DISCUSSION PAPER SERIES

IZA DP No. 9448

Understanding Peer Effects: On the Nature, Estimation and Channels of Peer Effects

Jan Feld
Ulf Zölitz

October 2015

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Understanding Peer Effects: On the Nature, Estimation and Channels of Peer Effects

Jan Feld
Victoria University of Wellington

Ulf Zölitz
IZA and Maastricht University

Discussion Paper No. 9448
October 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. 9448
October 2015

ABSTRACT

Understanding Peer Effects: On the Nature, Estimation and Channels of Peer Effects*

This paper estimates peer effects in a university context where students are randomly assigned to sections. While students benefit from better peers on average, low-achieving students are harmed by high-achieving peers. Analyzing students' course evaluations suggests that peer effects are driven by improved group interaction rather than adjustments in teachers' behavior or students' effort. We further show, building on Angrist (2014), that classical measurement error in a setting where group assignment is systematic can lead to substantial overestimation of peer effects. With random assignment, as is the case in our setting, estimates are only attenuated.

JEL Classification: I21, I24, J24

Keywords: peer effects, higher education, measurement error, estimation bias

Corresponding author:

Ulf Zölitz
Institute for the Study of Labor (IZA)
P.O. Box 7240
53072 Bonn
Germany
E-mail: zoelitz@iza.org

* The authors would like to thank two anonymous referees for providing valuable comments and suggestions. We would also like to thank Joshua Angrist, Sandra Black, Lex Borghans, Harold Cuffe, Gigi Foster, Andries de Grip, Monique de Haan, Thomas Dohmen, Andreas Dzemski, David Figlio, Bart Golsteyn, Jonathan Guryan, Daniel Hamermesh, Randi Hjalmarsson, Olivier Marie, Julie Moschion, Derek Stemple, Benedikt Vogt, participants at various seminars and conferences and especially Nicolás Salamanca for helpful discussions and comments. We further thank Joël Castermans, Sanne Klasen and Kim Schippers from the SBE Scheduling Department, Sylvie Kersten from the SBE Exams Office, and Jeannette Hommes and Paul Jacobs from the Educational Research and Development Department for providing data and valuable background information. We thank Sophia Wagner for providing research assistance.

1 Introduction

The promise of the peer effects literature is to provide policy makers with advice that can be used to increase overall performance by simply reorganizing peer groups. When looking at the by now substantial number of published articles that estimate peer effects in education, it becomes apparent that the literature has not yet delivered this promise. This can be seen, for example, in the recent review by Sacerdote (2011), who shows that size and even the sign of peer effects estimates notably differ between and even within primary, secondary and post-secondary education.

One potential reason why peer effects estimates are so varied is that there are a number of social and statistical forces that lead to similar outcomes between peers, even in the absence of causal peer effects (Angrist, 2014; Manski, 1993). Two well-known challenges to the identification of peer effects are the selection and reflection problem. The selection problem states that peer groups are usually formed endogenously and that it is empirically difficult to distinguish peer effects from selection effects. The reflection problem states that it is impossible to distinguish the effect of peers on the individual from the effect of the individual on peers if both are determined simultaneously. A number of recent peer effects studies (Carrell, Fullerton, & West, 2009; Carrell, Sacerdote, & West, 2013; Duflo, Dupas, & Kremera, 2011; Lyle, 2007) have convincingly addressed the selection and reflection problems by studying peer effects in a setting where students are exogenously assigned to peer groups and by using pre-treatment characteristics as measures for peer ability. Even these estimates, however, might be biased due to a mechanical relationship between the measures of own and peer ability as described in Angrist (2014). The nature of this bias is still poorly understood.

In the analytical part of this paper, we build on Angrist (2014) to analyze the role of measurement error in the estimation of peer effects. We show that classical measurement error —

which is usually associated with attenuation bias — can lead to *overestimation* of peer effects. We find that the size and direction of this bias depends on the true underlying peer effect and the group assignment mechanism. We show, both analytically and using Monte Carlo simulations, that in settings where peers are randomly assigned, classical measurement error will only lead to attenuation bias. With systematic student assignment, however, measurement error can lead to substantial overestimation of peer effects. This bias is distinct from selection bias.

Besides estimation bias, the large heterogeneity in peer effect estimates might reflect that peer effects are generated by mechanisms that are highly context specific. Take, for example, adjustment of teachers' behavior as one potential channel of peer effects. While Duflo et al. (2011) have shown that this channel matters in Kenyan classrooms, it is irrelevant in the settings where peer groups have no common teachers like in living communities (Carrell et al., 2009; Carrell et al., 2013; Lyle, 2007; Sacerdote, 2001; Zimmerman, 2003). A better understanding of the channels that drive peer effects could lead to a better understanding about why they differ across contexts. So far, there exists only limited empirical evidence on the channels of peer effects.[1]

In the empirical part of this paper we estimate peer effects in academic achievement in a setting where peers are randomly assigned to sections at the university level. The data consists of all students enrolled at the School of Business and Economics (SBE) at Maastricht University and their grades over a period of three years, which amounts to 7,672 students and 39,813 grades. Course participants are assigned to sections, groups of 10 to 15 students, which spend most of their contact hours together in one classroom. Our measure of student performance is course grades. Following the standard approach in the literature, to avoid the reflection problem, we use a pre-

[1] See, for example, Duflo et al. (2011) for evidence on channels in primary education, Lavy, Paserman and Schlosser (2012) in secondary education and Booij, Leuven and Osterbeek (2015) in post-secondary education.

treatment indicator of peer quality: the past GPA of the peers. To identify potential non-linearities in peer effects, we estimate heterogeneous effects in terms of student and peer achievements. Finally, we use data on students' individual level course evaluation to shed some light on which channels might be driving the observed peer effects in our setting.

