2 SD(GPA_{-i}) + \alpha_{12} \overline{GPA}_{-i} \times SD(GPA_{-i}) + \delta GPA_i + \mathbf{x}_i'\beta + \varepsilon_{ig} \qquad (4)$$

The peer variables $\overline{GPA}_{-i}$ and $SD(GPA_{-i})$ are recentered at mean zero in the regressions. This means that we can interpret α_1 as the effect of mean peer GPA at the sample average of the SD of peer GPA, and similarly α_2 as the effect of the SD of peer GPA at the sample average of mean peer GPA.

We also include a vector of control variables $\mathbf{x}_i$ that includes the randomization controls: a fully saturated set of dummies for each *GPA* category, type of mathematics in secondary education and cohort, interacted with application order. In addition, we add controls for gender, age, and a dummy for students that attended a professional college. The residual ε_{ig} can be heteroskedastic and is clustered at the individual's group level g. To investigate heterogeneous peer effects, we also estimate specifications in which the peer variables are interacted with students' own GPA.[14]

We use the estimates of the preferred specification to simulate the effects of different peer configurations. More specifically, we estimate the effects of two-way tracking (as in Duflo et al.), of three-way tracking, and of the bifurcation that Carrell et al. expected to be optimal (Track Middle) in comparison to the current practice in which students of different ability levels are randomly mixed. Subsection 4.2 describes the simulation procedure in detail.

4 Results

The results are presented in three subsections. Subsection 4.1 presents the estimates from different peer specifications. In Subsection 4.2 we use these results to compute the effects of alternative peer configurations. Subsection 4.3 presents results for other performance measures (average grade and Dropout). Appendix B assesses the robustness of the main findings.

[14]This can be seen as a second order approximation of equation (2) without the quadratic terms in GPA_i and $f(GPA_{-i})$. In an earlier working paper we estimated a specification including higher order terms. The current specification was suggested by a referee, and has the benefit of being more transparent while also describing the data well.

Table 4. Number of credits and peer group composition

	(1)	(2)	(3)	(4)	(5)
$\overline{GPA}_{-i}$	0.051	0.048	0.070	0.095**	0.148***
	(0.043)	(0.041)	(0.043)	(0.046)	(0.052)
$SD(GPA_{-i})$			-0.095	-0.121*	-0.185**
			(0.073)	(0.063)	(0.082)
$\overline{GPA}_{-i} \times SD(GPA_{-i})$				0.423**	0.343*
				(0.176)	(0.190)
GPA_i		0.314***	0.314***	0.317***	0.350***
		(0.033)	(0.034)	(0.034)	(0.035)
$GPA_i \times \overline{GPA}_{-i}$					-0.117***
					(0.042)
$GPA_i \times SD(GPA_{-i})$					0.104
					(0.075)
$GPA_i \times \overline{GPA}_{-i} \times SD(GPA_{-i})$					-0.287**
					(0.138)
Controls:					
- Randomization	✓	✓	✓	✓	✓
- Background		✓	✓	✓	✓
***F*-test (*p*-values):**					
- Peer variables = 0	0.242	0.254	0.222	0.014	0.003

Note: Columns (1) to (5) each present results from a separate OLS regression. Dependent variable is the number of collected credit points in the first year. The peer variables $\overline{GPA}_{-i}$ and $SD(GPA_{-i})$ are recentered at mean zero. All regressions include randomization controls: a saturated set of own GPA category, advanced math-, and cohort-dummies, interacted with application order. Background control variables are gender, age, and a dummy for professional college. Standard errors clustered at tutorial group level are in parentheses. N = 1,876, N clusters = 48. */**/*** denote significance at a 10/5/1% confidence level.

4.1 Peer effects regression estimates

Table 4 shows results from five increasingly flexible peer effects regressions, that illustrate the importance of accounting for peer group characteristics and peer effect heterogeneity. For reference, since this is the most commonly estimated peer effects specification in the literature, the first column presents results from the basic linear-in-means model where only the mean GPA of peers and the randomization controls are included. The point estimate of the coefficient on average peer ability equals 0.051 which, at face value, would imply that a one standard deviation increase of the mean GPA of peers raises the number of credit points a student collects by 5.1% of a standard deviation. The estimate is, however, not significantly different from zero (p-value=0.242). The second column shows that inclusion of control variables for own GPA, as well as gender, age and a dummy for professional college has only a minor impact on the estimated coefficient, as expected given the randomization of peer groups. The coefficient on own GPA is 0.314 and highly significant.

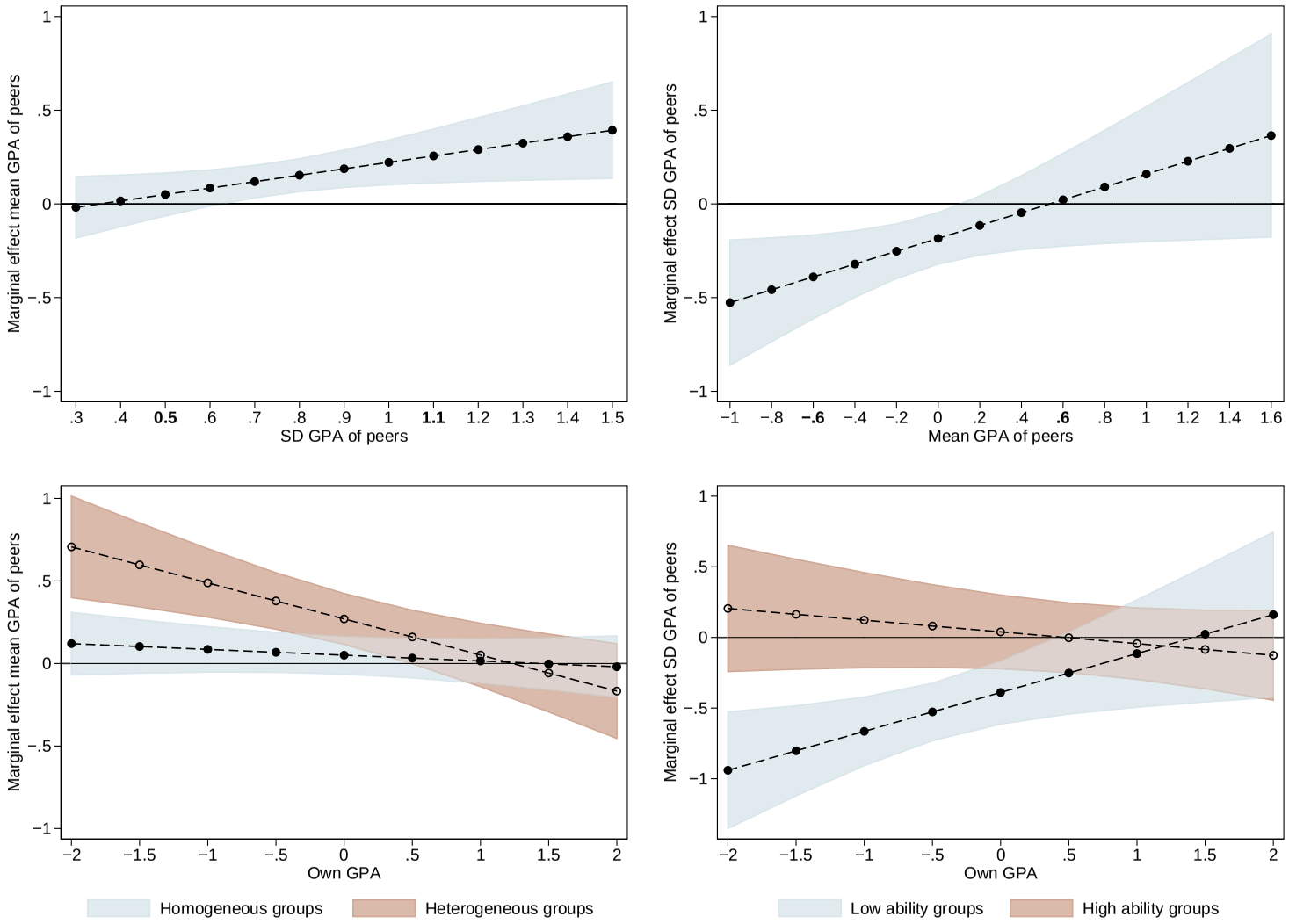
In a first step to account for a richer specification of peer group composition, column (3) adds the standard deviation of peers' GPA to the linear-in-means peer effects regression. The coefficient for mean GPA increases to 0.07. The coefficient of the standard deviation of peers' ability is -0.095. A smaller dispersion of GPA in a group thus increases the number of credit points students collect during the first year. Again, however, we cannot reject that the joint effect of the peer variables is equal to zero (p-value=0.222).

