DISCUSSION PAPER SERIES

IZA DP No. 8769

Ability Peer Effects in University: Evidence from a Randomized Experiment

Adam S. Booij
Edwin Leuven
Hessel Oosterbeek

January 2015

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Ability Peer Effects in University: Evidence from a Randomized Experiment

Adam S. Booij
University of Amsterdam, TIER and Tinbergen Institute

Edwin Leuven
University of Oslo, CEPR and IZA

Hessel Oosterbeek
University of Amsterdam, TIER, Tinbergen Institute, CESifo and FLACSO

Discussion Paper No. 8769
January 2015

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 8769
January 2015

ABSTRACT

Ability Peer Effects in University: Evidence from a Randomized Experiment*

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 ability – randomly. The data support a specification in which the group composition is captured by the mean and standard deviation of prior ability and their squares and interactions. Estimates from this specification imply that students of low and medium ability gain on average 0.2 SD units of achievement from switching from ability mixing to three-way tracking. Their dropout rate is reduced by 15 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 Classification: I22, I28

Keywords: peer effects, tracking, post-secondary education, field experiment

Corresponding author:

Edwin Leuven
Department of Economics
University of Oslo
P.O. Box 1095 Blindern
0317 Oslo
Norway
E-mail: edwin.leuven@econ.uio.no

* We gratefully acknowledge valuable comments from Dennis Epple, Erik Plug, and from seminar participants in various places.

1 Introduction

Can we improve student outcomes through ability grouping? The current paper aims to make progress on this question by analyzing data from a randomized evaluation in which the ability composition of tutorial groups for first-year students in economics was manipulated, and students were – conditional on their ability – randomly assigned to these groups. The manipulation of the composition of groups ensures that we can compare different ability groupings even in the presence of endogenous social interactions.

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 (f.e. Hoxby 2000; Carrell et al. 2009; Ammermueller and Pischke 2009; De Giorgi et al. 2012; Feld and Zölitz 2014) and a second, smaller, and more recent literature uses randomized experiments (Duflo et al., 2011; Carrell et al., 2013). While results are highly context dependent, the literature generally finds that peer effects are nonlinear and heterogeneous.[1]

Studies that are based on naturally occurring variation are likely to encounter support problems when translating their estimates into policy recommendations. A compelling illustration of this is provided by Carrell et al. (2013), who investigate how academic performance of freshmen at the US Air Force Academy depends on the ability composition of their peer group. They first estimate peer effects on data with naturally occurring (but non-manipulated) random variation in peer composition. The results suggest that students from the lowest one third of the prior ability distribution would gain from being grouped together with students from the highest one third of the ability distribution. They then conduct a randomized experiment to test this, and find that low-ability students are in fact harmed by the policy that was expected to benefit them.[2]

[1] Studies that document nonlinear and/or heterogenous peer effects include Hoxby (2000), Brodaty and Gurgand (2009), Lavy et al. (2012a), Lavy et al. (2012b), Burke and Sass (2013) and Black et al. (2013).

[2] Moreover, Angrist (2014) points out that peer effect studies based on naturally occurring variation may suffer

The studies that are based on randomized experiments do not encounter support problems when assessing the effect of the particular ability peer configuration they are interested in, but these studies are silent about the effects of alternative peer groupings. More specifically, Carrell et al. (2013) obtain credible estimates of the effects of grouping low-ability and high-ability students together relative to ability mixing, but they have no observations to estimate the effects of, for example, two-way tracking. Likewise, Duflo et al. (2011) present credible estimates of the effects of two-way tracking, but do not know what would happen when Carrell et al.'s low-high grouping would be introduced in their setting.

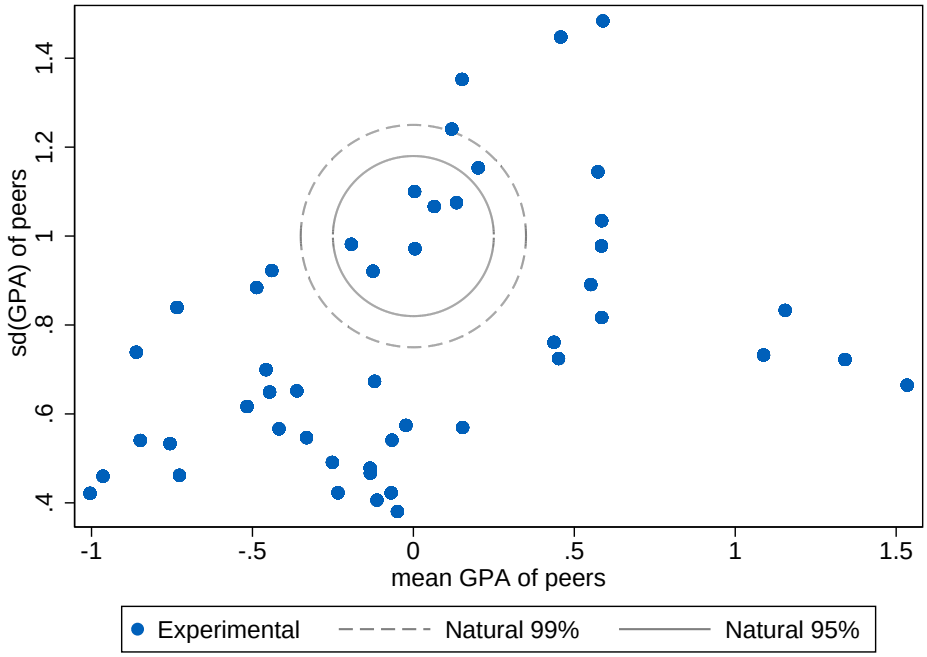
The context of our experiment is the first-year undergraduate program in economics and business at the University of Amsterdam. The around 600 students that enter each year are assigned to tutorial groups of around 40 students. The composition of these groups is fixed for the entire first year, and more than 60 percent 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 prior ability of students across tutorial groups, where ability was defined by students' grade point average on the nationwide final exams of secondary education (GPA). 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 ability 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 of standardized ability in a group, increased from [0.8, 1.2] to [0.3, 1.5].

The large support allows 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. Our estimates show that students benefit from being assigned to groups with more able peers, and also do better in more homogenous groups. These effects are larger for students with lower GPA.

from weak instrument type bias.

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



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

We use our estimates to contrast the predicted students outcomes for different peer group configurations. The results indicate that low-GPA and middle-GPA students would gain on average 0.19 percent of a standard deviation of realized credits from moving from mixing to three-way tracking. Dropout rates go down by around 15 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) 3 percent of a standard deviation and the achievement of middle-GPA students is boosted by 18 percent of a standard deviation. High-GPA students are again unaffected. These findings are qualitatively similar to the results that Carrell et al. obtain in their experiment.

We attribute the effect of the ability composition of tutorial groups to peer effects. A possible confounding factor for this interpretation is that due to the higher dropout rate among low-ability students, the average size of tutorial groups during the year is also affected by the ability composition. Our 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. The results show that group size is not an important factor.

We collected survey data to inform us about the 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 grouped into 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]

[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]In 2009 we drop two groups with late registrations, and both in 2010 and 2011 we drop a group of students

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. This shows that just over 60 percent 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 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 utilized to raise achievement.

Whether students pass a course and the grade they get, is solely determined by the exams that take place at the mid- and/or end-term of the course. The exam of a course is identical for all students and takes 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.

---

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

2.2 Design

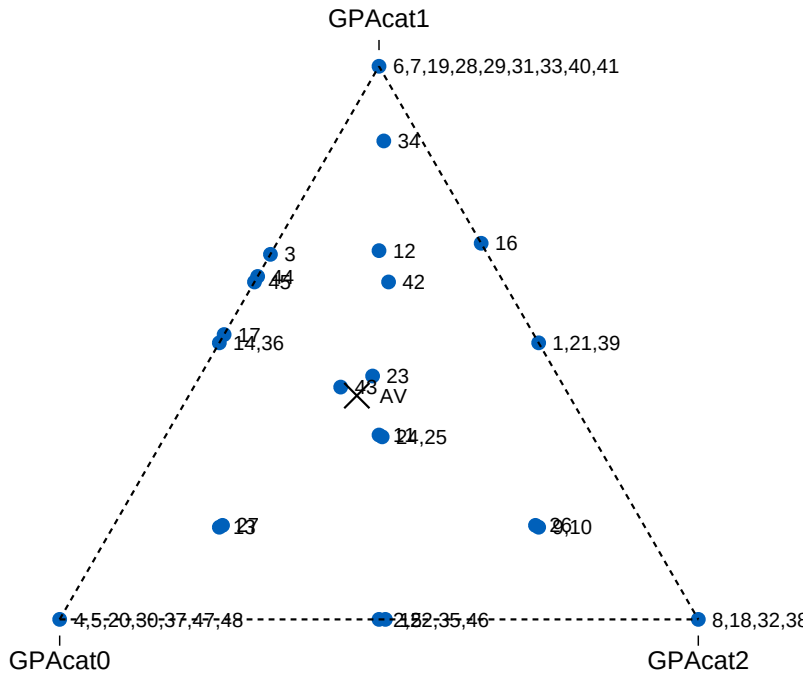
Assignment. To acquire information about the nature of ability peer effects in tutorial groups, we manipulated the ability composition of first-year tutorial groups, and randomly assigned students to these groups (conditional in their ability). As measure of a students' prior ability we use their GPA 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 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.