Our results for the linear-in-means specification show that being assigned to a section with, on average, higher-achieving peers increases students' grades in that course by a statistically significant but small amount. A one standard deviation increase in the average peer GPA causes an increase of 1.26 percent of a standard deviation in student grades. This result, however, masks important heterogeneity: low-achieving students are actually harmed by high-achieving peers. Analyzing students' course evaluations, we find suggestive evidence that the main channel of the observed peer effects is improved group interaction. We find no evidence for an adjustment in teachers' behavior or student effort driven by the section peer composition.

Taken together this article makes three main contributions. First, our discussion of the role of measurement error in the estimation of peer effects sheds light on a potential threat to the identification of peer effects that has so far been poorly understood. Second, we provide clean estimates of peer effects using a large dataset of randomly assigned students. Third, we are among the first to provide evidence of the underlying channels of peer effects using students' course evaluations.

The remainder of the paper is structured as follows. Section 2 summarizes the existing literature. Section 3 builds on Angrist (2014) to analyze how measurement error biases the estimation of peer effects. Section 4 describes the institutional environment studied and the assignment procedure of students into sections. Section 5 discusses the dataset. Section 6 provides evidence that the assignment to sections is random, conditional on scheduling constraints. Section

7 discusses the empirical strategy and the results. Section 8 investigates underlying channels using students' course evaluations. Section 9 concludes.

2 Literature Review

The size and nature of peer effects estimates across the vast literature are quite varied and appear highly context specific. In his review, Sacerdote (2011) shows that peer effects estimates at the mean vary from a decrease of 0.12 points (Vigdor & Nechyba, 2007) to an increase of 6.8 points (C. Hoxby, 2000) for a 1.0 point increase in the average test score of the peers in primary and secondary education.[2] Estimates of non-linear effects in these studies also differ substantially. While there is some evidence of high achieving students benefiting from high achieving peers, overall there is no reliable pattern across studies (see Sacerdote (2011) and studies cited within).

In the context of post-secondary education, a number of studies address the selection problem by exploiting exogenous assignment of students to sections (Booij et al., 2015; De Giorgi, Pellizzari, & Woolston, 2012), dorm rooms (e.g., Brunello, De Paola, & Scoppa, 2010; Sacerdote, 2001; Zimmerman, 2003) and living communities in military colleges (Carrell et al., 2009; Carrell et al., 2013; Lyle, 2007). Linear-in-mean point estimates in these environments are typically small and positive or statistically insignificant.

The inconsistencies of the peer effects literature have been attributed to endogeneity in peer group formation (Brock & Durlauf, 2001; Lin, 2010) and estimation bias (Angrist, 2014). We think that there are at least two other reasons why we might expect results to differ between studies: First, peer effects estimates are often presented in terms of standard deviations in a given sample, and

[2] See also Epple and Romano (2011) and Sacerdote (2014) for recent reviews.

estimates of non-linear effects usually rely on dividing students into achievement quantiles (Sacerdote, 2014). These results are not comparable across studies because reported peer effect estimates might actually reflect differences in achievement distributions rather than differences in peer effects. Second, peer effects can run through a number of potential channels which are likely to generate different peer effects depending on the contexts in which they are studied. While peer effects in primary and secondary education might partly run through classroom disruptions (Figlio, 2007; Lazear, 2001), these channels might be less pronounced in post-secondary education. Moreover, teachers' adjustments of pedagogical practice to the classroom composition might be an important channel for classroom settings (Duflo et al., 2011; Lavy et al., 2012), but this channel will not matter when studying peer effects in roommate settings. For living communities where peers spend a great deal of time together, we would expect student effort, i.e., adjustments in (joint) study hours, to be an important channel (Stinebrickner & Stinebrickner, 2006).

3 Measurement Error and the Estimation of Peer Effects

In a recent overview, Angrist (2014) discusses many threats to the identification of peer effects. In particular, he shows, building on earlier work by Acemoglu and Angrist (2001), that measurement error can lead to *overestimation* of peer effects.[3] This might seem counterintuitive since measurement error is usually associated with attenuation bias. The key problem in this context is that measurement error in own ability will automatically lead to measurement error in measured group ability because it is an aggregate of the individual ability measures. Since ability will always be measured with some error, a typical peer effects regression will thus contain *two* mismeasured

[3] See also Moffitt (2001) and Ammermueller and Pischke (2009) for a discussion on the role of measurement error in the estimation of peer effects.

independent variables, which makes the direction of the bias unclear. We phrase the discussion below in terms of the estimation of peer effects, but it generalizes to other settings where one independent variable is the group average of another mismeasured independent variable.[4] In these settings measurement error can lead to an upward bias in the group average coefficient.

We believe Angrist has uncovered an important source of bias that deserves further investigation since he does not explicitly show under which assignment mechanisms an upward bias exists or how the magnitude of this bias depends on the underlying parameters. In this section we will therefore first review Angrist's decomposition of the peer effects coefficient. Then, expanding upon this, we will show how classical measurement error can lead to overestimation of peer effects. Finally, we use Monte Carlo simulations to show how the size of this bias varies under different peer assignment regimes.