The specification in column (3) imposes that the mean and the standard deviation of peers' GPA are perfect substitutes in the production of student outcomes. Column (4) relaxes this assumption by including the interaction of the mean and the standard deviation of peers' GPA. The coefficient on mean peer GPA is now 0.095 and statistically significant at the 5% level. The interaction of the mean and the standard deviation of peers' GPA is significant at the 5% level. Since the regression includes this interaction, the main effect of mean peer GPA can now be interpreted as the effect of increasing average peer ability in a group with SD of peer GPA equal to the sample average SD of peer GPA (about 0.8). The coefficient on SD peer GPA is -0.121 and significant at the 10% level. It can be interpreted as the effect of increasing peer group heterogeneity in a group with an average mean peer GPA. A joint test of the null-hypothesis that there are no peer effects is now clearly rejected (p-value = 0.014).

The two top panels in Figure 3 show the marginal effects of the peer variables taking the interaction into account.[15] The top left panel reports the effect of mean peer GPA for different values of SD of peer GPA. We see that this effect is insignificant in homogenous groups and increases as SD of peer GPA increases. The top right panel shows the marginal effect of SD of peer GPA. Here we see that increasing peer group heterogeneity is detrimental to performance in groups with low mean peer GPA, and that this effect attenuates as peer quality increases and becomes insignificant.

To introduce peer effect heterogeneity, column (5) adds interaction terms with students' own GPA to the specification of column (4). The main effects of the mean and standard deviation of peers' GPA, which now also must be interpreted as the marginal effect for a student of average ability, increase in absolute size and are significantly positive and significantly negative,

[15]The figure is based on the estimates of the full model in column (5) of Table 4, but is essentially unchanged if we use the estimates in column (4).



Note: All marginal effects based on the specification in column (5) of Table 4. Shared areas are 90% confidence intervals. Homogeneous Groups refers to the 16 groups in our data with the lowest SD of peer GPA, while Heterogeneous Groups refers to the 16 groups with the highest SD of peer GPA. Similarly, Low Ability Groups refers to the 16 groups with the lowest mean peer GPA, and High Ability Groups to the 16 with the highest mean peer GPA.

Figure 3. Average marginal effects of peer group composition – Interaction effects

respectively. The signs of the interaction terms imply that more able students benefit less from more able peers, and are harmed less by peer group heterogeneity. The interaction between own GPA and the standard deviation of peers' ability is, however, not significantly different from zero. Jointly the peer variables in column (5) are significant at the 1%-level (p-value=0.003), and the interaction terms with students' own GPA are jointly significant at the 5%-level (p-value=0.03).[16]

The two bottom panels in Figure 3 illustrate the estimated peer effect heterogeneity. The bottom left panel illustrates how increasing average peer ability affects students of different ability in either homogenous groups or heterogeneous groups. In homogenous groups there is no evidence that more able peers benefit students, independently of their own ability. In contrast, low ability students benefit from more able peers in heterogenous groups, while more able students in such groups do not. The bottom right panel illustrates how increasing peer group heterogeneity affects students of different ability, separately for low and high ability groups. Here we see that there is no evidence for high ability groups that the heterogeneity in the group affects students. For low ability groups on the other hand we find that low ability students are negatively affected by more heterogeneity in the group, while again we find no peer effects for high ability students. In summary we find peer effects for low ability students, but not for high ability students. Low ability students benefit from being with more able peers but not in very homogenous groups, and they are harmed by heterogeneity unless they are placed in a high ability group.

4.2 Alternative assignments

In the previous subsection we presented the estimates of the peer effects regressions. Although these results are informative about peer effects at the margin, the implications for outcomes of assignment of students to groups at the student population level is less straightforward.

Consider two different group allocations $h(i)$ and $k(i)$, where $h(\cdot)$ (and $k(\cdot)$) maps each

[16] We have also estimated regressions including the squares of the mean and standard deviation of peers' GPA, with and without their interactions with own GPA. We cannot reject that the joint effects of these additional variables are equal to zero (p-value=0.489 for the specification without interaction terms with own GPA and p-value=0.409 for the specification with interaction terms with own GPA). The specification reported in column (5) is therefore our preferred specification.

student student to one of H (K) groups. Denote the difference in average achievement of these two allocations by $\bar{y}^{h(\cdot)} - \bar{y}^{k(\cdot)}$. It is well known that this difference equals zero in the simple linear-in-means model (1) since

$$\bar{y}^{h(\cdot)} - \bar{y}^{k(\cdot)} = \frac{1}{N}\sum_i \alpha_1 (\overline{GPA}_{h(i),-i} - \overline{GPA}_{k(i),-i}) = 0$$

and because for any grouping $h(i)$ the average peer is also the average person in the population: $N^{-1}\sum_i \overline{GPA}_{h(i),-i} = \overline{GPA}$. In a linear-in-means model, any grouping will thus produce the same average achievement.

In the model that adds the SD of peer GPA we get that

$$\bar{y}^{h(\cdot)} - \bar{y}^{k(\cdot)} = \frac{1}{N}\sum_i \alpha_2 (SD_{h(i),-i} - SD_{k(i),-i})$$

First note that $0 \leq \frac{1}{N}\sum_i SD_{k(i),-i} \leq SD(GPA)$, the upper bound of which is achieved by randomly assigning students to groups (mixing), and the lower bound by grouping identical students together (perfect tracking).[17] Since our estimate of α_2 is negative, this means that *any* grouping will improve average achievement compared to mixing, and that perfect tracking maximizes average achievement.

These unambiguous implications for group assignment disappear when $\overline{GPA}_{-i}$ and $SD(GPA_{-i})$ are no longer perfect substitutes. In this case both moments matter for average achievement because, at the population level, reassignment will affect $\overline{GPA}_{-i}$ *and* $SD(GPA_{-i})$ at the same time. Different allocations will thus result in different trade-offs between these two peer variables. How these balance out on average in the population depends on the underlying GPA distribution, the group assignment, and the coefficients in the outcome equation. Interactions with own GPA add further trade-offs.

We therefore use the results from the specification in column (5) of Table 4, to estimate average achievement under alternative group assignments compared to ability mixing.[18] To calculate these estimates we compute the means of the peer group variables for the different

[17] Ignoring small sample variation in the case of mixing.

[18] Finding the group assignment that maximizes average achievement is a mixed integer nonlinear programming problem for which no general analytical solutions exist. Note also that the "optimal" assignment may easily imply group compositions (i.e. perfect tracking) that are not in the support of our data.

assignment scenarios we consider:

- Ability mixing: students are grouped together irrespective of their GPA (i.e. randomly)
- Two-way tracking: students are grouped together depending on whether their GPA is above or below the median
- Three-way tracking: students are grouped together depending on whether their GPA is in the bottom, middle or top tertile
- Track Low (or Middle, or High): students from the bottom (or middle, or top) tertile are grouped together, while the remaining students are mixed.

For example, in mixed groups the mean of $\overline{GPA}_{-i}$ will be 0, and the mean of $SD(GPA_{-i})$ will be 1 given the normalization of GPA. Average values for the means and standard deviations of tutorial groups under tracking are obtained from the respective subsamples in our data. For example, for two-way tracking we set the mean and standard deviation of peer GPA in the bottom (top) tutorial group to the mean and standard deviation of the subsample of students with a GPA below (above) the median.[19] Given these mean values of the peer group characteristics in the different scenarios, we can then calculate the average predicted outcomes in our sample using our estimates as well as standard errors. All effects are relative to ability mixing.[20] Table 5 reports the results.

Column (1) of Table 5 reports the mean changes in the number of first-year credits from these alternative assignments in comparison to the current practice of mixing. In columns (2) to (4) the average changes are differentiated by GPA-groups. The first row shows that on average students gain 10% of a standard deviation in achievement from switching from mixing to two-way tracking. This gain is larger for students in the bottom half of the GPA distribution than for students in the top half. The second row shows that a switch to three-way tracking boosts the average achievement gain even further to 14% of a standard deviation. This gain is mainly

[19] We adjust the means of the peer variables for "leave-out". This has a negligible impact on the estimates.

[20] The tracking effects are thus equal to $(\bar{x}_{track} - \bar{x}_{mix})'\hat{\beta}$, and their standard errors to $\sqrt{(\bar{x}_{track} - \bar{x}_{mix})'V(\hat{\beta})(\bar{x}_{track} - \bar{x}_{mix})}$, where $\bar{x}_{track}$ and $\bar{x}_{mixed}$ are sample average covariate vectors including the leave-out means of the peer variables under different scenarios, and $\hat{\beta}$ the coefficients from the peer effects regressions.