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	0.27	0.33	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	0.36	0.28	0.25	0.30	0.22	0.26
		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.

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 GPA-category. We aimed at creating large variation in the ability composition of tutorial groups by



Note: Each dot in the triangle represents one tutorial group. The location of a dot in the triangle corresponds to the composition of the tutorial group in terms of the shares of students from categories 0, 1 and 2. Dots in the three corners resemble tutorial groups that consist of students from one category only. Dots on the line segments resemble tutorial groups that consist of two categories of students, and dots 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. Targeted tutorial group composition

covering the complete triangle in Figure 2. Each dot in the triangle resembles a tutorial group as a combination of different *GPAcat* shares. The fraction of each category in a group positions it on the triangle. Points on the vertices of the triangles correspond to groups that consist of only one category, such as *GPAcat*1 in case of the top one. Points on the edges combine the categories of the corresponding vertices. A point on the left edge for example combines only students from *GPAcat* 1 and *GPAcat* 0. Interior points combine all three categories.

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 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 corre-

lation 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] Also, as students with advanced math were traditionally grouped together, we treat tutorial groups with these students separately by including a full set of interaction terms in the regressions. 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 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 (no-shows). Since the decision of these students to not show up cannot have been affected by the composition of the group to which they were assigned, we eliminate these students from our data. For the empirical analyses we construct measures of the ability

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

composition of the peers in a tutorial group based on the students who actually started their study.[8] Figure 1 in the Introduction is based on this information.

Each dot in Figure 1 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, mean standardized GPA per group would, for 95% of the groups, vary between -0.3 and 0.3, while the standard deviation of standardized GPA in a group would, for 95% of the groups, vary between 0.8 and 1.2. The figure clearly shows that the GPA composition of many of the tutorial groups in our design are located outside the circles.

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. These statistics do not vary much across the three cohorts. 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.

Panel B reports summary statistics of students' GPA on the final exams in secondary school, the variable that is the basis for variation in the ability composition of tutorial groups. The high school GPA of students entering the department of economics and business of the University of Amsterdam ranges from 5.45 to 8.62,[10] with an average of about 6.65 and standard deviation

[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.

[9] We also collected additional data through a survey amongst 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

Table 2. Summary statistics

		Cohort					
		2009		2010		2011	
A: Background Characteristics	Range	mean	s.d.	mean	s.d.	mean	s.d.
Male	{0,1}	0.73	0.44	0.73	0.44	0.74	0.44
Age	[16.1, 30.1]	19.4	1.56	19.4	1.60	19.4	1.46
Professional college	{0,1}	0.05	0.22	0.04	0.20	0.04	0.19
Exact high school GPA	[5.45, 8.62]	6.71	0.47	6.65	0.46	6.61	0.47
B: Randomization controls							
Coarse high school GPA							
- GPAcat 0: $GPA < 6\frac{1}{2}$	{0,1}	0.33	0.47	0.40	0.49	0.44	0.50
- GPAcat 1: $6\frac{1}{2} \leq GPA < 7$	{0,1}	0.40	0.49	0.35	0.48	0.36	0.48
- GPAcat 2: $GPA \geq 7$	{0,1}	0.26	0.44	0.24	0.43	0.21	0.41
Advanced math	{0,1}	0.26	0.44	0.37	0.48	0.37	0.48
Application order	[0,1]	0.47	0.29	0.50	0.29	0.50	0.29
C: Treatment variables							
Mean GPA peers	[-1.03, 1.58]	-0.01	0.57	-0.01	0.54	0.01	0.63
SD GPA peers	[0.32, 1.52]	0.81	0.28	0.80	0.28	0.74	0.30
D: Outcome Variables							
Credits (raw)	[0,60]	32.1	22.1	34.7	23.3	32.3	24.8
Grades (raw)	[1.00 9.32]	5.26	1.36	5.57	1.42	5.32	1.71
Dropout	{0,1}	0.55	0.50	0.46	0.50	0.46	0.50
Number of tutorial groups		14		17		17	
Number of students		606		668		602	

Note: The table reports means and standard deviations of main variables by cohort. Panel A: Professional college is dummy equal to one if student entered university through professional college instead of the pre-university track in secondary education. Exact high school GPA is grade point average on high school exit exam; scale from 1 to 10. Panel B: The GPA-categories are dummies indicating if a student belongs to that group. Advanced math is dummy equal to one if student took advanced math in high school zero otherwise. Application order is a student's percentile rank in the application order. Lower application order for students who applied earlier. Panel C: Mean and SD of GPA of peers in tutorial group are the treatment variables. Both are based on standardized exact high school GPA. Panel D: Credits is the number of credits points that a student collects in the first academic year, 60 is the maximum. Grade is the grade point average of the tests that the student did in the first year. Exams that the student missed are not included. Dropout is a dummy variable equal to one if the student collected more than 45 out of 60 credit points in the first year, zero otherwise. The threshold of 45 is required to be allowed to continue.

close to 0.5. 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 of their GPA in secondary 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 write all first-year exams during the first year. The other 54% of the students miss at least one exam, and on average they miss 6.2 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. We refer to this variable as "Dropout". This 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]

from those part of the high school passing criterion.

[11] In the further analyses both the number of credit points and the average grade are standardized to mean zero 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

	Treatment		Outcomes		
	(1) Mean GPA peers	(2) SD GPA peers	(3) Credits	(4) Grade	(5) Dropout
Male	-0.01	0.00	-0.18	-0.10	0.11
	(0.02)	(0.01)	(0.04)***	(0.04)**	(0.02)***
Youngest $\frac{1}{3}$	0.02	-0.01	0.11	0.13	-0.03
	(0.02)	(0.01)	(0.05)**	(0.03)***	(0.02)
Oldest $\frac{1}{3}$	0.00	-0.02	-0.12	-0.04	0.04
	(0.02)	(0.01)	(0.05)**	(0.05)	(0.02)*
Professional college	-0.01	0.01	0.05	0.12	-0.04
	(0.04)	(0.03)	(0.12)	(0.12)	(0.06)
GPA	-0.01	0.00	0.31	0.46	-0.15
	(0.02)	(0.01)	(0.03)***	(0.03)***	(0.02)***
Randomization controls	✓	✓	✓	✓	✓
$\bar{y}$	0.00	0.79	0.00	0.00	0.49
$sd(y)$	0.58	0.29	1.00	1.00	0.50
χ^2-stat coeff.$= 0$		4.16	44.14	58.44	30.69
p-value		0.94	0.00	0.01	0.00
R^2	0.51	0.28	0.27	0.35	0.23
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 GPAcat-, advanced math-, and cohort-dummies, interacted with application order. Robust standard errors are in parentheses in columns (1) and (2). Group clustered standard errors are in parentheses in columns (3) to (5). */**/*** denote significance at a 10/5/1% confidence level.

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 no systematic patterns, as expected (p-value = 0.94). At the same time, columns (3) to (5) show that the 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 (cf. Carrell et al., 2013; Duflo et al., 2011). This allowed the authors to estimate 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 variation in peer configurations in multiple directions, including the treatments considered in the previous experimental studies. To analyze the rich variation that our design generates, we have to impose some structure and need to capture the peer composition of tutorial groups in a limited number of "treatment" variables. 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. Inclusion of the mean of peers' prior ability concurs with the canonical linear-in-means model. It reflects the idea that being surrounded by smarter peers is beneficial. Inclusion of the standard deviation of peers' prior ability is less common. It has, however, recently been rationalized by Tincani (2014a) (see also Tincani, 2014b) in a model in which students care about their rank in the class.[12] The estimation equation is then:

$$y_{ig} = \delta \overline{GPA}_{g-i} + \gamma SD(GPA_{g-i}) + \beta' \mathbf{X}_i + u_g + \varepsilon_{ig} \tag{1}$$

where y_{ig} is the outcome of student i in group g (credits, grade, Dropout), $\overline{GPA}_{g-i}$ and $SD(GPA_{g-i})$ are the mean and the standard deviation of the prior ability of the other students in the group to which student i is assigned. $\mathbf{X}_i$ is a vector of control variables including the randomization controls: a fully saturated set of dummies for each *GPAcat*, type of mathematics in secondary education and cohort, interacted with application order. In addition, we include control variables for gender, age, professional college, and own GPA.