3.1 Decomposition of the Peer Effects Estimator

Consider the following OLS regression model

$$y_{ig} = \mu + \pi_0 x_i + \pi_1 \bar{x}_g + \xi_i, \qquad (1)$$

where y_{ig} is the grade of student i in group g, x_i is the measure of student ability, $\bar{x}_g$ is the average of x_i in group g, ξ_i is an error term and $E[\xi_i | x_i] = E[\xi_i | \bar{x}_g] = 0$. For simplicity we will discuss group average measures of ability as opposed to the leave-out means (i.e., the group average excluding student *i*) as this distinction matters little econometrically. In particular, Angrist (2014)

[4] The problem we describe here for the estimation of peer effects also arises in other contexts, for example, when including both own household income and the average household income in a geographic area in the same regression. Whenever "own status" and some group average of this status are included in the same regression measurement error can lead to upward bias in the estimated group coefficient.

has shown that the peer effects estimator, when using leave-out means instead of group averages, differs only by a factor of $\frac{N_g}{N_g-1}$, where N_g is the size of group g. Acemoglu and Angrist (2001) have shown that in this setup the population parameter π_1 is equal to

$$\pi_1 = \phi(\psi_{IV} - \psi_{OLS}) \approx \psi_{IV} - \psi_{OLS}, \quad (2)$$

where ψ_{OLS} is equal to the population coefficient from a bivariate regression of y_{ig} on x_i, and ψ_{IV} is the population coefficient of ability in a two-stage least squares IV regression of y_{ig} on ability using group dummies as instruments for ability. $\phi = \frac{1}{1-R^2}$, where R^2 is the R-squared from the first stage of the above IV regression.[5] As this R-squared is typically close to zero, the peer effects estimator is approximately equal to the difference between the IV and OLS estimator of grades on own ability.

One can see from Equation (2) that not only peer effects, but all factors that lead to a difference between ψ_{IV} and ψ_{OLS} will affect π_1. We will focus here on measurement error in x_i as a plausible reason why ψ_{IV} and ψ_{OLS} would differ even in the absence of peer effects.[6] But how can measurement error lead to an overestimation of peer effects? The intuition behind this is as follows: If there is measurement error in x_i, both ψ_{IV} and ψ_{OLS} are attenuated. When student assignment to peer groups is systematic, ψ_{IV} is, in the absence of actual peer effects, less attenuated than ψ_{OLS}, which leads to an overestimation of peer effects. While Angrist also argues that the direction of the bias depends on the assignment mechanism, he is—probably due to the condensed

[5] Definitions of ψ_{OLS}, ψ_{IV} and R^2 can be found in Appendix A1.1.
[6] Angrist further mentions weak instrument bias: If students are randomly assigned to groups an IV with a weak first stage might bias ψ_{IV} towards ψ_{OLS} which would bias peer effects estimates towards zero.

nature of his overview—not explicit on why this is the case. We show this bias analytically in the next section.

3.2 Measurement Error and Bias of π_1

We model the grade data generating process as follows:

$$y_{ig} = \delta + \beta_0 x_i^* + \beta_1 \bar{x}_g^* + u_i, \quad (3)$$

where y_{ig} is the grade of student *i* in group *g*, x_i^* is student *i's* latent ability, x_g^* is the group average of x_i^* in group g, δ is a constant, β_0 is the causal effect of ability on grade and the parameter of interest, β_1, is the causal effect of group average ability on grade. We, however, only observe a noisy measure of individual ability, $x_i = x_i^* + \varepsilon_i$, where ε_i is classical measurement error, which has a mean of zero and is independent of $x_i^*, \bar{x}_g^*$ and u_i. Measurement error in the measure of individual ability will automatically lead to measurement error in the measure of group ability so that $\bar{x}_g = \bar{x}_g^* + \bar{\varepsilon}_g$, where $\bar{x}_g^*$ and $\bar{\varepsilon}_g$ are the respective averages of x_i^* and ε_i in group *g*. To focus this discussion on the role of measurement error, we assume that $E[u_i | x_i^*] = E[u_i | \bar{x}_g^*] = 0$, which means that if we would perfectly observe x_i^* and $\bar{x}_g^*$ (i.e., if $\varepsilon_i = 0$), π_1 would be equal to β_1.[7] We take Equation (2) as a starting point and further decompose the peer effects estimator given the definition of variables and data generating process defined above. For brevity, we do the analytical step-by-step decomposition in Appendix A1.1, where we show that:

$$\pi_1 = \beta_1 \phi\left(\frac{Var(\bar{x}_g^*)}{Var(\bar{x}_g)} - \frac{Var(\bar{x}_g^*)}{Var(x_i)}\right) + \beta_0 \phi\left(\frac{Var(\bar{x}_g^*)}{Var(\bar{x}_g)} - \frac{Var(x_i^*)}{Var(x_i)}\right). \quad (4)$$

[7] Note that throughout this paper we understand ability very broadly as all stable factors that influence grades, including innate ability, motivation, and access to academic resources.

Note that ϕ, which is equal to $\frac{1}{1-R^2}$, is always larger than one and usually close to one. We define W $\equiv \frac{Var(\bar{x}_g^*)}{Var(\bar{x}_g)} - \frac{Var(\bar{x}_g^*)}{Var(x_i)}$ and Q $\equiv \frac{Var(\bar{x}_g^*)}{Var(\bar{x}_g)} - \frac{Var(x_i^*)}{Var(x_i)}$ and can thus rewrite Equation (4) like this:

$$\pi_1 = \beta_1 \phi W + \beta_0 \phi Q. \quad (5)$$

We show in Appendix A1.2 that without measurement error π_1 is equal to β_1. We further show that ϕW will always range between 0 and 1, and in the case of random assignment to peer groups, it will be equal to the test reliability of ability $\frac{Var(x_i^*)}{Var(x_i)}$. So ϕW alone would only lead to an attenuation of π_1. To understand any potential upward bias of π_1, we need to understand the relationship between the ratio of variances of latent to measured group average ability $\frac{Var(\bar{x}_g^*)}{Var(\bar{x}_g)}$ and the ratio of variances of latent to measured individual ability $\frac{Var(x_i^*)}{Var(x_i)}$ in Q. When all individual ability measures have the same variance, we can rewrite Q as (see Appendix A1.2):

$$Q = \frac{Var(x_i^*)}{Var(x_i^*) + \frac{Var(\varepsilon_i)}{(1+(N_g-1)\rho)}} - \frac{Var(x_i^*)}{Var(x_i^*)+Var(\varepsilon_i)}, \quad (6)$$

where ρ is the average correlation of the distinct student abilities in group g,[8] and $Var(\varepsilon_i)$ is the variance of the measurement error.