Table 5. Estimated tracking effects on first-year credits compared to mixing

			Student GPA category		
		ATE	L (B)	M	H (A)
		(1)	(2)	(3)	(4)
Two-way tracking	{B},{A}	0.099***	0.131***		0.067
		(0.029)	(0.039)		(0.042)
Three-way tracking	{L},{M},{H}	0.138***	0.219***	0.162***	0.034
		(0.037)	(0.070)	(0.059)	(0.058)
Track Low	{L},{M,H}	0.121***	0.219***	0.124***	0.019
		(0.032)	(0.070)	(0.036)	(0.034)
Track Middle	{M}, {L,H}	0.042***	-0.023	0.162***	-0.012
		(0.011)	(0.022)	(0.059)	(0.025)
Track High	{L,M},{H}	0.063**	0.089***	0.064*	0.034
		(0.025)	(0.028)	(0.039)	(0.058)
$\bar{y}$		0.000	-0.505	-0.017	0.523
$SD(y)$		1.000	0.920	0.955	0.845

Note: The table reports (conditional) average treatment effects of different tracking configurations relative to mixing based on the estimates from Table 4, column (5); Student GPA groups are L(ow), M(iddle), H(igh) in case of three-way tracking, and for two-way tracking B(elow) and A(bove). The curly brackets indicate the grouping of GPA groups.

concentrated at students in the lower two thirds of the GPA distribution, who on average gain 19% of a standard deviation of achievement. The other three tracking systems in which students from two thirds of the GPA distribution are mixed, all have a smaller impact on achievement than three-way tracking.

Of special interest are the results of the Track Middle assignment. This is the same assignment as the one that Carrell et al. (2013) expected to be optimal on the basis of the results from the pre-intervention cohorts. Low-GPA students are mixed with high-GPA students and middle-GPA students are kept apart. Our results indicate that this has a slight but insignificant negative effect on low-GPA students, no effect on the high-GPA students and a substantial and significantly positive effect of 16% of a standard deviation on middle-GPA students. These findings are comparable to the unexpected results that Carrell et al. found in their experiment.

4.3 Other outcomes

In this subsection we report and discuss ability peer effects on two other measures of student performance: average grade and collecting less than 45 credit points (Dropout). Table A3 in the appendix reports the regression results, where the first two columns repeat the results

Table 6. Estimated tracking effects compared to mixing

	Outcome variable							
	Av. Grade				Dropout			
	ATE	L (B)	M	H (A)	ATE	L (B)	M	H (A)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Two-way tracking	0.072**	0.100*		0.044	-0.059***	-0.097***		-0.021
	(0.032)	(0.055)		(0.035)	(0.021)	(0.033)		(0.022)
Three-way tracking	0.114***	0.160	0.159***	0.023	-0.077***	-0.161***	-0.070**	-0.001
	(0.041)	(0.100)	(0.054)	(0.047)	(0.028)	(0.055)	(0.035)	(0.029)
Track Low	0.086**	0.160	0.106***	-0.008	-0.071***	-0.161***	-0.043**	-0.010
	(0.037)	(0.100)	(0.034)	(0.029)	(0.023)	(0.055)	(0.020)	(0.017)
Track Middle	0.031***	-0.050**	0.159***	-0.016	-0.022***	-0.003	-0.070**	0.008
	(0.011)	(0.023)	(0.054)	(0.022)	(0.007)	(0.010)	(0.035)	(0.013)
Track High	0.052**	0.096**	0.036	0.023	-0.037**	-0.055**	-0.054**	-0.001
	(0.023)	(0.045)	(0.033)	(0.047)	(0.017)	(0.025)	(0.022)	(0.029)
$\bar{y}$	0.000	-0.567	-0.097	0.623	0.488	0.717	0.499	0.248
$SD(y)$	1.000	0.853	0.877	0.885	0.500	0.451	0.500	0.432

Note: Using estimates from Table A3, specifications 4 and 6.

from columns (4) and (5) of Table 4. The patterns of peer effects are very similar for the other outcome variables. Most variables in Table A3 have the same signs in the regressions for the number of credit points (columns 1 and 2) and average grade (columns 3 and 4), and the opposite signs in the regressions for Dropout (columns 5 and 6). The results for the F-tests in the bottom part of the table also lead to the same conclusions: ability peer effects are nonlinear and heterogeneous.

To better compare the peer effects for different outcome variables taking into account the distribution of GPA in the sample, Table 6 reports effects of different tracking scenarios relative to ability mixing, similar to those reported in Table 5. For both alternative outcome variables, results concur very well with those in Table 5. Switching from ability mixing to three-way tracking reduces the dropout rate of low-GPA students by 16 percentage points, relative to an average dropout rate for this group of 72%. For middle-GPA students the reduction in the dropout rate is 7 percentage points, relative to an average dropout rate for this group of 50%. These are rather substantial reductions in student dropout.

To summarize, the results from the peer effects models in Tables 4 and A3 show that in our data achievement of low-GPA students increases with mean GPA of peers and decreases with the standard deviation of peers' GPA. The positive impact of mean GPA of peers is larger in

more heterogeneous groups, while the negative impact of peer heterogeneity is smaller when average peer quality is higher. There is no evidence that high-GPA students are affected by the composition of their tutorial group. We assessed the aggregate implications for group assignment in counterfactual simulations based on our peer effects regressions. These results show that, given the composition of the incoming students, substantial increases in the performance of low-GPA students can be expected from tracking on the basis of prior ability.

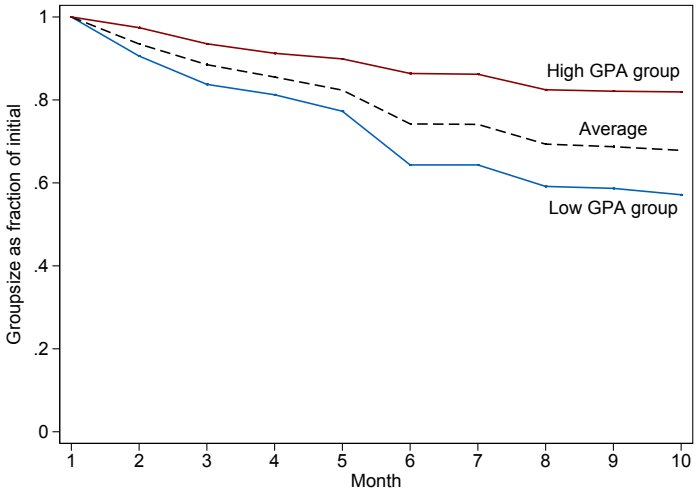
We investigated the robustness of these findings through four sets of results (reported in Appendix B). First, we compare the estimates from the specification where the composition of tutorial groups is measured in terms of the mean and standard deviation of peers' GPA, with those from a specification where the composition of tutorial groups is measured in terms of the shares of students coming from the bottom and top one thirds of the GPA distribution. Second, we compare the results based on three cohorts with results based on only two cohorts (2009 and 2010, 2009 and 2011, and 2010 and 2011). Third, we present results from regressions where we have included other peer characteristics (share of boys, average age and average application order in the tutorial group). Finally, we present results from regressions in which elements of the interaction terms in the specifications of columns (4) and (5) of Table 4 are measured as binary variables instead of continuous variables. Our findings hold up to each of these specification changes.

5 Mechanisms

To gain further insight into the driving forces of the ability peer effects that we have documented, this section examines the relevance of possible mechanisms. In the first subsection we assess to what extent endogenous variation in group size can explain our results. In the next subsection we use data that we obtained through short questionnaires to assess the role of teachers and the influences of peers.

5.1 Group size

Low-GPA students are more likely than others to drop out during the year. Consequently, tutorial groups with more low-GPA students will for most of the year be smaller than groups with



Note: The graph shows the changes in group size during the first year for the 16 groups in our data with the lowest, middle, and highest average GPA at the start of the year.

Figure 4. Group size throughout the year

fewer low-GPA students. This is illustrated in Figure 4 which shows how group size evolves over the year for the 16 groups in our data with the lowest, and highest mean GPA at the start of the year.[21] Compared to the start of the year high mean GPA groups see group size fall with about 20%. Low mean GPA groups experience a much larger decrease in group size, reflecting that low ability students are more likely to drop out, and at the end of the year group size is reduced by 40%. This illustrates that if the size of tutorial groups has an independent effect on student achievement, then part of our findings should be attributed to the group size effect instead of a pure ability peer effect.

To assess the role of differences in group size between tutorial groups with a different GPA composition, we include the average size of a group during the year as an additional regressor in our main peer effects specification. Average group size is however potentially endogenous since it is an outcome of the ability composition of the group. We will therefore instrument the size of a tutorial group with the number of students that was assigned to that tutorial group but never showed up. The average number of no-shows per group is 2.1, with a standard deviation of 2.3.[22] A regression of actual average group size during the year on the number of no-shows

[21] The graph assumes that students dropped out when they stopped taking exams.