Since our experimental design generates sufficient variation in the peer variables, we extend equation (1) by including higher order terms of the peer variables to examine nonlinearities. To investigate heterogeneous peer effects, we will in addition present results from specifications in

[12] Inclusion of the standard deviation is also implied by social cognitive learning theory which strongly suggests that achievement may benefit from the presence of similar classmates (Bandura, 1986; Schunk, 1991).

which the peer variables are interacted with student's own GPA.

We use the estimates of the preferred specification to simulate the effects of different peer configurations. More specifically, we will 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.

We assess the robustness of our main findings in a number of ways. First, we present results from regressions in which the composition of peers is captured by the shares of low, middle, and high-GPA students in the tutorial group (and their interactions). Second, we present results from regressions in which the composition of peers is captured by the median GPA and interquartile range of GPA of students in the group (and their interactions). Finally, we compare the main findings to the results obtained through nonparametric local linear regression. While the results from the share-based and quartile-based estimations are not entirely in line with the results from the moment-based estimations, the nonparametric results are. This indicates that the moment-based estimations on which the main results are based, are sufficiently flexible to capture the main patterns in the data.

4 Results

The results are presented in four subsections. Subsection 4.1 presents the main results. 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) and Subsection 4.4 assesses the robustness of the main findings.

4.1 Main results

The top part of Table 4 shows results from six increasingly flexible moment-based specifications. The bottom part of the table presents p-values of F-tests. These test the hypotheses that:

1. the coefficients of the peer variables in the respective column are jointly equal to zero: cf(Peer variables)=0

2. the coefficients of the terms that were added in comparison with the previous column are jointly equal to zero: cf(Added terms)=0

3. the coefficients of the nonlinear peer terms are jointly equal to zero: cf(Nonlinear terms)=0

4. the coefficients of the higher order peer variables are jointly equal to zero: cf(Higher order terms)=0

5. the peer variables in columns (5) and (6) are the same for students with different own GPA: cf(Peer variables)=homogenous.

The first column of Table 4 presents results from the basic linear-in-means model where only the mean of peers' GPA and the randomization controls are included. The point estimate equals 0.051, which at face value would imply that a one standard deviation increase of the mean of peers' GPA raises the number of credit points a student collects by 5.1 percent of a standard deviation. The estimate is, however, not significantly different from zero (p=0.242). The second column shows that inclusion of control variables for gender, age and a dummy for professional college has only a minor impact on the estimated coefficient and its standard error.

Column (3) reports results from a specification that includes the mean as well as the standard deviation of peers' GPA in a tutorial group. The coefficient for the mean of peers' GPA is positive and that of the standard deviation of peers' GPA is negative. A higher mean and a smaller dispersion of GPA in a group increase 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=0.222).

Column (4) increases the number of nonlinear terms and adds the squares of the mean and the standard deviation of peers' GPA as well as their interaction. The coefficients of the separate terms of this specification are hard to interpret. The statistics from the F-tests show, however, that in this specification we can reject that the joint effect of the peer variables equals zero (p=0.048). The coefficients of the nonlinear terms are jointly significant at the 5%-level and the coefficients of the higher order terms are jointly significant at the 10%-level.

In column (5) we estimate the same specification as in column (4) but add interactions of the peer variables and own GPA. The F-test for the significance of the added interaction terms,

Table 4. Peer effects on number of credits; various specifications

	(1)	(2)	(3)	(4)	(5)		(6)		
Peer Prior *GPA*					Main	$\times GPA_i$	Main	$\times GPA_i$	$\times GPA_i^2$
$\overline{GPA}_{\neg i}$	0.051	0.048	0.070	0.20	0.21	-0.16	0.24	-0.16	-0.020
	(0.043)	(0.041)	(0.043)	(0.083)**	(0.086)**	(0.081)*	(0.096)**	(0.11)	(0.057)
$sd\left(GPA_{\neg i}\right)-1$			-0.095	-0.12	-0.16	0.089	-0.13	0.20	-0.0055
			(0.073)	(0.13)	(0.14)	(0.086)	(0.14)	(0.13)	(0.075)
$\overline{GPA}_{\neg i}^2$				-0.040	0.031	-0.092	-0.046	-0.039	0.013
				(0.061)	(0.066)	(0.059)	(0.074)	(0.11)	(0.053)
$\left(sd\left(GPA_{\neg i}\right)-1\right)^2$				0.11	0.14	0.32	0.22	0.48	-0.20
				(0.22)	(0.21)	(0.33)	(0.27)	(0.45)	(0.26)
$\overline{GPA}_{\neg i} \times \left(sd\left(GPA_{\neg i}\right)-1\right)$				0.40	0.14	-0.40	0.16	-0.60	0.069
				(0.18)**	(0.22)	(0.21)*	(0.32)	(0.35)*	(0.15)
Randomization controls	✓	✓	✓	✓	✓		✓		
Controls		✓	✓	✓	✓		✓		
$\bar{y}$					0.00				
$sd\left(y\right)$					1.00				
$N_{cluster}$					48				
N					1876				
R^2	0.22	0.27	0.27	0.27	0.27		0.27		
F-tests (p-values)									
cf(Peer variables) = 0	0.242	0.254	0.222	0.048	0.000		0.000		
cf(Added terms) = 0		0.000	0.202	0.065	0.003		0.863		
cf(Nonlinear terms) = 0			0.202	0.026	0.001		0.110		
cf(Higher order terms) = 0				0.065	0.062		0.346		
cf(Peer variables) = homogenous					0.003		0.000		

Note: Columns (1) to (6) each present results from a separate OLS regression. Dependent variable is number of collected credit points in the first year. Main explanatory variables are mean and standard deviation of standardized GPA of tutorial group peers. Column (5) reports estimates of peer effects where peer variables are interacted with own GPA. Column (6) reports estimates of peer effects where peer variables are interacted with own GPA and with own GPA squared. F-tests are reported for null-hypotheses that joint effect of peer variables equals zero (cf(Peer variables) = 0), that terms that added in comparison with previous column have no effect (cf(Added terms) = 0), that joint effect of nonlinear peer variables equals zero (cf(Nonlinear terms) = 0), that the joint effect of the higher order peer variables equals zero (cf(Higher order terms)=0), and that peer effects are the same for different own GPA in columns (5) and (6) (cf(Peer variables) = homogenous). Randomization controls are a saturated set of own GPAcat-, advanced math-, and cohort-dummies, interacted with application order. Other control variables are own GPA, gender, age, and a dummy for professional college. Group clustered standard errors in parentheses. */**/*** denote significance at a 10/5/1% confidence level.

show that these are jointly significant (p=0.003). The data thus reject that ability peer effects are homogenous with respect to students' own GPA. Column (6) takes the interaction effects one step further and allows also for interaction effects of the ability peer variables and the square of own GPA. The F-test reported in the bottom of this column shows that we cannot reject that the coefficients of these higher order interaction terms are jointly equal to zero (p=0.863). We therefore base our further analysis of the results on the specification in column (5).

Figure 3 illustrates the results from column (5) graphically. The top graphs show the relation between the mean of peers' GPA and performance separately for students with below and above-median GPA. The bottom graphs show the relation between the standard deviation of peers' GPA and performance, again separately for students with below and above-median GPA.[13] Comparison of the two graphs for below-median GPA students and the two graphs for above-median GPA students, reveals that it is mainly the performance of below-median GPA students that is affected by the peer group composition. These students benefit from an increase in the mean GPA of their peers, and they are harmed by an increase in the standard deviation of peers' GPA.