Understanding the role of ρ is central in understanding a potential upward bias in π_1. Importantly, when students are randomly assigned, ρ will be zero and the first and second term of Equation (6) will be equal in size. This means that Q will be zero and π_1 will only be attenuated. When students of similar ability tend to be grouped together, ρ will be positive, and the first term

[8] More specifically, $\rho = \frac{1}{N_g(N_g-1)} \sum_{i \neq j}^{N_g} \frac{Cov(x_i^*, x_j^*)}{Var(x_i^*)}$.

of Equation (6) will be larger than the second one; hence, Q will be positive. In this case—assuming that the effect of own ability on grade β_0 is positive—$\pi_1$will be upward biased in the absence of peer effects. Given that ρ is positive, the bias will increase with the group size N_g. Note that a positive ρ can also be driven by non-random grouping at a higher level. In our setting, for example, students' course selection is non-random, while assignment to sections within courses is random. Fortunately, including course fixed effects eliminates this problem by taking out the correlation of individual abilities, which is driven by systematic assignment of students to courses.[9]

Figure (1) visualizes our analytical findings by describing the bias as a function of the true peer effects for different values of ρ. To get an idea about the potential size of the bias we use plausible values for the variables and parameters in Equation (4). In particular, we set $Var(x_i^*) = 1$ and $Var(\varepsilon_i) = 1/9$ to get a high test reliability of 0.9, which is approximately the test reliability of the SAT for measured ability (The College Board, 2014). The effect of own ability on grade β_0 is set to 0.6, which is approximately equal to the coefficient of the measure of own ability in our setting. We also assume that own ability, peer ability and grade are measured with the same scale so that we can compare the magnitude of the bias with typical peer effects estimates from the review of Sacerdote (2011). For $\rho = 0$, as is the case under random assignment, peer effect estimates are only attenuated. For $\rho > 0$ peer effects estimates are upward biased in the absence of peer effects, and this bias declines as peer effects increase.

[9] When including course fixed effects, Equation (4) holds if we replace x_i with the residuals of a regression of x_i on course dummies, $\tilde{x}_i$, throughout. $\tilde{x}_i$ is the ability measure demeaned at the course level (see: Angrist, 2014; Frisch & Waugh, 1933).

Figure 1: Bias as a Function of the Actual Peer Effects for Different Values of ρ

Note: This Figure is based on Equation (4), with $\beta_0 = 0.6$, $Var(\bar{x}_g^*) = 1$ and $Var(\varepsilon_i) = 1/9$.

The overall direction of the bias depends on whether the upward bias caused by $\beta_0 \phi Q$ exceeds the attenuation bias caused by $\beta_1 \phi W$ (see Equation (5)). The potential magnitude of the bias is substantial: In the absence of peer effects, for example, and with a group assignment that leads to ρ = 0.1, OLS estimates would, on average, wrongly suggest that a one point increase in peer ability leads to an approximately 0.04 point increase in grade. This bias would be even larger for more noisy ability measures: a test reliability of 0.6, for example, would lead to an upward bias of approximately 0.12. Differences in measurement error, degree of assortative assignment and group size between studies can thus contribute to the large heterogeneity in peer effects estimates

through the mechanism described above. For comparison, the linear in means peer effects estimates reported in the review by Sacerdote (2011) range from -0.12 to 6.8 with a median of 0.3.[10]

3.3 Results from Monte Carlo Simulations

To confirm our analytical results, we estimate Equation (3) using Monte Carlo simulations.[11] We use the same values for variables and parameters as for Figure 1[12] and show how results vary by the student assignment mechanism. In particular, we simulate an environment with 1,500 students who are divided into 10 courses. Each course has 10 sections, and each section has 15 students. To see how estimates depend on the true peer effect, we set β_1 to 0 and 0.3. We test three different assignment mechanisms: 1) Random assignment to courses and sections, 2) Assortative course assignment and random section assignment, and 3) Random course assignment and assortative section assignment. Assortative assignment here means that students are assigned (or self-select) to peers with similar abilities. Under the random assignment, ρ is approximately 0 and assignment mechanisms 2 and 3 lead to a ρ of approximately 0.1. The assortative assignment of students to sections is based on one variable which is correlated with ability. In practice this "assignment variable" may or may not be observable to the researcher.

Table 1 shows the average difference between estimated and actual peer effects—an indication of the estimation bias—using OLS estimation with 1,000 Monte Carlo replications. The

[10] This refers to the peer effects estimates for primary and secondary schools reported in Table 4.2 in Sacerdote (2011).
[11] The Stata Do-file of this simulation be downloaded under: http://ulfzoelitz.com/publications/extra-material
[12] This means that $\beta_0 = 0.6$, ability $x_i^* \sim N(0,1)$, measurement error $\varepsilon_i \sim N(0,1/9)$ and group ability $\bar{x}_g^*$ and its measure $\bar{x}_g$ are calculated as laid out above. We set the error term of the model to $u_i \sim N(0,1)$.

first rows of Panels A, B and C show that peer effect estimates are unbiased in the absence of measurement error for all assignment mechanisms. This confirms that the bias discussed here is not driven by selection bias in the classical sense.