[22] Cohort 2009 does not have any no-shows because there our estimation sample consists of groups that were formed after knowing the actual presence of students, by redistributing students from the smallest groups to the remaining ones. Hence, this cohort does not contribute to the identification of the tutorial group size effect.

gives a first stage coefficient of -0.66 (s.e. 0.12; F-value 31.4).[23] Because no-shows were never exposed to the intended peers in their tutorial group or informed about them, their decision not to start the program cannot have been influenced by the GPA composition of their tutorial group.

Table 7 reports results from three regressions. To ease comparison, column (1) repeats column (5) of Table 4, our preferred specification. Column (2) reports results from an IV-regression that includes the average size of a group during the year as additional control, which is instrumented by the number of no-shows. The estimates of the peer group effects are virtually unchanged when group size is included; the coefficients in columns (1) and (2) are almost the same and so are the estimated tracking effects.

One threat to the validity of the coefficient of average group size in column (2) is that the GPA composition of a group is possibly affected by the number of no-shows (the instrument for group size). This would make the ability peer variables potentially bad control variables when one is interested in the effect of group size. To address this concern, we instrument actual GPA composition of the group at the start of the year (excluding the no-shows) with the the assigned GPA composition of the group before the start of the year (including the no-shows). The results are presented in column (3) of Table 7. Column (4) reports the F-statistics of the first-stage relationships between the instrumented variables and the instruments, indicating instrument relevance. The coefficient of group size remains small and statistically insignificant, indicating that average group size during the year has no impact on student outcomes. Consistent with this, we see that the estimates of the peer variables in column (3) are very similar to those in column (2), although in some cases less precise. Jointly, the peer variables in column (3) are statistically significant (p-value=0.014). Homogeneity of the peer effects cannot be rejected at conventional levels (p-value=0.15). Most importantly, the simulated effects from a switch from ability mixing to ability tracking in columns (3) and (2) are very similar.

In sum, the higher dropout rates of low-GPA students cause a reduction of the average size of

[23]The average size of a tutorial group during the year is calculated as the sum over all students, where a student that drops out after the first month (i.e. was observed in the exam-records only in the first month) counts for 1/10, 2nd month 2/10 et cetera (see Figure 4). A student that is still observed at the end of the year has 10/10 and counts for 1. Note that ultimate dropout is higher than what is observed in the 10th month in Figure 4 because some students fail the re-sits in August.

Table 7. Results on credits in the first year controlling for group size

	OLS	IV		
	(1)	(2)	(3)	(4)
$\overline{GPA}_{-i}$	0.148***	0.121	0.090	[1579.303]
	(0.052)	(0.084)	(0.085)	
$SD(GPA_{-i})$	-0.185**	-0.175**	-0.132	[1873.850]
	(0.082)	(0.085)	(0.084)	
$\overline{GPA}_{-i} \times SD(GPA_{-i})$	0.343*	0.305	0.376*	[409.093]
	(0.190)	(0.215)	(0.212)	
GPA_i	0.350***	0.350***	0.351***	
	(0.035)	(0.035)	(0.035)	
$GPA_i \times \overline{GPA}_{-i}$	-0.117***	-0.115***	-0.106**	[1776.946]
	(0.042)	(0.041)	(0.046)	
$GPA_i \times SD(GPA_{-i})$	0.104	0.100	0.042	[2919.354]
	(0.075)	(0.072)	(0.081)	
$GPA_i \times \overline{GPA}_{-i} \times SD(GPA_{-i})$	-0.287**	-0.298**	-0.267	[522.098]
	(0.138)	(0.146)	(0.196)	
Group size		0.005	0.003	
		(0.014)	(0.015)	
		[31.423]	[28.593]	
Controls				
Randomization	✓	✓	✓	
Background	✓	✓	✓	
F-tests (p-values):				
- Peer variables = 0	0.003	0.001	0.014	
- Peer effects homogenous	0.030	0.017	0.150	
Predicted Tracking Effect:				
Two-way tracking	0.099***	0.099***	0.082***	
	(0.029)	(0.029)	(0.032)	
Three-way tracking	0.138***	0.139***	0.113***	
	(0.037)	(0.037)	(0.041)	

Note: Dependent variable is the number of collected credit points in the first year. The peer variables $\overline{GPA}_{-i}$ and $SD(GPA_{-i})$ are recentered at mean zero. All regressions include randomization controls: a saturated set of own GPA category, advanced math-, and cohort-dummies, interacted with application order. Background control variables are gender, age, and a dummy for professional college. Standard errors clustered at tutorial group level are in parentheses, and first stage F-statistics in brackets. N = 1,876, N clusters = 48. */**/*** denote significance at a 10/5/1% confidence level.

groups with many low-GPA students. Average group size does not, however, have a significant effect on student performance, and consistent with that the peer effect estimates are unaffected by the inclusion of group size in the analysis. We interpret this as showing that the peer effects we estimate in column (1) are not contaminated by group size effects.

5.2 Teachers and peers

Three months after the start of the academic years covered in the analysis, we carried out a survey among the students asking them about the learning environment in their tutorial groups, their interaction with other students and with teachers of their tutorial group, as well as the teaching style of these teachers. We chose to do this after three months to strike a balance between students being able to give informed responses to the questions and not too many students having dropped out already. Table C1 in Appendix C lists the items that were included in the surveys together with the number of respondents, the scale on which they are measured and the means and standard deviations of the responses. Since the survey questions were somewhat changed between the first and second cohorts, it also indicates which questions were asked to which cohorts. The response rate to the survey is around 70% in all three years. The first column in Table A4 in the appendix shows that response to the surveys is independent of the ability composition of the tutorial groups (p-value=0.659).

To summarize the information from the 26 survey items, we constructed six index variables which each are the unweighted sum of three or more items (an item is never used for more than one index variable). These sums were normalized to have mean zero and standard deviation one. We label these six index variables as follows (Table C1 in the appendix reports the items that were used to construct these variables):

- Too fast: tutorial group teachers are too fast, spend too little time on simple things or give complicated answers;
- Too slow: tutorial group teachers are too slow, spend too much time on simple things or focus too much on weak students;
- Stimulating: learns a lot from tutorial group teachers, group meetings are stimulating or

teacher asks questions to test our understanding;

- Conducive: good atmosphere in tutorial group, learns from students in tutorial group, tutorial group influences performance positively;
- Interaction: study together, help other students or is helped by other students;
- Involved: Me or others frequently ask questions; level of other students demotivates me (-), dislike to ask questions (-); unquietness makes it difficult to concentrate (-).

For respondents who did not answer all items, we assigned mean values from the other respondents to these items. In particular, we imputed mean values from the respondents in 2010 and 2011 to the respondents in 2009 for the items that were not included in the 2009 survey. The first three variables are related to the (perceived) behavior of teachers, the last three variables capture elements of the direct influence of peers.

Table A4 in the appendix reports results from peer effects regressions in which each of the six constructed index variables are the dependent variables and in which we use the same specification as in column (5) of Table 4. Table 8 reports how tracking affects reported peer and teacher behavior based on the estimates of Table A4. The top part of the table reports the average effect from a switch from ability mixing to two-way tracking. This is reported for all students together and separately for students with GPA below and above the median GPA. There is no evidence that the (perceived) behavior of teachers is significantly affected by tracking. This is different for the influence of peers. Students from the bottom half of the GPA distribution experience more positive interactions with the other students in their tutorial group, and they are more involved. The magnitudes of these effects are 23% and 22% of a standard deviation of the dependent variables.

The bottom part of Table 8 shows the average effect of a switch from ability mixing to three-way tracking. Estimates are reported for all students together and for the lowest, middle, and highest one thirds of the GPA distribution. Low-GPA and middle-GPA students feel more involved in a tracked group than in a mixed group. Low-GPA students also experience more positive interactions with the other students in their tutorial group. For high-GPA students there is no impact on the reported influence of peers. Again, there is no indication that the (perceived)

Table 8. Mechanisms

	Teachers			Peers		
	Too fast	Too Slow	Stimulating	Conducive	Interaction	Involved
	(1)	(2)	(3)	(4)	(5)	(6)
A. Two-way tracking						
ATE	0.059	0.047	0.034	-0.019	0.134**	0.114*
	(0.075)	(0.068)	(0.120)	(0.061)	(0.063)	(0.063)
- Below	0.131	0.132	0.105	0.023	0.227**	0.218***
	(0.114)	(0.102)	(0.186)	(0.083)	(0.096)	(0.084)
- Above	-0.012	-0.039	-0.036	-0.061	0.042	0.009
	(0.077)	(0.083)	(0.088)	(0.070)	(0.060)	(0.068)
B. Three-way tracking						
ATE	0.071	0.068	0.080	-0.036	0.171**	0.168**
	(0.095)	(0.088)	(0.155)	(0.080)	(0.085)	(0.082)
- Low	0.188	0.240	0.114	0.038	0.339**	0.346***
	(0.186)	(0.166)	(0.276)	(0.109)	(0.152)	(0.122)
- Middle	0.056	-0.002	0.193	-0.057	0.156	0.202*
	(0.102)	(0.110)	(0.169)	(0.118)	(0.114)	(0.114)
- High	-0.031	-0.036	-0.069	-0.088	0.019	-0.045
	(0.093)	(0.114)	(0.110)	(0.100)	(0.075)	(0.080)

Note: */**/*** denote significance at a 10/5/1% confidence level.

behavior of teachers is affected by tracking.