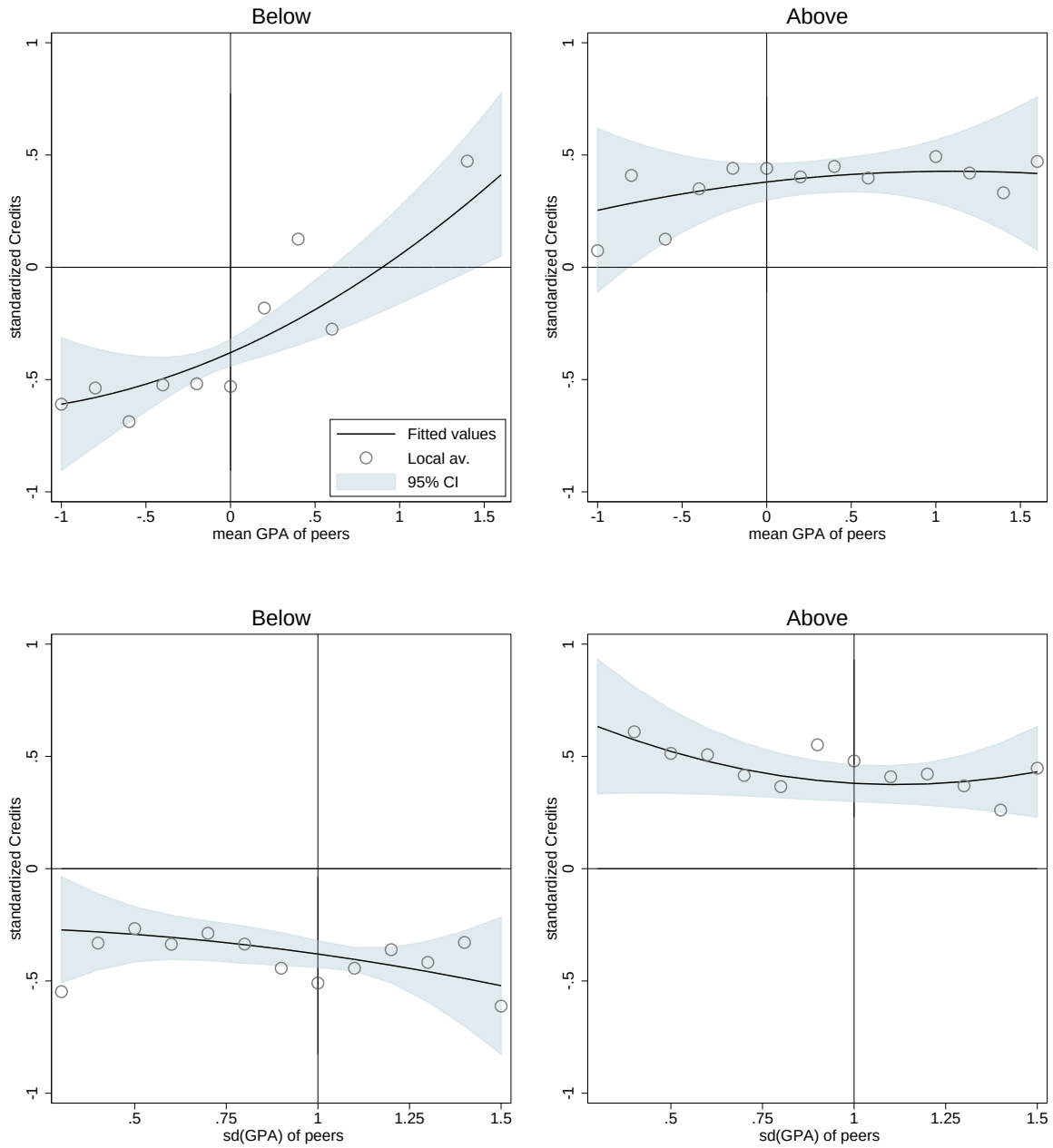
4.2 Alternative assignments

In this subsection we use the results from model (5) in Table 4, to calculate how the predicted number of credit points for each student is affected by a change from the current practice of ability mixing to various alternative assignments. The pattern of results in Figure 3 suggests that it is optimal to place each low-GPA student in a group that otherwise only consists of high-GPA students. This is obviously not feasible because there are too few high-GPA students (or too many low-GPA students) to do so. We therefore only consider alternative assignments that take the current distribution of students' GPA as given. We evaluate five such assignments:[14]

- Two-way tracking, where each group consists only of students with below-median GPA or only of students with above-median GPA;

[13] To construct the graphs, own GPA is set equal to the mean of the below and above median GPA students (-0.77 and +0.77, respectively). In the graphs for mean GPA, the standard deviation of peers' GPA is set equal to 1. In the graphs for the standard deviation of GPA, the mean of peers' GPA is set equal to 0.

[14] For computational ease we take the GPA distribution of the full estimation sample to represent one cohort, and calculate the effect of the alternative assignments neglecting "leave-out". Neglecting "leave-out" should have a negligible impact on the estimates given an average group-size of about 39.



Note: Each dot in the top-left graph is the local average of standardized credits for below-median GPA students per 0.1 bin of mean GPA of tutorial group peers. Top-right graph for above-median GPA students. Each dot in the bottom-left graph is the local average of standardized credits for below-median GPA students per 0.2 bin of the standard deviation of GPA of tutorial group peers. Bottom-right graph for above-median GPA students. The lines are based on the estimates reported in column (5) of Table 4, where in the top graphs the standard deviation of peers' GPA is set equal to one and in the bottom graph the mean of peers' GPA is set equal to zero. Average own GPA below equals -0.77; average own GPA above equals 0.77

Figure 3. Effect of peers' mean GPA and s.d. on first year credits, by student prior ability

- Three-way tracking, where each group consists only of students from the lowest one third of the GPA distribution, or only of students from the middle one third of the GPA distribution, or only of students from the highest one third of the GPA distribution;
- Track Low, where each group consists only of students from the lowest one third of the GPA distribution, or only of students from the highest two thirds of the GPA distribution;
- Track Middle, where each group consists only of students from the middle one third of the GPA distribution, or only of students from the lowest or highest one thirds of the GPA distribution;
- Track High, where each group consists only of students from the lowest two thirds of the GPA distribution, or only of students from the highest one third of the GPA distribution.

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 the same for students in the bottom and top halves of the GPA distribution. The second row shows that a switch to three-way tracking boosts the average achievement gain even further to 15% of a standard deviation. This gain is mainly concentrated by students in the lower two thirds of the GPA distribution, who on average gain 19% of a standard deviation. 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 18% of a standard deviation on middle-GPA students. These findings are similar to the unexpected results that Carrell et al. found in their experiment.

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

		(1)	(2)	(3)	(4)
Tracking		ATE	Below/Low	Middle	Above/High
Two-way tracking	{B},{A}	0.10	0.10		0.10
		(0.03)***	(0.06)*		(0.05)**
Three-way tracking	{L},{M},{H}	0.15	0.20	0.18	0.08
		(0.05)***	(0.09)**	(0.08)**	(0.06)
Track Low	{L},{M,H}	0.13	0.20	0.14	0.04
		(0.03)***	(0.09)**	(0.04)***	(0.05)
Track Middle	{M}, {L,H}	0.05	-0.03	0.18	0.01
		(0.04)	(0.04)	(0.08)**	(0.03)
Track High	{L,M},{H}	0.06	0.06	0.04	0.08
		(0.03)**	(0.05)	(0.04)	(0.06)
$\bar{y}$		0.00	-0.51	0.00	0.50
$sd\,(y)$		1.00	0.92	0.96	0.85

Note: Using estimates from Table 4, specification 5.

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). We use the specifications from columns (4) and (5) of Table 4. Table A3 in the appendix reports the results, where the first two columns repeat the results from columns (4) and (5) of Table 4. The patterns of peer effects appear to be 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). Also the results for the F-tests lead to the same conclusions: ability peer effects are nonlinear and heterogenous.

To better compare the peer effects for different outcome variables, Table 6 reports results from simulations similar to those reported in Table 5. When average grade is the outcome variable, the estimated effects of tracking have almost always the same sign as in Table 5, but fewer effects are significantly different from zero. When Dropout is the outcome measure, results concur very well with those in Table 5. Switching from ability mixing to three-way tracking reduces the dropout rates of low-GPA students by 17 percentage points, relative to an average dropout rate for this group of 0.72. For middle-GPA students the reduction in dropout rates is 13 percentage points, relative to an average dropout rate for this group of 0.49. These are rather substantial reductions in student dropout.

Table 6. Estimated tracking effects compared to mixing

	Outcome variable							
	Av. Grade				Dropout			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Tracking	ATE	B/Low	Middle	T/High	ATE	B/Low	Middle	T/High
Two-way tracking	0.06	0.07		0.06	-0.07	-0.10		-0.04
	(0.03)*	(0.07)		(0.04)	(0.02)***	(0.04)***		(0.02)*
Three-way tracking	0.06	0.13	0.08	-0.02	-0.11	-0.17	-0.13	-0.03
	(0.07)	(0.12)	(0.11)	(0.05)	(0.03)***	(0.06)***	(0.05)***	(0.03)
Track Low	0.10	0.13	0.12	0.04	-0.08	-0.17	-0.04	-0.02
	(0.04)***	(0.12)	(0.03)***	(0.04)	(0.02)***	(0.06)***	(0.02)**	(0.02)
Track Middle	-0.02	-0.07	0.08	-0.06	-0.06	-0.03	-0.13	-0.01
	(0.05)	(0.04)*	(0.11)	(0.03)**	(0.02)***	(0.02)	(0.05)***	(0.02)
Track High	0.02	0.08	0.00	-0.02	-0.04	-0.05	-0.05	-0.03
	(0.03)	(0.06)	(0.04)	(0.05)	(0.02)**	(0.03)*	(0.02)**	(0.03)
$\bar{y}$	0.00	-0.57	-0.08	0.61	0.49	0.72	0.49	0.26
$sd\,(y)$	1.00	0.85	0.88	0.89	0.50	0.45	0.50	0.44

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

To summarize, the results from the peer effects models in Tables 4 and A3 show that in our data peer effects are nonlinear and heterogenous, with low-GPA students benefiting from a higher mean and a lower standard deviation of peers' GPA. High-GPA students are basically unaffected by the composition of their tutorial group. 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. These results hold for different outcome variables and are only somewhat less prominent when average grade is used as outcome. Recall that the number of collected credits and passing the threshold of 45 credit points are arguably the more relevant outcome measures, and average grade the less relevant (and possibly selective).