The second rows of Panels A, B and C show the results with measurement error. All results confirm our analytical discussion. Under random assignment to courses and sections, measurement error leads to attenuation bias (see Columns (3) and (4) of Panel A). With assortative assignment, however, peer effects are overestimated (see Columns (1) and (3) of Panels B and C). The size of the bias is as predicted by our analytical results: in the absence of peer effects, for example, and with course or section assignment that leads to a $\rho = 0.1$, the bias is approximately 0.04 (see Column (1) of Panels B and C). As expected, the inclusion of course fixed effects eliminates the upward bias caused by assortative assignment to courses, and peer effects estimates are then only attenuated (see Column (4) of Panel B).

In many practically relevant settings, non-random assignment to sections (i.e., the peer group of interest) is likely to lead to a positive ρ. In the (rare) case where this assortative assignment is based on an observable assignment variable, controlling for this variable eliminates the potential upward bias. Estimates will then again only be attenuated (see Column (4) of Panel C). This applies, for example, when students are tracked within schools based on past performance observable to the researcher. In this case controlling for students' past performance will eliminate the potential upward bias by the mechanism described above.

Table 1: Simulation Results on Bias in Peer Effects Estimates

	(1)	(2)	(3)	(4)
	Actual peer effect = 0.00		**Actual peer effect = 0.30**	
Panel A	Random course and section assignment ($\rho = 0$)			
Average estimate – actual peer effect (Without measurement error)	-0.0021 [-0.0074 ; 0.0056]	-0.0035 [-0.0086 ; 0.0049]	-0.0021 [-0.0085 ; 0.0042]	-0.0035 [-0.0102 ; 0.0032]
Average estimate – actual peer effect (With measurement error)	0.0004 [-0.0066 ; 0.0059]	0.0011 [-0.0077 ; 0.0055]	-0.0302 [-0.036 ; -0.0239]	-0.0310 [-0.0377 ; -0.0242]
Course fixed effects	NO	YES	NO	YES
Panel B	Assortative course assignment & random section assignment ($\rho = .1$)			
Average estimate – actual peer effect (Without measurement error)	-0.0015 [-0.0030 ; 0.0061]	-0.0016 [-0.0058 ; 0.0091]	-0.0015 [-0.0030 ; 0.0061]	-0.0016 [-0.0058 ; 0.0091]
Average estimate – actual peer effect (With measurement error)	0.0401 [0.0356 ; 0.0445]	-0.0017 [-0.0054 ; 0.0088]	0.0268 [0.0223 ; 0.0313]	-0.0307 [-0.0380 ; -0.0234]
Course fixed effects	NO	YES	NO	YES
Panel C	Random course assignment & assortative section assignment ($\rho = .1$)			
Average estimate – actual peer effect (Without measurement error)	0.0035 [-0.0090 ; 0.0080]	0.0038 [-0.0060 ; 0.0080]	0.0035 [-0.0090 ; 0.0080]	0.0038 [-0.0060 ; 0.0080]
Average estimate – actual peer effect (With measurement error)	0.0421 [0.0377 ; 0.0464]	-0.0032 [-0.0012 ; 0.0074]	0.0287 [0.0243 ; 0.0332]	-0.0116 [-0.0161 ; -0.0071]
Controlling for assignment variable	NO	YES	NO	YES

Note: Monte Carlo simulations based on 1,000 repetitions for each reported estimate. 95% confidence intervals are in brackets [;].

Given that ability can only be measured with some degree of error, these findings have different implications for studies from non-experimental and (quasi) experimental settings. Because in non-experimental settings assortative assignment is likely, measurement error can lead to an upward bias on top of any potential selection bias. Studies that use data from (quasi) experimental settings, on the contrary, do not have these problems. It has been already well-established that random assignment or systematic assignment based observables eliminates selection bias. We have now added to this by showing that in these settings measurement error will only lead to attenuation bias. In the remainder of this paper, we will present new evidence on the structure of peer effects in a setting where students were randomly assigned to university sections.

4 Background

4.1 Institutional Environment

The data we collected for this paper comes from the School of Business and Economics (SBE) of Maastricht University, which is located in the south of the Netherlands.[13] Currently there are approximately 4,200 students at the SBE enrolled in bachelor's, master's, and PhD programs. Because of its proximity to Germany, the SBE has a large German student population (53 percent) mixed with students of Dutch (33 percent) and other nationalities. Approximately 37 percent of the students are females. The academic year at the SBE is divided into four regular teaching periods of two months and two skills periods of two weeks. Students usually take two courses at the same time in the regular periods and one course in the skills period. We exclude courses in skills periods

[13] See also Feld, Salamanca and Hamermesh (2015) for a detailed description of the institutional background and examination procedure at the SBE.

from our analysis because these are often not graded and we could not always identify the relevant peer group.[14]

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

The SBE differs from other universities with respect to its focus on Problem Based Learning.[15] The general Problem Based Learning setup is that students generate questions about a topic at the end of one session and then try to answer these questions through self-study. In the next session, the findings are discussed with other students in the section. In the basic form of PBL, the teacher plays only a guiding role, and most of the studying is done by the students independently. Courses, however, differ in the extent to which they give guidance and structure to the students, depending on the nature of the subject covered, with more difficult subjects usually requiring more guidance, and the preference of the course coordinator and teacher.

Compared to the traditional lecture system, the Problem Based Learning system is arguably more group focused because most of the teaching happens in small groups in which group discussions are the central part of the learning process. Much of the students' peer interaction

[14] In some skills courses, for example, students are scheduled in different sections but end up sitting together in the same room. Furthermore, skills courses have no exam at the end of the skill period, and in many skills course, students do not receive a GPA-relevant grade but only a "pass" or a "fail" grade.
[15] See http://www.umpblprep.nl/ for a more detailed explanation of PBL at Maastricht University.

happens with members of their section, either in the sessions or while completing homework and in study groups.