To summarize, students from the lower end of the GPA distribution report more positive interaction with their tutorial group peers and are more involved in a tracked group than in a mixed ability group. To the extent that positive interaction with peers and feeling more involved contribute to student achievement, these two mechanisms help explain the ability peer effects that we identified in Section 4.

6 Conclusion

This paper reports on an experiment that manipulated the prior ability composition of tutorial groups for undergraduate students in economics. The design of the experiment was aimed to create variation in peer ability over a large support which allows us to estimate flexible contextual peer effects regressions. The data are consistent with a specification where student outcomes depend on the mean and the standard deviation of peers' prior ability, their interaction and interactions with students' own prior ability. While these results are directly informative about the effects of marginal changes in peer group composition, we also assess their implica-

tions for aggregate effects of different forms of ability grouping. Compared to previous experimental studies, our design permits us to compare more than two ability groupings. Compared to previous quasi-experimental studies, we do not have to extrapolate beyond the support of our data.

We find substantial positive effects from ability grouping on the achievement of students from the lower two thirds of the GPA distribution. In terms of credits points these students gain on average 0.19 SD units of achievement from switching from ability mixing to three-way ability tracking. The dropout rate of these students is reduced by around 12 percentage points (relative to a mean of 0.60). High-GPA students are unaffected. Analysis of survey data points to two underlying mechanisms. In tracked groups, low-ability students i) have more positive interaction with other students, and ii) are more involved. We find no evidence that teachers adjust their teaching to the composition of tutorial groups.

Our findings are broadly consistent with results in Duflo et al. (2011) and Carrell et al. (2013). Like Duflo et al. (2011) we find that group homogeneity benefits performance and that low-ability students gain from tracking. Like Carrell et al. (2013) we also find that low-ability students perform worse when assigned to groups with high variance in ability. Carrell et al. infer that this is due to the formation of subgroups. Our finding that positive peer interaction decreases with group heterogeneity is in line with that inference. Like Carrell et al., but unlike Duflo et al., we find that high-ability students are unaffected by the ability composition of their group. Interestingly, when we simulate the group composition that Carrell et al. believed to be optimal – place low-ability and high-ability students together and keep middle-ability students separate – we reproduce what Carrell et al. find in their experiment: low-ability students are harmed and middle-ability students benefit. This demonstrates the value-added of the wide variation in group composition in our study.

The similarity in findings across the different studies is remarkable given the large contextual differences. Duflo et al. (2011) study peer effects among students in the first grade of primary school in Kenya, Carrell et al. (2013) look at a quite specific population of entering freshmen at the United States Air Force Academy, while our study examines ability peer effects among first year students in a non-selective undergraduate program in economics. Even the

non-selective undergraduate program in economics in our study recruits its students from the top 20% of the ability distribution of their age cohorts. The low-ability students in our sample are therefore only of low ability relative to the other students in our sample, not to the Dutch population. It also implies that tracking is beneficial even in an already homogenous group of students.

Many education institutes (primary schools, secondary schools, universities, air force academies) have incoming cohorts that are divided into subgroups (sections, tutorial groups, squadrons). Our results show that there is a potential gain in assigning students to these subgroups in a systematic way. This gain comes at no financial cost, nor do we find indications that some groups of students are harmed in the form of lower achievement.

References

Ammermueller, A. and Pischke, J.-S. (2009). Peer effects in European primary schools: Evidence from the Progress in International Reading Literacy Study. *Journal of Labor Economics*, 27(3):315–348.

Angrist, J. D. (2014). The perils of peer effects. *Labour Economics*, 30:98–108.

Bandura, A. (1986). *Social Foundations of Thought and Action: A Social Cognitive Theory*. Prentice Hall.

Black, S., Devereux, P., and Salvanes, K. (2013). Under pressure? The effect of peers on outcomes of young adults. *Journal of Labor Economics*, 31(1):119–153.

Brodaty, T. and Gurgand, M. (2016). Good peers or good teachers? evidence from a french university. *Economics of Education Review*, 54:62 – 78.

Burke, M. A. and Sass, T. R. (2013). Classroom peer effects and student achievement. *Journal of Labor Economics*, 31(1):51–82.

Carrell, S. E., Fullerton, R. L., and 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., and West, J. E. (2013). From natural variation to optimal policy? The importance of endogenous peer group formation. *Econometrica*, 81:855–882.

De Giorgi, G., Pellizzari, M., and Woolston, W. G. (2012). Class size and class heterogeneity. *Journal of the European Economic Association*, 10(4):795–830.

Duflo, E., Dupas, P., and Kremer, M. (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya. *American Economic Review*, 101(5):1739–1774.

Feld, J. and Zölitz, U. (2016). On the nature of peer effects in academic achievement. *Journal of Labor Economics*, 35(2).

Glaeser, E., Sacerdote, B., and Scheinkman, J. (2003). The social multiplier. *Journal of the European Economic Association*, 1:345–353.

Graham, B. S., Imbens, G. W., and Ridder, G. (2010). Measuring the effects of segregation in the presence of social spillovers: A nonparametric approach. NBER Working Paper 16499.

Hanushek, E. A., Kain, J. F., Markman, J. M., and Rivkin, S. G. (2003). Does peer ability affect student achievement? *Journal of Applied Econometrics*, 18(5):527–544.

Hoxby, C. (2000). Peer effects in the classroom: Learning from gender and race variation. NBER Working Paper 7867.

Hurder, S. (2012). Evaluating econometric models of peer effects with experimental data. Unpublished working paper.

Lavy, V., Paserman, M. D., and Schlosser, A. (2012a). Inside the black box of ability peer effects: Evidence from variation in the proportion of low achievers in the classroom. *Economic Journal*, 122(559):208–237.

Lavy, V., Silva, O., and Weinhardt, F. (2012b). The good, the bad, and the average: Evidence on ability peer effects in schools. *Journal of Labor Economics*, 30(2):pp. 367–414.

Lyle, D. S. (2009). The effects of peer group heterogeneity on the production of human capital at West Point. *American Economic Journal: Applied Economics*, 1(4):69–84.

Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *The Review of Economic Studies*, 60(3):531–542.

Sacerdote, B. (2001). Peer effects with random assignment: Results for Dartmouth roommates. *Quarterly Journal of Economics*, 116(2):681–704.

Sacerdote, B. (2014). Experimental and quasi-experimental analysis of peer effects: Two steps forward? *Annual Review of Economics*, 6(1):253–272.

Schunk, D. H. (1991). *Learning Theories: An Educational Perspective*. Merrill, New York.

Tincani, M. M. (2014). Heterogeneous peer effects and rank concerns: Theory and evidence. Unpublished working paper.

Appendices

A Additional tables

Appendix A present four additional tables. Table A1 gives an overview of the courses in the first year of the undergraduate program in economics and business at the University of Amsterdam in the years 2009-2011. Per course it reports the term in which it is taught, the number of total teaching hours, the number of teaching hours in tutorial groups and the number of credit points.

Table A2 shows for each tutorial group in each year what shares of students from GPA-categories 0, 1 and 2 were assigned to it. It also reports whether the group was an Advanced Math group or not.

Table A3 shows the estimation results from the specification in columns (5) and (6) of Table 4 for two alternative outcome variables: Average grade and Dropout. For ease of comparison, columns (1) and (2) repeat the results from columns (5) and (6) of Table 4. The results in columns (4) and (6) are the basis for the simulation results reported in Table 6 of the main text.

Table A4 shows the estimation results from the specification in columns (5) and (6) of Table 4 for six variables that were constructed from the survey responses (columns 2 to 7), and for an indicator of response to the survey (column 1). The results in columns (2) to (7) are the basis for the simulation results reported in Table 8 of the main text.

Table A1. Overview of the first-year courses in the economics and business program

Course	Term	Total teaching hours	Tutorial group hours	Credit points
Financial accounting	1	28	14	5
Organization	1	12	12	5
Orientation fiscal economics	1	6	0	2
Mathematics 1	1 and 2	56	28	5
Academic skills 1	1 and 2	28	28	2
Management accounting	2	28	14	4
Microeconomics	2	42	28	7
Organization and management	3	28	14	6
Statistics	3	42	14	5
Mathematics 2	3 and 4	56	28	4
Academic skills 2	3 and 4	28	28	3
Finance	4	21	21	5
Macroeconomics	4	42	28	7
Total		417	257	60

Note: The table gives information about the courses in the first year of the undergraduate program of economics and business at the University of Amsterdam. It reports the terms in which courses are taught (1 to 4), total number of teaching hours per course, number of tutorial group hours per course and number of credit points per course.