4.4 Other specifications

In the models reported in Tables 4 and A3, the ability composition of the peer group is expressed in terms of the mean and standard deviation of peers' GPA. In this subsection we assess the robustness of the main findings with regard to three alternative specifications of the main estimation equation. In the first alternative specification the GPA composition of peers is measured by the median and the interquartile range of the GPA of peers in the tutorial group. In the second alternative specification 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. Finally, we compare the main findings with results from nonparametric local linear

regressions.

Table A4 in the appendix reports results from the quantile-based and the share-based specifications.[15] For ease of comparison, column (1) of that table repeats the results from the moments-based specification from column (5) in Table 4. The results from the quantile-based specification in column (2) are very much in line with those from the moment-based specification. The main effects of the median and interquartile range of GPA are positive and negative respectively, while their interactions with own GPA have the opposite signs. This means that low-GPA students benefit from higher median GPA of the peers in their tutorial group and from a smaller interquartile range, and that these effects get smaller (closer to zero) when own GPA increases. The effects of various peer configurations relative to the current practice of mixing are reported in Table A5 in the appendix. Almost all estimated effects obtained from the quantile-based specification have the same sign as those obtained using the moment-based specification, but are typically smaller in size and less likely to differ significantly from zero.

Estimates from the share-based specification are reported in column (3) of Table A4 in the appendix. These estimates are transformed into effects of various peer configurations in columns (9) to (12) of Table A5 in the appendix. Also these effects are smaller and less precise than the effects obtained through the moment-based specification. There is also an important qualitative difference. The results from the moment-based and quartile-based specification point to three-way tracking as the configuration that gives the largest gain in comparison to the current practice. In contrast, the results from the share-based specification indicate that introduction of Track Low is more beneficial and that it is specifically favorable for students from the middle third of the GPA distribution.

The different results from the different specifications are reason for concern. Theory gives no guidance as to which specification to prefer, and also the statistical tests reported in the bottom part of Table A4 give no decisive answer about which specification to prefer (see also Hurder, 2012). We therefore turn to nonparametric results which were obtained using local linear regressions.[16] Columns (4) to (6) of Table 7 report the effects of different peer config-

[15] Shares and quantiles are also calculated using the leave-out approach.

[16] We used a (normalized) tricubic kernel function with a span of 0.70 (flexible bandwidth covering 70% of the data), implemented by the "locfit" package written for R.

Table 7. Comparing parametric (OLS) estimates to nonparametric (local linear regression) estimates

	Parametric			Nonparametric		
	Moment based (1)	Quantile based (2)	Share based (3)	Moment based (4)	Quantile based (5)	Share based (6)
Two-way Tracking	0.10	0.03	0.09	0.14	0.02	0.10
	(0.03)***	(0.02)	(0.06)	(0.07)	(0.06)	(0.08)
Three-way Tracking	0.15	0.06	0.09	0.21	0.03	0.20
	(0.05)***	(0.03)*	(0.07)	(0.11)	(0.07)	(0.14)
Track Low	0.13	0.04	0.18	0.21	0.07	0.19
	(0.03)***	(0.02)**	(0.05)***	(0.08)	(0.06)	(0.13)
Track Middle	0.05	0.00	-0.01	0.16	-0.02	0.14
	(0.04)	(0.05)	(0.09)	(0.11)	(0.10)	(0.20)
Track High	0.06	0.02	0.01	0.08	-0.02	0.07
	(0.03)**	(0.02)	(0.06)	(0.07)	(0.06)	(0.09)

Note: Results in columns (1) to (3) repeat results from Table A3 in the appendix. Results in columns (4) to (6) are based on local linear regressions.

urations implied by these nonparametric results. The local linear regressions used either the mean and standard deviation of peers' GPA (column (4)), or the median and interquartile range (column (5)), or the shares of low-GPA and high-GPA peers (column (6)) as explanatory variables. Columns (1) to (3) repeat the results from Table 5. There are three things to note from this table: i) The results in columns (4) and (6) are strikingly similar. For the nonparametric estimates it does not matter whether peer composition is measured in moments or in shares; ii) The nonparametric results in columns (4) and (6) are much closer to the moment-based parametric estimates in column (1) than to quantile-based and share-based parametric estimates (in columns 2 and 3). (The only exception is the effect of Track Low.) This indicates that the moment-based specification is sufficiently flexible to track the nonparametric results, whereas the share-based and quantile-based models are not; iii) The nonparametric estimates are, not surprisingly, less precise than the parametric estimates. These findings are not only true for the overall effects of different peer configurations, but also for the effects of the subgroups of students from the lowest, middle, and highest one thirds of the GPA distribution. These results are reported in Table A6 in the appendix. Most notably, the results confirm that low-GPA students benefit substantially from three-way tracking.[17]

[17]Tables A7 and A8 in the appendix reports two further robustness checks. In Table A7 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 A7 show that the ability peer estimates are robust to the inclusion of these variables. In Table A8 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. The results indicate that this is not the case.

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.

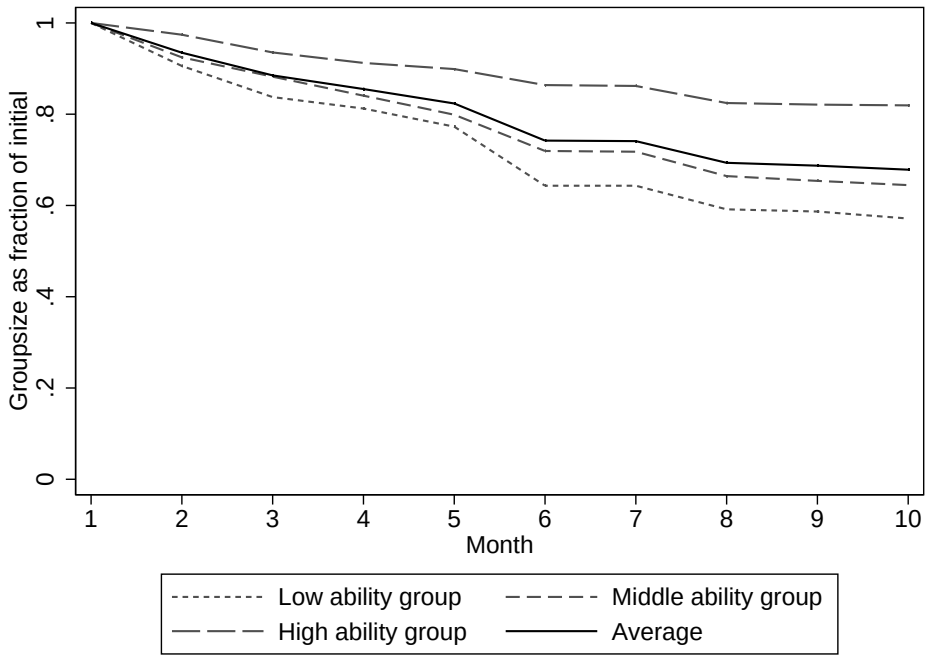
5.1 Group size

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 fewer low-GPA students. This is illustrated in Figure 4 which shows how group size evolves over the year for the 12 groups in our data with the lowest, middle, and highest average GPA at the start of the year. The graph assumes that students dropped out when they stopped writing exams. If the size of tutorial groups has an independent effect on student achievement, part of our findings should be attributed to the group size effect instead of a pure ability peer effect.

To assess the importance 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. Since this average size of a group is potentially an outcome of the ability composition of the group and therefore a "bad" control variable, we instrument it. As instrument for the size of a tutorial group we use 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.[18] A regression of actual average group size during the year on the number of no-shows gives a first stage estimate of -0.70 (s.e. 0.12; F-value 34.6).[19] Because no-shows were never exposed to the intended peers in their tutorial group or informed about them, their decision

[18]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 size effect.

[19]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 dropout is higher than what is observed in the 10th month in Figure 4 because some students fail the re-sits in August.



Note: The graph shows the changes in group size during the first year for the 12 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

not to start the program cannot have been influenced by the GPA composition of their tutorial group.

Table 8 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 simulated effects from switching from ability mixing to ability tracking.