4.2 Students' Course Evaluations

Two weeks before the exam, students are invited by email to evaluate the courses they are currently taking in an online questionnaire.[16] Students receive up to three email reminders, and the questionnaire closes before the day of the exam. Students are assured that their individual answers will not be passed on to anyone involved in the course. The teaching staff receives no information about the evaluation before they have submitted the final course grades to the examination office.[17] This "double blind" procedure is implemented to avoid a situation where either of the two parties retaliates with negative feedback in the form of lower grades or evaluations. The exact length and content of the online questionnaires differ by course. The questionnaire typically contains 19-25 closed questions and two open questions. For our analysis, we use the nine core questions that are assessed in most courses that allow us to investigate the effect of peers on group functioning, student effort and teacher functioning. These questions ask students about how they perceived the instructor, how many hours they studied for the course, and about the interaction with their fellow students. Data on students' course evaluations at the individual level were provided by the Department of Educational Research and Development of the SBE. The course evaluation data are described in greater detail in Section 8.

[16] For more information, see the course evaluation home page: http://iwio-sbe.maastrichtuniversity.nl/default.asp.
[17] After exam grades are published, teaching staff receive the results of the courses evaluations aggregated at the section level.

4.3 Assignment of Students to Sections

The Scheduling Department of the SBE assigns students to sections, teachers to sections, and sections to time slots. Before each period, there is a time frame in which students can register online for the courses they want to take. After the registration deadline, the scheduler is given a list of registered students and allocates the students to sections using a computer program. About ten percent of the slots in each group are initially left empty and are filled with students who register late.[18] This procedure balances the amount of late registration students over the sections. Before the start of the academic year 2010/11, the section assignment for master's courses and for bachelor's courses was conducted with the program Syllabus Plus Enterprise Timetable using the allocation option "allocate randomly" (see Figure A1 in the Appendix). Since the academic year 2010/11, all bachelor's sections have been stratified by nationality with the computer program SPASSAT.[19] Some bachelor's courses are also stratified by exchange student status. After the assignment of students to sections, the sections are assigned to time slots, and the program Syllabus Plus Enterprise Timetable indicates scheduling conflicts.[20] Scheduling conflicts arise for approximately 5 percent of the initial assignments. If the computer program indicates a scheduling

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

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

[20] There are four reasons for scheduling conflicts: (1) the student takes another regular course at the same time; (2) the student takes a language course at the same time; (3) the student is also a teaching assistant and needs to teach at the same time; and (4) the student indicated non-availability for evening education. By default, all students are recorded as available for evening sessions. Students can opt out of this default position by indicating this in an online form. Evening sessions are scheduled from 6 p.m. to 8 p.m., and approximately three percent of all sessions in our sample are scheduled for this time slot.

conflict, the scheduler manually moves students between different sections until all scheduling conflicts are resolved. After all sections have been allocated to time slots, the scheduler assigns teachers to the sections.[21] The section and teacher assignment is then published. After publication, the scheduler receives information on late-registering students and allocates them to the empty slots. The schedulers do not know the students nor do they observe their previous grades.

Only 20-25 students (less than one percent) officially switch sections per period. Switching sections is possible only through a student advisor and is allowed only for medical reasons or due to a conflict with sports practice for students who are on a list of top athletes.[22] Students sometimes switch their section unofficially when they have extra appointments. This type of switching is usually limited to one session, and students rarely switch sections permanently.[23]

There are some exceptions to this general procedure. First, when the number of late registering student exceeds the number of empty spots, the scheduler creates a new section that mainly consists of late registering students. We excluded eight late registration sections from the analysis.[24] Second, for some bachelor's courses, there are special sections consisting mainly of repeating students. Whether a repeater section is created depends on the preference of the course coordinator and the number of repeat students. We excluded 34 repeater sections from the analysis. Third, in some bachelor's courses students who are part of the Maastricht Research Based Learning

[21] Approximately ten percent of teachers indicate time slots when they are not available for teaching. They do so before they are scheduled, and the signature of the department chair is required.
[22] We do not have a record for these students and therefore cannot exclude them. However, section switching in these rare cases is mostly due to conflicts with medical and sports schedules and therefore unrelated to section peers.
[23] It is difficult to obtain reliable numbers on unofficial switching. From our own experience and consultation with teaching staff, we estimate that session switching happens in less than 1 percent of the sessions, and permanent unofficial class switching happens for less than 1 percent of the students.
[24] Students who register late, for example, generally have a lower GPA and might be particularly busy or stressed during the period in which they registered late, which may also affect their performance. This dynamic might create a spurious relationship between GPA and grades.

(MARBLE) program are assigned to separate sections where they often are assigned to more experienced teacher. Students of this program are typically the highest performing students of their cohort. We excluded 15 sections that consist of MARBLE students from the analysis.[25] Fourth, in six courses, the course coordinator or other education staff influenced the section composition.[26] We excluded these courses from our analysis. Fifth, some master's tracks have part-time students. Part-time students are scheduled mostly in evening classes, and there are special sections with only part-time students. We excluded 95 part-time students from the analysis. Sixth, we excluded the first-year-first-period courses of the two largest bachelor's programs (International Business and Economics) because in these courses only particular students, such as repeating student, have previous grades. Seventh, we excluded sections for which fewer than five students had a past GPA. For these courses, peer GPA does not reliably capture the peer quality of the students in the section. Eighth, we excluded sections with more than 16 students (two percent) because the official class size limit according to scheduling guidelines is 15, and in special cases 16. Sections with more than 16 students are a result of room availability constraints or special requests from course coordinators. We also excluded 36 courses from the estimation sample in which part of the final grade might have consisted of group graded components, such as joint papers or other jointly graded projects. After removing these exceptions, neither students nor teachers, and not even course coordinators, should have any influence the composition of the sections in our estimation sample.