Table A2. Group assignment probabilities conditional on GPA category and advanced math

	Cohort												
	2009				2010				2011				
	GPAcat			*Advanced*		*GPAcat*			*Advanced*		*GPAcat*			*Advanced*
Group	0	1	2	*Math*	Group	0	1	2	*Math*	Group	0	1	2	*Math*
1	0.00	0.43	0.50	1	15	0.25	0.00	0.28	1	32	0.00	0.00	0.63	1
2	0.60	0.00	0.50	1	16	0.00	0.31	0.17	1	33	0.00	0.42	0.00	1
3	0.40	0.57	0.00	1	17	0.25	0.23	0.00	1	34	0.03	0.37	0.05	1
4	0.25	0.00	0.00	0	18	0.00	0.00	0.55	1	35	0.24	0.00	0.32	1
5	0.25	0.00	0.00	0	19	0.00	0.46	0.00	1	36	0.24	0.21	0.00	1
6	0.00	0.25	0.00	0	20	0.51	0.00	0.00	1	37	0.49	0.00	0.00	1
7	0.00	0.25	0.00	0	21	0.00	0.10	0.18	0	38	0.00	0.00	0.41	0
8	0.00	0.00	0.33	0	22	0.15	0.00	0.18	0	39	0.00	0.11	0.20	0
9	0.04	0.04	0.22	0	23	0.09	0.09	0.10	0	40	0.00	0.21	0.00	0
10	0.04	0.04	0.22	0	24	0.10	0.07	0.12	0	41	0.00	0.21	0.00	0
11	0.08	0.08	0.11	0	25	0.10	0.07	0.12	0	42	0.05	0.13	0.09	0
12	0.04	0.17	0.06	0	26	0.05	0.03	0.24	0	43	0.09	0.09	0.09	0
13	0.17	0.04	0.06	0	27	0.20	0.03	0.06	0	44	0.10	0.13	0.00	0
14	0.12	0.12	0.00	0	28	0.00	0.20	0.00	0	45	0.10	0.13	0.00	0
					29	0.00	0.20	0.00	0	46	0.13	0.00	0.20	0
					30	0.30	0.00	0.00	0	47	0.26	0.00	0.00	0
					31	0.00	0.20	0.00	0	48	0.26	0.00	0.00	0

Note: The table shows for each tutorial group in each year which share of students from GPA-categories 0, 1 and 2 are assigned to the group. Assignment is separate for Advanced Math groups and regular groups (indicated in the columns Advanced Math). Due to rounding shares per "cohort*GPA category*Advanced math"-category do not always add up to 1.

Table A3. Results on other outcomes

	Outcome variable					
	Credits		Average grade		Dropout	
Peer Prior *GPA*	(1)	(2)	(3)	(4)	(5)	(6)
$\overline{GPA}_{-i}$	0.095**	0.148***	0.062	0.124**	-0.031	-0.042
	(0.046)	(0.052)	(0.043)	(0.047)	(0.023)	(0.027)
$SD(GPA_{-i})$	-0.121*	-0.185**	-0.157***	-0.203***	0.045	0.066
	(0.063)	(0.082)	(0.057)	(0.074)	(0.036)	(0.045)
$\overline{GPA}_{-i} \times SD(GPA_{-i})$	0.423**	0.343*	0.101	-0.019	-0.306***	-0.293***
	(0.176)	(0.190)	(0.202)	(0.200)	(0.087)	(0.090)
GPA_i	0.317***	0.350***	0.459***	0.489***	-0.152***	-0.168***
	(0.034)	(0.035)	(0.032)	(0.035)	(0.018)	(0.018)
$GPA_i \times \overline{GPA}_{-i}$		-0.117***		-0.109***		0.042*
		(0.042)		(0.035)		(0.021)
$GPA_i \times SD(GPA_{-i})$		0.104		0.162*		-0.008
		(0.075)		(0.081)		(0.038)
$GPA_i \times \overline{GPA}_{-i} \times SD(GPA_{-i})$		-0.287**		-0.339**		0.134*
		(0.138)		(0.143)		(0.074)
Randomization controls		✓		✓		✓
Controls		✓		✓		✓
$\bar{y}$		0.000		0.000		0.488
$sd(y)$		1.000		1.000		0.500
$N_{cluster}$		1876		1753		1876
N		48		48		48
F-tests (p-values)						
Peer variables = 0	0.014	0.003	0.052	0.017	0.011	0.022
Peer effects linear	0.005	0.002	0.022	0.009	0.004	0.013
Peer effects homogenous		0.030		0.006		0.096

Note: Columns (1) and (2) repeat columns (5) and (6) in Table 4. Other columns have the same specification but different dependent variables. Grade is the average grade students received for exams they wrote in the first year, weighted by the number of credits of the exam. Not all students write all exams. Some students write no exam at all; this explains the smaller number of observations for this outcome. Dropout equals one if the student collected fewer than 45 credits points in the first year, not allowing the student to continue in the second year. The peer variables $\overline{GPA}_{-i}$ and $SD(GPA_{-i})$ are recentered at mean zero. F-tests are reported for null-hypotheses that joint effect of peer variables equals zero (cf(Peer variables) = 0), that joint effect of nonlinear peer variables equals zero (cf(Non-lin. terms) = 0), and that peer effects are the same for different GPA groups in (cf(Peer var.) = homo.). Randomization controls are a saturated set of own GPA category, advanced math-, and cohort-dummies, interacted with application order. Other control variables are own GPA, gender, age, and a dummy for professional college. Standard errors clustered at tutorial group level are in parentheses. */**/*** denote significance at a 10/5/1% confidence level.

Table A4. Mechanisms

	Responded	Teachers			Peers		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Peer Prior *GPA*		Too fast	Too Slow	Stim.	Conducive	Interact	Involved
$\overline{GPA}_{-i}$	0.011	-0.027	-0.019	-0.024	-0.057	0.052	0.065
	(0.044)	(0.084)	(0.081)	(0.075)	(0.065)	(0.092)	(0.095)
$SD(GPA_{-i})$	-0.057	-0.039	0.026	-0.268	0.095	-0.150	-0.229
	(0.067)	(0.127)	(0.141)	(0.206)	(0.151)	(0.143)	(0.146)
$\overline{GPA}_{-i} \times SD(GPA_{-i})$	-0.179	0.211	-0.059	-0.360	0.032	0.417	0.326
	(0.113)	(0.307)	(0.289)	(0.348)	(0.283)	(0.254)	(0.255)
GPA_i	0.034*	-0.117**	0.102	-0.005	0.025	0.014	0.076**
	(0.019)	(0.053)	(0.071)	(0.060)	(0.035)	(0.052)	(0.034)
$GPA_i \times \overline{GPA}_{-i}$	-0.029	0.013	0.013	-0.097	0.019	-0.022	-0.134**
	(0.026)	(0.051)	(0.084)	(0.074)	(0.056)	(0.068)	(0.058)
$GPA_i \times SD(GPA_{-i})$	0.028	0.054	0.155	0.060	0.064	0.121	0.050
	(0.045)	(0.127)	(0.166)	(0.165)	(0.122)	(0.109)	(0.086)
$GPA_i \times \overline{GPA}_{-i} \times SD(GPA_{-i})$	-0.050	0.025	-0.329	-0.281	0.066	-0.089	-0.456**
	(0.084)	(0.249)	(0.330)	(0.344)	(0.294)	(0.243)	(0.189)
Randomization controls	✓	✓	✓	✓	✓	✓	✓
Controls	✓	✓	✓	✓	✓	✓	✓
$\overline{y}$	0.715	0.000	0.000	0.000	0.000	0.000	0.000
$sd(y)$	0.451	1.000	1.000	1.000	1.000	1.000	1.000
$N_{cluster}$	48	47	47	47	47	47	47
N	1876	1342	1342	1342	1342	1342	1342
F-tests (p-values)							
- Peer variables = 0	0.659	0.629	0.670	0.023	0.944	0.206	0.004
- Peer effects homogenous	0.716	0.870	0.497	0.520	0.871	0.741	0.026

Note: Columns (1) to (7) each present results from a separate OLS regression. Dependent variable in first column is an indicator which equals 1 if student responded to the survey, otherwise 0. Other dependent variables are constructed variables based on survey. Main explanatory variables are mean and standard deviation of standardized GPA of peers in tutorial group. The peer variables $\overline{GPA}_{-i}$ and $SD(GPA_{-i})$ are recentered at mean zero. F-tests are reported for null-hypotheses that joint effect of peer variables equals zero (cf(Peer variables) = 0), and that peer effects are the same for different GPA groups (cf(Peer var.) = homo.). Randomization controls are a saturated set of own GPA category, advanced math-, and cohort-dummies, interacted with application order. Other control variables are own GPA, gender, age, and a dummy for professional college. Standard errors clustered at tutorial group level are in parentheses. */**/*** denote significance at a 10/5/1% confidence level.