The coefficient of average group size in column (2) cannot be given a causal interpretation. The reason is that the GPA composition of a group is possibly affected by the number of no-shows (the instrument for group size) and the ability peer variables are therefore potentially bad control variables when one is interested in the effect of group size. To address this problem, we instrument the variables that capture the GPA composition of the group at the start of the year (excluding the no-shows) with the corresponding variables based on the GPA composition of the group before the start of the year (including the no-shows). The results are presented

Table 8. Results on credits in the first year controlling for group size

	Baseline OLS		IV		IV			
	(1)		(2)		(3)		(4)	
					Coefficient		F-stat	
Peer Prior *GPA*	Main	$\times GPA_i$	Main	$\times GPA_i$	Main	$\times GPA_i$	Main	$\times GPA_i$
$\overline{GPA}_{\neg i}$	0.21	-0.16	0.16	-0.16	0.12	-0.15	[1361.8]	[1333.0]
	(0.086)**	(0.081)*	(0.13)	(0.082)**	(0.13)	(0.098)		
$sd\left(GPA_{\neg i}\right)-1$	-0.16	0.089	-0.14	0.086	-0.053	0.027	[1396.3]	[1616.9]
	(0.14)	(0.086)	(0.14)	(0.083)	(0.13)	(0.096)		
$\overline{GPA}_{\neg i}^2$	0.031	-0.092	0.044	-0.092	0.061	-0.10	[1640.7]	[1666.6]
	(0.066)	(0.059)	(0.073)	(0.058)	(0.070)	(0.068)		
$\left(sd\left(GPA_{\neg i}\right)-1\right)^2$	0.14	0.32	0.13	0.32	0.27	0.32	[410.0]	[196.1]
	(0.21)	(0.33)	(0.20)	(0.32)	(0.19)	(0.31)		
$\overline{GPA}_{\neg i} \times \left(sd\left(GPA_{\neg i}\right)-1\right)$	0.14	-0.40	0.080	-0.41	0.061	-0.38	[1556.7]	[850.2]
	(0.22)	(0.21)*	(0.25)	(0.20)**	(0.27)	(0.22)*		
Group size			0.0074	[34.6]	0.0069		[35.5]	
			(0.013)		(0.014)			
Randomization controls	✓		✓			✓		
Controls	✓		✓			✓		
R^2	0.27		0.27			0.27		
F-tests (p-values)								
cf(Peer variables) = 0	0.001		0.000		0.000			
cf(Peer var.) = homo.	0.003		0.000		0.002			
Predicted Tracking Effect								
Difference (2Track - Mix)	0.10		0.11		0.09			
s.e.(2Track - Mix)	(0.03)***		(0.03)***		(0.03)***			
Difference (3Track - Mix)	0.15		0.16		0.15			
s.e.(3Track - Mix)	(0.05)***		(0.05)***		(0.05)***			

Note: Group clustered standard errors in parenthesis, and first stage F-statistics in brackets. */**/*** denote significance at a 10/5/1% confidence level. Randomization controls are a saturated set of own GPAcat-, advanced math-, and cohort-dummies, interacted with application order. Other control variables are own GPA, gender, age, and a dummy for professional college.

in column (3) of Table 8. 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 (which can now be given a causal interpretation) is small and not statistically significant, 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 quite similar to those in column (1), although less precise. Jointly, the peer variables in column (3) are statistically significant (p=0.000) and we reject homogeneity of the peer effects (p=0.002). Most importantly, the simulated effects from a switch from ability mixing to ability tracking in columns (3) and (1) are almost identical.

In sum, the higher dropout rates of low-GPA students cause a reduction of the average size of 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. The peer effects that we estimate are therefore 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, and the teaching style of these teachers. The period of three months was chosen 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 A10 in the appendix lists the items that have been 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 percent in all three years. The first column in Table A9 in the appendix shows that response to the surveys is not selective with regard to the ability composition of tutorial groups (p=0.763).

To summarize the information from 26 survey items we constructed six variables which

each are the unweighted sum of three or more items (one item is never used for more than one constructed variable). These sums were normalized to have mean zero and standard deviation one. We label these six variables as follows (in Table A10 in the appendix we indicate which items are used for the construction of which 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: learn a lot from tutorial group teachers, group meetings are stimulating or teacher asks questions to test our understanding;
- Conducive: atmosphere in tutorial group, learn from students in tutorial group, tutorial group influences performance positively;
- Interaction: study together, help other students or are 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 A9 in the appendix reports results from peer effect regressions in which each of the six constructed variables are the dependent variables and in which we use the same specification as in column (5) of Table 4. The results in the bottom rows show that the peer variables are jointly significant for each of the dependent variables. We find for two out of six dependent variables that the peer effects are heterogenous (Too slow and Involved). Table 9 reports results from simulations based on the estimates of Table A9. The top part of the table reports the average change in the dependent variable from a switch from ability mixing to two-way tracking. This

Table 9. Mechanisms

	Teachers			Peers		
	(1)	(2)	(3)	(4)	(5)	(6)
Tracking	Too fast	Too Slow	Stimulating	Conducive	Interaction	Involved
Two-way tracking						
ATE	0.06	0.02	0.03	-0.05	0.11	0.12
	(0.08)	(0.07)	(0.12)	(0.06)	(0.06)*	(0.07)*
- Below	0.12	0.14	0.06	-0.03	0.21	0.23
	(0.12)	(0.11)	(0.18)	(0.09)	(0.10)**	(0.11)**
- Above	0.01	-0.10	-0.00	-0.06	0.01	0.01
	(0.08)	(0.08)	(0.09)	(0.07)	(0.07)	(0.07)
Three-way tracking						
ATE	0.07	0.04	0.00	-0.19	0.12	0.19
	(0.12)	(0.10)	(0.14)	(0.12)	(0.11)	(0.14)
- Low	0.20	0.19	0.07	-0.02	0.30	0.36
	(0.19)	(0.16)	(0.29)	(0.11)	(0.14)**	(0.15)**
- Middle	0.06	0.01	-0.05	-0.29	0.18	0.21
	(0.16)	(0.17)	(0.18)	(0.25)	(0.20)	(0.27)
- High	-0.05	-0.08	-0.01	-0.25	-0.13	0.00
	(0.12)	(0.11)	(0.11)	(0.09)***	(0.08)	(0.09)

Note: Columns (1) to (6) each present results from a separate OLS regression. Dependent variables are constructed variables based on survey. Main explanatory variables are mean and standard deviation of peers' standardized GPA. 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(NL terms) = 0), and that peer effects are the same for different GPA groups in columns (5) and (6) (cf(Peer var.) = homo.). All regressions include a saturated set of own GPAcat-, advanced math-, and cohort-dummies, interacted with application order. Other control variables are own GPA, gender, age, and a dummy for professional college. Group clustered standard errors in parenthesis. */**/*** denote significance at a 10/5/1% confidence level.

is reported for all students together and separately for students with GPA below and above the median GPA. This shows that the (perceived) behavior of teachers is not significantly affected by tracking. This is different for the influence of peers. Students from the bottom half of the GPA distribution experience more positive interaction with the other students in their tutorial group, and they are more involved. The magnitudes of these effects are 21 and 23 percent of a standard deviation of the dependent variable.

The bottom part of Table 9 shows the results from simulations from 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. High-GPA students find the interaction with peers less conducive under tracking than under mixing. Low-GPA students feel more involved in a tracked group than in a mixed group and experience more positive interaction with the other students in their tutorial group.

To summarize, students from the lower end of the GPA distribution have 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

We documented substantial positive effects from ability grouping on the achievement of students from the lower part of the GPA distribution. In terms of credits points these students gain on average 0.2 SD units of achievement from switching from ability mixing to (three-way) ability tracking. The dropout rate of these students is reduced by around 15 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 ef-

fects 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 percent 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 cost, neither in terms of financial expenditures nor in lower achievement for some students who are harmed by the regrouping.

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. (2009). Teacher and peer effects in higher education: Evidence from a French university. Unpublished working paper.

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.

Davidson, R. and MacKinnon, J. G. (1982). Some non-nested hypothesis tests and the relations among them. *Review of Economic Studies*, 49(4):551–565.

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. (2014). On the nature of peer effects in academic achievement. Working Papers in Economics 596, University of Gothenburg.

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

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

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.

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

Tincani, M. M. (2014b). On the nature of social interactions in education: An explanation for recent puzzling evidence. Unpublished working paper.