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

[26] The schedulers informed us about these courses.

5 Data

We obtained data for all students taking courses at the SBE during the academic years 2009/2010, 2010/2011 and 2011/2012. Scheduling data were provided by the Scheduling Department of the SBE. The scheduling data include information on section assignment, the allocated teaching staff and the day and time the sessions took place, as well as a list of late registrations for our sample period. In total, we have 7,672 students, 395 courses, 3,703 sections and 39,813 grades in our estimation sample. Panel A of Table 2 provides an overview of courses, sections and students in the different years.[27]

The data on student grades and student background, such as gender, age and nationality, were provided by the Examinations Office of the SBE. The Dutch grading scale ranges from 1 to 10, with 5.5 being the lowest passing grade.

Table 2: Descriptive Statistics

Panel A

Academic year	Number of courses	Number of unique students	Number of sections	Average number of students per section	Number of grades
2009 / 10	110	3,819	1,134	13.21	11,925
2010 / 11	141	4,018	1,346	13.08	13,768
2011 / 12	144	4,131	1,223	14.18	14,120
All years	395	7,672	3,703	13.50	39,813

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

Table 2: Descriptive Statistics (continued)

Panel B

	Mean	S.D.	Min	25p	Median	75p	Max	Obs.
Student level information								
Course dropout	0.0841	0.278	0	0	0	0	1	43,471
Grade first attempt	6.537	1.886	1	6	7	8	10	39,813
Final grade	6.745	1.687	1	6	7	8	10	40,601
GPA	6.790	1.206	1	6.122	6.983	7.583	10	43,471
Section level information								
Number of registered students per section	13.50	1.317	5	13	14	14	16	43,471
Number of students that dropped class	2.291	1.977	0	1	2	3	14	43,471
Peer GPA	6.757	0.471	4.897	6.441	6.776	7.096	8.500	43,471
Within section SD of peer GPA	1.126	0.365	0.104	0.869	1.105	1.360	2.799	43,471
Student Background information								
Age	20.74	2.203	16.19	19.17	20.42	21.98	41.25	40,469
Female	0.369	0.483	0	0	0	1	1	40,469
Dutch	0.312	0.463	0	0	0	1	1	43,471
German	0.496	0.500	0	0	0	1	1	43,471
Bachelor student	0.793	0.406	0	1	1	1	1	43,471
BA International Business	0.403	0.491	0	0	0	1	1	43,471
BA Economics	0.280	0.449	0	0	0	1	1	43,471
Exchange student	0.0619	0.241	0	0	0	0	1	43,471

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

Figure 2: Distribution of Grades after the First Examination

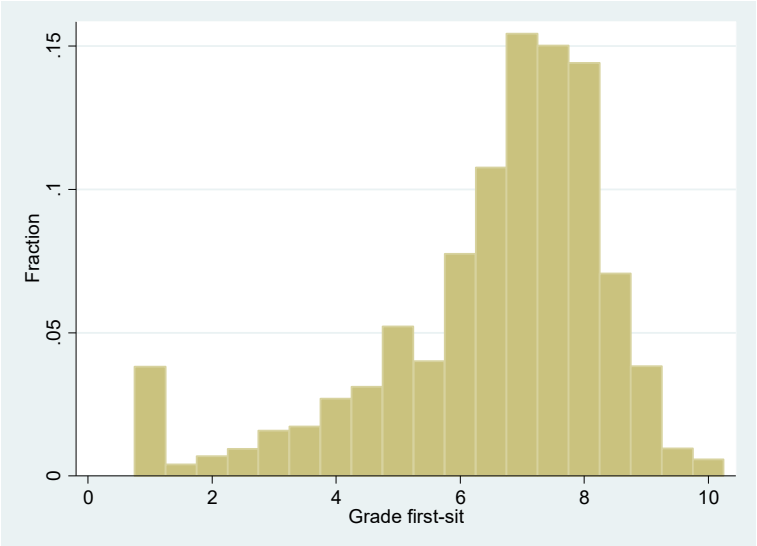


Figure 2 shows the distribution of final grades in our estimation sample. The final course grade is often calculated as the weighted average of multiple graded components, such as the final exam grade, participation grade, presentation grade or midterm paper grade.[28] The graded components and their respective weights differ by course, with most courses giving most of the weight to the final exam grade. If the final course grade of a student after taking the final exam is lower than 5.5, the student fails the course and has the option of taking a second and third attempt at the exam. We observe final grades after each attempt separately. For our analysis, we use only the final grade after the first exam attempt as an outcome measure because first- and second-attempt grades are not comparable.[29] For the construction of the student GPA, we use the final grades after the last attempt.[30]

Panel B of Table 2 shows some descriptive statistics for our estimation sample. Our sample contains 43,471 student course registrations. Out of these, 3,658 (8 percent) dropped out of the course during the course period. We therefore observe 39,813 course grades after the first attempt. The average course grade after the first attempt is 6.54. Approximately one fifth of the graded students obtain a course grade lower than 5.5 after the first attempt and therefore fail the course. The average final course grade (including grades from second and third attempt) is 6.75, and the average GPA is 6.79. Figure 3 shows the distribution of the GPAs based on final grades.

The peer GPA is the section-average GPA excluding the grades of the student of interest. Figure 4 shows the distribution of peer ability, measured as the average GPA of all other students in the section.