B Robustness

In Tables 4 and A3 in the main text, the ability composition of the peer group is expressed in terms of the mean and standard deviation of peers' GPA. In this appendix we present results from a specification where the GPA composition of peers in the tutorial group is measured by the shares from the bottom, middle, and top one thirds of the population of incoming students.

Table B1 reports results from the share-based specification.[24] In accordance with the findings of Graham et al. (2010) we include third order polynomials of the shares of peers in the tutorial group from the lowest one third and the highest one third of the overall GPA-distribution. Table B2 translates the estimates into the effects of different forms of tracking. Columns (1) to (4) repeat the results from the main text, columns (5) to (8) report the results based on the specification in terms of shares. It is reassuring that the simulation results based on these two different specifications are rather similar. The results based on the specification with mean and standard deviation are more precise, which is unsurprising because the share based specification ignores information on exact GPA's.

Tables B3 to B4 report two further robustness checks. In Table B3 we have included other peer characteristics as additional regressors. These are the share of boys in the tutorial group, the average age of students in the tutorial group and the average application order. Each of these are correlated with prior GPA and the ability peer effects that we report may potentially be due to these characteristics. The results in Table B3 show that the ability peer estimates are robust to the inclusion of these variables. In Table B4 we report estimates that are based on two instead of three years of data. This assesses whether our results are driven by one specific cohort. These results indicate that data for any two of the three cohorts gives the same qualitative conclusions.

The first column in Table B5 reports results from a regression in which continuous SD of peer GPA is replaced by dummies indicating whether SD of peer GPA is below or above the median of the distribution of the SD of peer GPA. The results show that the coefficient of mean GPA is larger when SD is in the top half than when SD is in the bottom half. This is consistent with the pattern in the upper-left graph in Figure 3. The difference in coefficients is, however, not statistically significant. The second column in Table B5 reports results from a regression in

[24] Shares are also calculated using the leave-out approach.

which continuous mean peer GPA is replaced by dummies indicating whether mean peer GPA is below or above the median of the distribution of the mean peer GPA. The results show that the coefficient of the SD of peer GPA is larger (less negative) when mean GPA is in the top half than when mean GPA is in the bottom half. This is consistent with the pattern in the upper-right graph in Figure 3. This difference in coefficients is statistically significant (p-value=0.060).

Table B6 reports results from a regression in which continuous own GPA in the interactions with peer variables is replaced by indicators of own GPA being below or above the median of the distribution of own GPA. The results in this table show that peer effects are much stronger for students with GPA below the median than for students with GPA above the median. The difference in coefficients is statistically significant (p-value=0.055). For students with below median GPA the main effect of mean peer GPA is positive, the main effect of SD of peer GPA is negative for them, and the interaction of the two peer variables has a positive coefficient . These patterns are consistent with the results in column (5) of Table 4 and with the bottom graphs in Figure 3.

Table B1. Regression of credits on shares of peer from bottom and top one thirds of GPA distribution

Peer Prior *GPA*	(1)	(2)	(3)	(4)	(5)	
					Main	$\times GPA_i$
$FLow_{g-i}$	-0.095	-0.252*	-0.391*	-0.306	-0.426*	-0.083
	(0.140)	(0.148)	(0.205)	(0.271)	(0.243)	(0.232)
$FHigh_{g-i}$	0.048	0.071	0.255	0.499**	0.230	-0.342
	(0.121)	(0.133)	(0.200)	(0.231)	(0.207)	(0.209)
$FLow^2_{g-i}$		0.942***	-0.210	-0.533	-0.071	-0.098
		(0.299)	(0.657)	(0.603)	(0.661)	(0.766)
$FHigh^2_{g-i}$		-0.350	1.043	0.305	1.121	-1.299*
		(0.291)	(0.826)	(0.956)	(0.860)	(0.750)
$FLow^3_{g-i}$			1.904	0.019	1.621	0.866
			(1.541)	(3.429)	(2.154)	(2.120)
$FHigh^3_{g-i}$			-2.333	-6.888**	-1.935	2.377**
			(1.421)	(2.926)	(1.419)	(1.163)
$FLow^4_{g-i}$				3.112		
				(5.355)		
$FHigh^4_{g-i}$				8.118		
				(5.113)		
Randomization controls	✓	✓	✓	✓	✓	
Controls	✓	✓	✓	✓	✓	
F-tests (p-values)						
Peer variables = 0	0.449	0.014	0.000	0.000	0.000	
Added terms = 0	0.449	0.003	0.067	0.260	0.015	

Note: The table reports results from a regression of standardized number of credits on third order polynomials of the fractions of peers from lowest and highest one thirds of overall GPA distribution, interacted with students' own GPA. The peer variables are recentered at mean zero. Randomization controls are a saturated set of own GPA category, advanced math-, and cohort-dummies, interacted with application order. Other control variables are own GPA, gender, age, and a dummy for professional college. Standard errors clustered at tutorial group level are in parentheses. */**/*** denote significance at a 10/5/1% confidence level.

Table B2. Estimated tracking effects compared to mixing, based on OLS estimates

	Peer statistic							
	Mean and standard deviation				Shares from bottom and top one thirds			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Tracking	ATE	L (B)	M	H (A)	ATE	L (B)	M	H (A)
Two-way tracking	0.099***	0.131***		0.067	0.090	0.125		0.056
	(0.029)	(0.039)		(0.042)	(0.098)	(0.108)		(0.137)
Three-way tracking	0.138***	0.219***	0.162***	0.034	0.155***	0.240*	0.210	0.015
	(0.037)	(0.070)	(0.059)	(0.058)	(0.044)	(0.126)	(0.138)	(0.073)
Track Low	0.121***	0.219***	0.124***	0.019	0.135	0.240*	0.145	0.019
	(0.032)	(0.070)	(0.036)	(0.034)	(0.085)	(0.126)	(0.098)	(0.169)
Track Middle	0.042***	-0.023	0.162***	-0.012	0.062	0.089	0.210	-0.113
	(0.011)	(0.022)	(0.059)	(0.025)	(0.072)	(0.075)	(0.138)	(0.119)
Track High	0.063**	0.089***	0.064*	0.034	0.091	0.185	0.072	0.015
	(0.025)	(0.028)	(0.039)	(0.058)	(0.082)	(0.131)	(0.119)	(0.073)
$\bar{y}$	0.000	-0.503	-0.015	0.519	0.000	-0.503	-0.015	0.519
$sd(y)$	1.000	0.921	0.954	0.848	1.000	0.921	0.954	0.848

Note: Using estimates from Table 4 column (5) and TableB1 column (5) respectively.

Table B3. Results on credits including other peer characteristics

Peer Prior *GPA*	(1)	(2)	(3)	(4)	(5)
$\overline{GPA}_{-i}$	0.148***	0.161***	0.141***	0.124**	0.128**
	(0.052)	(0.058)	(0.052)	(0.053)	(0.058)
$SD(GPA_{-i})$	-0.185**	-0.186**	-0.227**	-0.176**	-0.214**
	(0.082)	(0.083)	(0.093)	(0.080)	(0.095)
$\overline{GPA}_{-i} \times SD(GPA_{-i})$	0.343*	0.342*	0.401**	0.317*	0.371**
	(0.190)	(0.190)	(0.183)	(0.181)	(0.177)
GPA_i	0.350***	0.350***	0.349***	0.350***	0.350***
	(0.035)	(0.036)	(0.035)	(0.036)	(0.036)
$GPA_i \times \overline{GPA}_{-i}$	-0.117***	-0.117***	-0.114***	-0.111**	-0.110**
	(0.042)	(0.041)	(0.040)	(0.043)	(0.042)
$GPA_i \times SD(GPA_{-i})$	0.104	0.108	0.092	0.099	0.091
	(0.075)	(0.075)	(0.076)	(0.075)	(0.077)
$GPA_i \times \overline{GPA}_{-i} \times SD(GPA_{-i})$	-0.287**	-0.300**	-0.250*	-0.295**	-0.268*
	(0.138)	(0.148)	(0.138)	(0.137)	(0.151)
$FBoys_{\neg i}$		✓			✓
$\overline{Age}_{\neg i}$			✓		✓
$\overline{App.Order}_{\neg i}$				✓	✓
Randomization controls	✓	✓	✓	✓	✓
Controls	✓	✓	✓	✓	✓
F-tests (p-values)					
Peer variables = equal to (1)		0.999	0.748	0.917	0.756

Note: Columns (1) to (5) each present results from a separate OLS regression. Dependent variable is the number of collected credit points in the first year. The peer variables $\overline{GPA}_{-i}$ and $SD(GPA_{-i})$ are recentered at mean zero. All regressions include randomization controls: a saturated set of own GPA category, advanced math-, and cohort-dummies, interacted with application order. Background control variables are gender, age, and a dummy for professional college. Standard errors clustered at tutorial group level are in parentheses. N = 1,876, N clusters = 48. */**/*** denote significance at a 10/5/1% confidence level.