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

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: Group

Table A3. Results on other outcomes

	Outcome variable								
	Credits			Average grade			Dropout		
	(1)	(2)		(3)	(4)		(5)	(6)	
Peer Prior *GPA*	Main	Main	$\times GPA_i$	Main	Main	$\times GPA_i$	Main	Main	$\times GPA_i$
$\overline{GPA}_{\neg i}$	0.20	0.21	-0.16	0.20	0.20	-0.13	-0.067	-0.061	0.043
	(0.083)**	(0.086)**	(0.081)*	(0.088)**	(0.087)**	(0.058)**	(0.042)	(0.044)	(0.040)
$sd\left(GPA_{\neg i}\right)-1$	-0.12	-0.16	0.089	-0.35	-0.33	0.10	-0.017	-0.014	0.025
	(0.13)	(0.14)	(0.086)	(0.13)***	(0.14)**	(0.075)	(0.065)	(0.071)	(0.041)
$\overline{GPA}^2_{\neg i}$	-0.040	0.031	-0.092	-0.13	-0.090	-0.062	-0.012	-0.045	0.056
	(0.061)	(0.066)	(0.059)	(0.059)**	(0.059)	(0.052)	(0.029)	(0.032)	(0.025)**
$\left(sd\left(GPA_{\neg i}\right)-1\right)^2$	0.11	0.14	0.32	-0.33	-0.29	0.016	-0.20	-0.25	-0.045
	(0.22)	(0.21)	(0.33)	(0.28)	(0.27)	(0.30)	(0.11)*	(0.10)**	(0.15)
$\overline{GPA}_{\neg i} \times \left(sd\left(GPA_{\neg i}\right)-1\right)$	0.40	0.14	-0.40	0.39	0.13	-0.37	-0.19	-0.053	0.11
	(0.18)**	(0.22)	(0.21)*	(0.24)	(0.27)	(0.19)*	(0.092)**	(0.11)	(0.11)
Randomization controls		✓			✓			✓	
Controls		✓			✓			✓	
$\bar{y}$		0.00			0.00			0.49	
$sd\left(y\right)$		1.00			1.00			0.50	
$N_{cluster}$		1876			1753			1876	
N		48			48			48	
R^2	0.27	0.27		0.35	0.35		0.24	0.24	
F-tests (p-values)									
cf(Peer variables) = 0	0.048	0.000		0.091	0.000		0.036	0.000	
cf(Added terms) = 0		0.003			0.008			0.016	
cf(Non-lin. terms) = 0	0.026	0.001		0.054	0.004		0.021	0.003	
cf(Order terms) = 0		0.062			0.099			0.003	
cf(Peer var.) = homo.		0.003			0.008			0.008	

Note: Column (1) repeats column (5) 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. Quit early equals one if the student deregisters in or before February in the first academic year. Continue equals one if the student collected more than 45 credits points in the first year, allowing the student to continue in the second year. 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(NL terms) = 0), and that peer effects are the same for different GPA groups in (cf(Peer var.) = homo.). Bottom rows report results from comparison of peer effects from tracking versus mixing. Randomization controls are a saturated set of own GPAcat-, advanced math-, and cohort-dummies, interacted with application order. Other control variables are own GPA, gender, age, and a dummy for professional college. Group clustered standard errors in parenthesis. */**/*** denote significance at a 10/5/1% confidence level.

Table A4. Results on credits of other peer statistics

Moment based			Quantile based			Share based		
(1)			(2)			(3)		
Peer Prior *GPA*	Main	$\times GPA_i$	Peer Prior *GPA*	Main	$\times GPA_i$	Peer Prior *GPA*	Main	$\times GPA_i$
$\overline{GPA}_{\neg i}$	0.21	-0.16	$MED_{\neg i}$	0.13	-0.13	$FB_{\neg i} - \frac{1}{3}$	-0.41	-0.011
	(0.086)**	(0.081)*		(0.050)**	(0.045)***		(0.19)**	(0.17)
$sd\left(GPA_{\neg i}\right) - 1$	-0.16	0.089	$IQR_{\neg i} - 1.35$	-0.070	0.021	$FT_{\neg i} - \frac{1}{3}$	0.0030	-0.16
	(0.14)	(0.086)		(0.060)	(0.047)		(0.20)	(0.21)
$\overline{GPA}^2_{\neg i}$	0.031	-0.092	$MED^2_{\neg i}$	0.031	-0.070	$\left(FB_{\neg i} - \frac{1}{3}\right)^2$	0.71	1.27
	(0.066)	(0.059)		(0.059)	(0.053)		(0.63)	(1.03)
$\left(sd\left(GPA_{\neg i}\right) - 1\right)^2$	0.14	0.32	$\left(IQR_{\neg i} - 1.35\right)^2$	0.045	0.077	$\left(FT_{\neg i} - \frac{1}{3}\right)^2$	0.017	0.99
	(0.21)	(0.33)		(0.064)	(0.094)		(0.50)	(0.78)
$\overline{GPA}_{\neg i} \times \left(sd\left(GPA_{\neg i}\right) - 1\right)$	0.14	-0.40	$MED_{\neg i} \times \left(IQR_{\neg i} - 1.35\right)$	-0.059	-0.15	$\left(FB_{\neg i} - \frac{1}{3}\right) \times \left(FT_{\neg i} - \frac{1}{3}\right)$	-0.43	3.01
	(0.22)	(0.21)*		(0.10)	(0.066)**		(0.90)	(1.50)*
Randomization controls	✓			✓			✓	
Controls	✓			✓			✓	
R^2	0.272			0.270			0.272	
AICc	4822.9			4825.8			4821.2	
Cross-Validation MRSE	0.892			0.893			0.890	
J-test (p-value) prediction from (1)				0.005			0.048	
Res-test (p-value) using model (1)				0.733			0.462	
J-test (p-value) prediction from (2)	0.538						0.067	
Res-test (p-value) using model (2)	0.866						0.032	
J-test (p-value) prediction from (3)	0.003			0.000				
Res-test (p-value) using model (3)	0.732			0.506				
F-tests (p-values)								
cf(Peer variables) = 0	0.000			0.000			0.000	
cf(Order terms) = 0	0.062			0.230			0.000	
cf(Peer var.) = homo.	0.003			0.002			0.002	
Predicted Tracking Effect								
Difference (2Track - Mix)	0.10			0.03			0.09	
s.e.(2Track - Mix)	(0.03)***			(0.02)			(0.06)	
Difference (3Track - Mix)	0.15			0.06			0.09	
s.e.(3Track - Mix)	(0.05)***			(0.03)*			(0.07)	

Note: "Cross-Validation RMSE" is the root mean squared error for leave-one-group-out cross-validation. The Davidson and MacKinnon (1982) J-test uses the predicted values of the alternative model as covariate, and the Res-test gives the p-value of the F-statistic from the regression of the residual on the alternative model. Group clustered standard errors in parenthesis. */**/*** denote significance at a 10/5/1% confidence level.

Table A5. Estimated tracking effects compared to mixing, based on OLS estimates

	Peer statistic											
	Moment based				Quantile based				Share based			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Tracking	ATE	B/Low	Middle	T/High	ATE	B/Low	Middle	T/High	ATE	B/Low	Middle	T/High
Two-way tracking	0.10	0.10		0.10	0.03	-0.01		0.07	0.09	0.02		0.16
	(0.03)***	(0.06)*		(0.05)**	(0.02)	(0.04)		(0.03)**	(0.06)	(0.06)		(0.09)*
Three-way tracking	0.15	0.20	0.18	0.08	0.06	0.01	0.10	0.08	0.09	0.09	0.08	0.10
	(0.05)***	(0.09)**	(0.08)**	(0.06)	(0.03)*	(0.06)	(0.06)*	(0.05)	(0.07)	(0.11)	(0.14)	(0.08)
Track Low	0.13	0.20	0.14	0.04	0.04	0.01	0.08	0.03	0.18	0.09	0.24	0.21
	(0.03)***	(0.09)**	(0.04)***	(0.05)	(0.02)**	(0.06)	(0.02)***	(0.03)	(0.06)***	(0.11)	(0.06)***	(0.14)
Track Middle	0.05	-0.03	0.18	0.01	0.00	-0.16	0.10	0.08	-0.01	-0.18	0.08	0.08
	(0.04)	(0.04)	(0.08)**	(0.03)	(0.05)	(0.11)	(0.06)*	(0.07)	(0.10)	(0.11)	(0.14)	(0.16)
Track High	0.06	0.06	0.04	0.08	0.02	-0.01	-0.01	0.08	0.01	-0.05	-0.05	0.10
	(0.03)**	(0.05)	(0.04)	(0.06)	(0.02)	(0.03)	(0.03)	(0.05)†	(0.06)	(0.09)	(0.10)	(0.08)
$\bar{y}$	0.00	-0.57	-0.08	0.61	0.00	-0.57	-0.08	0.61	0.00	-0.57	-0.08	0.61
$sd\,(y)$	1.00	0.85	0.88	0.89	1.00	0.85	0.88	0.89	1.00	0.85	0.88	0.89

Note: Using estimates from Table A4.