Table B4. Results on credits excluding different years

		Excluded Cohort		
	(1)	(2)	(3)	(4)
Peer Prior *GPA*	All	2009	2010	2011
$\overline{GPA}_{-i}$	0.148***	0.165**	0.167**	0.137**
	(0.052)	(0.064)	(0.068)	(0.063)
$SD(GPA_{-i})$	-0.185**	-0.173**	-0.156	-0.239*
	(0.082)	(0.078)	(0.111)	(0.118)
$\overline{GPA}_{-i} \times SD(GPA_{-i})$	0.343*	0.612***	0.363	0.108
	(0.190)	(0.154)	(0.283)	(0.240)
GPA_i	0.350***	0.406***	0.321***	0.319***
	(0.035)	(0.037)	(0.046)	(0.046)
$GPA_i \times \overline{GPA}_{-i}$	-0.117***	-0.174***	-0.047	-0.140***
	(0.042)	(0.043)	(0.053)	(0.050)
$GPA_i \times SD(GPA_{-i})$	0.104	0.061	0.097	0.142*
	(0.075)	(0.090)	(0.112)	(0.078)
$GPA_i \times \overline{GPA}_{-i} \times SD(GPA_{-i})$	-0.287**	-0.489***	-0.131	-0.292
	(0.138)	(0.136)	(0.187)	(0.184)
Randomization controls	✓	✓	✓	✓
Controls	✓	✓	✓	✓
F-tests (p-values)				
Peer variables = equal to (1)		0.412	0.228	0.465

Note: Columns (1) to (4) each present results from a separate OLS regression. Dependent variable is the number of collected credit points in the first year. The peer variables $\overline{GPA}_{-i}$ and $SD(GPA_{-i})$ are recentered at mean zero. All regressions include randomization controls: a saturated set of own GPA category, advanced math-, and cohort-dummies, interacted with application order. Background control variables are gender, age, and a dummy for professional college. Standard errors clustered at tutorial group level are in parentheses. */**/*** denote significance at a 10/5/1% confidence level.

Table B5. Number of credits and peer group composition, categorial interactions

	(1)	(2)
$\overline{GPA}_{-i} \times SD_{BELOW}$	0.060	
	(0.070)	
$\overline{GPA}_{-i} \times SD_{ABOVE}$	0.088*	
	(0.052)	
SD_{ABOVE}	-0.067	
	(0.051)	
$SD(GPA_{-i}) \times \overline{GPA}_{BELOW}$		-0.387***
		(0.112)
$SD(GPA_{-i}) \times \overline{GPA}_{ABOVE}$		-0.090
		(0.107)
$\overline{GPA}_{ABOVE}$		0.151***
		(0.055)
Coeffs of continuous peer variable is equal for SD ($\overline{GPA}$) below and above median(p-value)	0.711	0.060*

Note: Columns (1) and (2) each present results from a separate OLS regression. Dependent variable is the number of collected credit points in the first year. The peer variables $\overline{GPA}_{-i}$ and $SD(GPA_{-i})$ are recentered at mean zero. Subscripts $BELOW$ and $ABOVE$ indicate that groups belong to the bottom half or top half of the respective peer variable. Both regressions include randomization controls (a saturated set of own GPA category, advanced math-, and cohort-dummies, interacted with application order) and background control variables (gender, age, and a dummy for professional college). Standard errors clustered at tutorial group level are in parentheses. N = 1,876, N clusters = 48. The p-values at the bottom test whether the coefficients of mean peer GPA (SD of peer GPA) are the same in groups with below and above median value of SD of peer GPA (mean peer GPA). */**/*** denote significance at a 10/5/1% confidence level.

Table B6. Number of credits and peer group composition, own GPA category interactions

Variable	(1)
$\overline{GPA}_{-i} \times GPA_{BELOW}$	0.265**
	(0.109)
$\overline{GPA}_{-i} \times GPA_{ABOVE}$	0.030
	(0.066)
$SD(GPA_{-i}) \times GPA_{BELOW}$	-0.326**
	(0.156)
$SD(GPA_{-i}) \times GPA_{ABOVE}$	0.056
	(0.105)
$\overline{GPA}_{-i} \times SD(GPA_{-i}) \times GPA_{BELOW}$	0.690**
	(0.279)
$\overline{GPA}_{-i} \times SD(GPA_{-i}) \times GPA_{ABOVE}$	-0.236
	(0.236)
Coeffs of peer variables are equal for GPA-categories (p-value)	0.055*

Note: Results are from an OLS regression. Dependent variable is the number of collected credit points in the first year. The peer variables $\overline{GPA}_{-i}$ and $SD(GPA_{-i})$ are recentered at mean zero. Subscripts $BELOW$ and $ABOVE$ indicate that student belongs to the bottom half or top half of own GPA. The regressions includes randomization controls (a saturated set of own GPA category, advanced math-, and cohort-dummies, interacted with application order) and background control variables (gender, age, and a dummy for professional college). Standard errors clustered at tutorial group level are in parentheses. N = 1,876, N clusters = 48. The p-value at the bottom tests whether the coefficients of the three peer variables are the same for students with below and above median own GPA. */**/*** denote significance at a 10/5/1% confidence level.

C Survey questions

Table C1 lists the questions that are included in the surveys that were administrated during the academic years. It reports the scales on which the answers are measured, the years in which the questions were included in the survey, the number of respondents for each specific question and the mean and standard deviation per question. The final column indicates for which variable used in Section 5 each question is used.

Table C1. Survey questions

		Scale	2009	2010/1	N	Mean	S.D.	Variable
1	What grade do you give for the atmosphere in your tutorial group?	1-10	√	√	1272	7.50	1.20	conducive
2	The composition of my tutorial group influences my performance positively	1-6	√	√	1163	3.76	1.19	conducive
3	Sometimes I study together with other students	1-6	√	√	1291	3.12	1.69	interact
4	Sometimes I help other students with exercises or study material in general	1-6	√	√	1309	3.56	1.40	interact
5	Sometimes other students help me with exercises or study material in general	1-6	√	√	1306	3.63	1.41	interact
6	I learn a lot from the students in my tutorial group	1-6		√	956	3.19	1.23	conducive
7	I'm motivated to study	1-6		√	974	4.37	1.22	
8	The students around me are better students than me and that demotivates me	1-6	√	√	1291	2.40	1.36	involved (-)
9	I'm a better student than the students around me and that demotivates me	1-6		√	944	2.09	1.14	involved (-)
10	I learn a lot from the tutorial group teachers	1-6		√	949	4.17	1.12	stimulate
11	The tutorial group meetings are stimulating	1-6		√	955	3.83	1.11	stimulate
12	I frequently ask questions during tutorial group meetings	1-6		√	947	3.14	1.28	involved
13	Other students frequently ask questions during tutorial group meetings	1-6		√	939	3.84	1.08	involved
14	The tutorial group teachers are too fast for me	1-6		√	950	2.74	1.22	fast
15	The tutorial group teachers are too slow for me	1-6		√	943	2.85	1.23	slow
16	The tutorial group teachers spend too much time on simple things	1-6		√	935	3.16	1.28	slow
17	The tutorial group teachers spend too little time on simple things	1-6		√	934	2.60	1.13	fast
18	The tutorial group teachers ask questions to test whether we understand	1-6	√	√	1270	3.79	1.19	stimulate
19	I dislike to ask questions during the tutorial group meetings	1-6	√	√	1271	2.82	1.47	involved (-)
20	The tutorial group teachers often give complicated answers	1-6	√	√	1269	2.97	1.17	fast
21	It is often unquiet during tutorial group meetings making it difficult to concentrate	1-6	√	√	1275	2.61	1.13	involved (-)
22	The tutorial group teachers focus too much on the weak students	1-6	√	√	1238	2.74	1.20	slow
23	The tutorial group teachers provide examples to memorize the material	1-6	√		333	4.23	0.98	
24	The tutorial group teachers challenge us to approach questions in different ways	1-6	√		327	3.49	1.06	
25	The tutorial group teachers discuss material that is not in the textbook	1-6	√		302	3.35	1.19	
26	The tutorial group teachers gave confidence that if I work, I pass the exam	1-6	√		326	4.41	1.07	

Note: Survey questions posed after 3 months (December). All items have an integer scale.