Table A6. Estimated tracking effects compared to mixing, based on local linear regressions

	Peer statistic											
	Moment Based				Quantile Based				Share Based			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Tracking	ATE	Low	Middle	High	ATE	Low	Middle	High	ATE	Low	Middle	High
Two-way tracking	0.15	0.25	-0.25	-0.09	0.02	0.15	-0.10	-0.01	0.10	0.25	-0.04	-0.08
	(0.07)**	(0.17)	(0.34)	(0.13)	(0.06)	(0.12)	(0.15)	(0.11)	(0.08)	(0.18)	(0.19)	(0.19)
Three-way tracking	0.21	0.28	0.40	-0.06	0.03	0.10	-0.11	-0.00	0.20	0.46	0.34	-0.06
	(0.11)**	(0.24)	(0.45)	(0.13)	(0.07)	(0.14)	(0.14)	(0.13)	(0.14)	(0.24)	(0.60)	(0.13)
Track Low	0.21	0.41	-0.14	-0.17	0.07	0.16	-0.16	-0.37	0.19	0.12	0.34	-0.82
	(0.08)***	(0.20)**	(0.74)	(0.61)	(0.06)	(0.10)	(0.32)	(0.32)	(0.13)	(0.49)	(0.49)	(0.67)
Track Middle	0.16	0.12	0.40	-0.01	-0.02	0.09	-0.11	-0.20	0.14	0.13	0.34	-0.06
	(0.11)	(0.09)	(0.45)	(0.62)	(0.10)	(0.31)	(0.14)	(0.40)	(0.20)	(0.33)	(0.60)	(1.38)
Track High	0.08	0.27	-1.64	-0.06	-0.02	0.21	-0.41	-0.00	0.07	0.46	-0.53	-0.06
	(0.07)	(0.13)**	(1.26)	(0.13)	(0.06)	(0.13)	(0.57)	(0.13)	(0.09)	(0.31)	(0.80)	(0.13)
$\bar{y}$	0.00	-0.57	-0.08	0.61	0.00	-0.57	-0.08	0.61	0.00	-0.57	-0.08	0.61
$sd\,(y)$	1.00	0.85	0.88	0.89	1.00	0.85	0.88	0.89	1.00	0.85	0.88	0.89

Note: The results in this table are obtained from local linear regressions using a (normalized) tricubic kernel function with a span of 0.70 (flexible bandwidth covering 70% of the data).

Table A7. Results on credits including other peer characteristics

Peer Prior *GPA*	(1)	(2)	(3)	(4)	(5)
$\overline{GPA}_{\neg i}$	0.21	0.22	0.21	0.14	0.16
	(0.09)**	(0.09)**	(0.09)**	(0.09)	(0.10)
$sd\left(GPA_{\neg i}\right)-1$	-0.16	-0.16	-0.20	-0.12	-0.15
	(0.14)	(0.14)	(0.15)	(0.14)	(0.16)
$\overline{GPA}_{\neg i}^2$	0.03	0.03	0.03	0.07	0.06
	(0.07)	(0.07)	(0.07)	(0.07)	(0.08)
$\left(sd\left(GPA_{\neg i}\right)-1\right)^2$	0.14	0.13	0.15	0.19	0.18
	(0.21)	(0.22)	(0.21)	(0.22)	(0.22)
$\overline{GPA}_{\neg i} \times \left(sd\left(GPA_{\neg i}\right)-1\right)$	0.14	0.15	0.20	0.02	0.08
	(0.22)	(0.23)	(0.23)	(0.24)	(0.27)
$\times GPA_i$ interactions	✓	✓	✓	✓	✓
$FBoys_{\neg i}$		✓			✓
$\overline{Age}_{\neg i}$			✓		✓
$\overline{App.Order}_{\neg i}$				✓	✓
Randomization controls	✓	✓	✓	✓	✓
Controls	✓	✓	✓	✓	✓
R^2	0.27	0.27	0.27	0.27	0.27
$N_{cluster}$	48	48	48	48	48
N	1876	1876	1876	1876	1876
F-tests (p-values)					
cf(Peer variables) = equal to (1)		1.000	0.977	0.899	0.896

Note: Group clustered standard errors in parenthesis. */**/*** denote significance at a 10/5/1% confidence level.

Table A8. Results on credits excluding different years

		Excluded Cohort		
	(1)	(2)	(3)	(4)
Peer Prior *GPA*	All	2009	2010	2011
$\overline{GPA}_{\neg i}$	0.21	0.24	0.24	0.18
	(0.09)**	(0.11)**	(0.13)*	(0.09)*
$sd\left(GPA_{\neg i}\right)-1$	-0.16	-0.08	-0.14	-0.27
	(0.14)	(0.17)	(0.21)	(0.17)
$\overline{GPA}_{\neg i}^2$	0.03	0.08	0.04	-0.03
	(0.07)	(0.10)	(0.09)	(0.07)
$\left(sd\left(GPA_{\neg i}\right)-1\right)^2$	0.14	0.18	0.11	0.09
	(0.21)	(0.23)	(0.31)	(0.28)
$\overline{GPA}_{\neg i} \times \left(sd\left(GPA_{\neg i}\right)-1\right)$	0.14	0.38	0.15	0.01
	(0.22)	(0.28)	(0.29)	(0.29)
$\times GPA_i$ interactions	✓	✓	✓	✓
Randomization controls	✓	✓	✓	✓
Controls	✓	✓	✓	✓
R^2	0.27	0.29	0.26	0.27
$N_{cluster}$	48	34	31	31
N	1876	1270	1208	1274
F-tests (p-values)				
cf(Peer variables) = equal to (1)		0.596	0.614	0.673

Note: Group clustered standard errors in parenthesis. */**/*** denote significance at a 10/5/1% confidence level.

Table A9. Mechanisms

	Responded	Teachers						Peers					
	(1)	(2)		(2)		(3)		(4)		(5)		(6)	
		Too fast		Too Slow		Stimulating		Conducive		Interact		Involved	
Peer Prior *GPA*		Main	$\times GPA_i$	Main	$\times GPA_i$	Main	$\times GPA_i$	Main	$\times GPA_i$	Main	$\times GPA_i$	Main	$\times GPA_i$
$\overline{GPA}_{\neg i}$	-0.03	0.03	0.10	0.07	-0.22	-0.10	-0.19	0.20	0.09	0.34	-0.02	0.06	-0.26
	(0.09)	(0.18)	(0.11)	(0.13)	(0.16)	(0.16)	(0.16)	(0.17)	(0.12)	(0.17)**	(0.13)	(0.20)	(0.12)**
$sd\,(GPA_{\neg i}) - 1$	-0.05	-0.05	-0.05	-0.10	0.35	-0.35	0.10	-0.30	-0.01	-0.39	0.10	-0.13	0.08
	(0.14)	(0.24)	(0.15)	(0.17)	(0.18)*	(0.31)	(0.20)	(0.25)	(0.13)	(0.22)*	(0.14)	(0.25)	(0.13)
$\overline{GPA}^2_{\neg i}$	-0.004	-0.03	-0.09	-0.13	0.20	0.14	-0.05	-0.27	-0.04	-0.31	0.05	0.11	0.016
	(0.06)	(0.13)	(0.09)	(0.08)*	(0.09)**	(0.08)*	(0.09)	(0.09)***	(0.09)	(0.10)***	(0.10)	(0.08)	(0.08)
$(sd\,(GPA_{\neg i}) - 1)^2$	-0.04	0.02	-0.17	-0.20	0.26	-0.49	0.34	-0.96	-0.19	-0.31	-0.13	0.13	0.02
	(0.29)	(0.51)	(0.32)	(0.39)	(0.47)	(0.54)	(0.51)	(0.65)	(0.28)	(0.49)	(0.34)	(0.65)	(0.41)
$\overline{GPA}_{\neg i} \times (sd\,(GPA_{\neg i}) - 1)$	-0.14	0.09	0.11	0.46	-0.60	-0.27	-0.33	0.69	-0.02	0.90	-0.25	0.19	-0.40
	(0.24)	(0.52)	(0.29)	(0.42)	(0.37)	(0.54)	(0.44)	(0.55)	(0.28)	(0.48)*	(0.28)	(0.54)	(0.30)
Randomization controls	✓	✓		✓		✓		✓		✓		✓	
Controls	✓	✓		✓		✓		✓		✓		✓	
$\bar{y}$	0.72	0.00		0.00		0.00		0.00		0.00		0.00	
$sd\,(y)$	0.45	1.00		1.00		1.00		1.00		1.00		1.00	
$N_{cluster}$	48	47		47		47		47		47		47	
N	1876	1342		1342		1342		1342		1342		1342	
R^2	0.13	0.05		0.10		0.03		0.09		0.08		0.05	
F-tests (p-values)													
cf(Peer variables) = 0	0.763	0.027		0.040		0.017		0.003		0.020		0.003	
cf(Peer var.) = homo.		0.742		0.078		0.292		0.496		0.562		0.041	

Note: Columns (1) to (6) each present results from a separate OLS regression. Dependent variables are constructed variables based on survey. Main explanatory variables are mean and standard deviation of standardized GPA of peers in tutorial group. 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(NL terms) = 0), and that peer effects are the same for different GPA groups in columns (5) and (6) (cf(Peer var.) = homo.). Randomization controls are a saturated set of own GPAcat-, advanced math-, and cohort-dummies, interacted with application order. Other control variables are own GPA, gender, age, and a dummy for professional college. Group clustered standard errors in parenthesis. */**/*** denote significance at a 10/5/1% confidence level.

Table A10. 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.