[28] We excluded 36 courses in which part of the final grade might have consisted of group graded components from the estimation sample (see Section 4.3).
[29] The second-attempt exam usually takes place two months after the first exam.
[30] We decided to use the GPA calculated from final grades because this is closer to the popular understanding of GPA.

Figure 3: Distribution of Student GPA

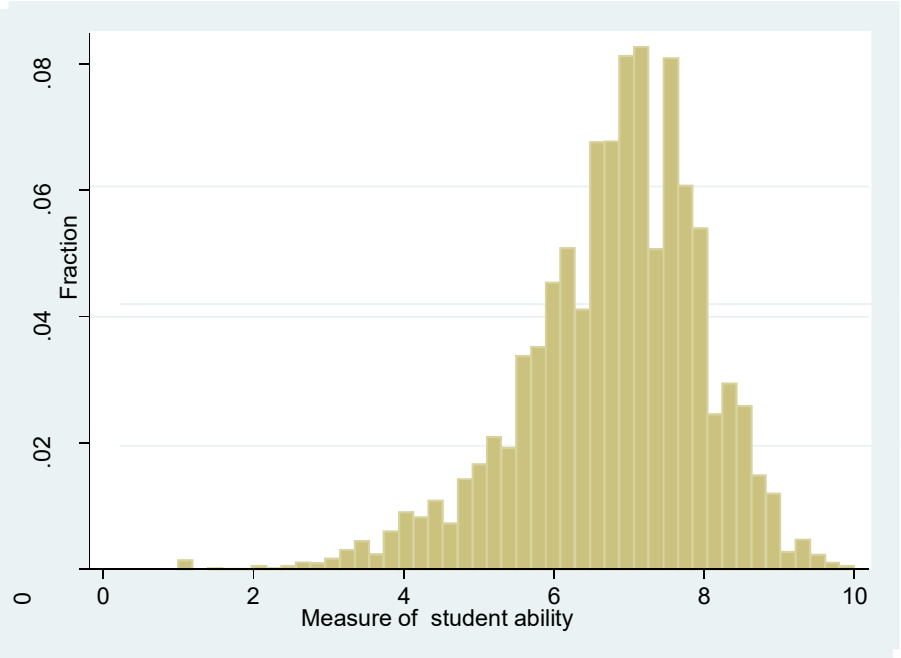
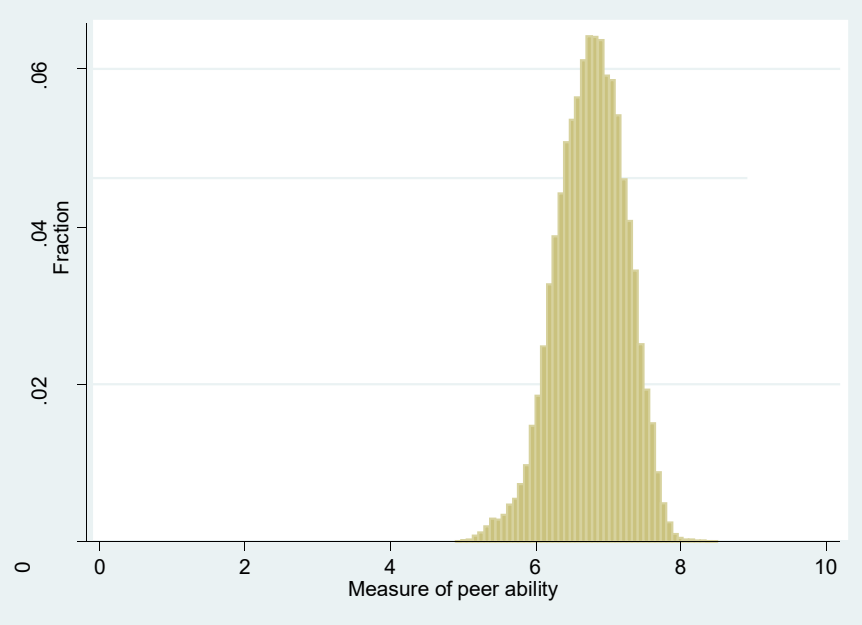


Figure 4: Distribution Peer GPA



6 Test for Random Assignment of Students to Sections

The scheduling procedure we describe in Section 4.3 shows that section assignment is random. Nevertheless, we test whether section assignment has the properties that one would expect under random assignment. In the spirit of standard randomization checks in experiments, we test whether section dummies jointly predict student pre-treatment characteristics when controlling for scheduling and balancing indicators. The pre-treatment characteristics we consider are GPA, age, gender, and student ID rank.[31] For each course in our sample, we run a regression of pre-treatment characteristics on section dummies as well as scheduling and balancing controls, and we F-test for joint significance of the section dummies. Thus, for each pre-treatment characteristic, we run approximately 400 regressions. Under conditional random assignment, the p-values of the F-tests of these regressions should be uniformly distributed with a mean of 0.5 (Murdoch, Tsai, & Adcock, 2008). Furthermore, if students are randomly assigned to sections within each course, the F-test should reject the null hypothesis of no relation between section assignment and students' pre-treatment characteristics at the 5 percent, 1 percent and 0.1 percent significance level in close to 5 percent, 1 percent and 0.1 percent of the cases, respectively.

The results of these randomization tests confirm that the section assignment is random (Section A3 in the Appendix provides a more detailed description on our randomization check). The average of the p-values of the F-tests is close to 0.5 (see Table A1 in the Appendix), and the p-values are roughly uniformly distributed (see Figure A2 in the Appendix). Table 2 shows in how

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

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

Table 3: Randomization Check of Section Assignment

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Dependent variable	Total number of courses	Number significant	Percent significant	Number significant	Percent significant	Number significant	Percent significant
Joint F-test significant:		...at the 5 percent level		...at the 1 percent level		...at the 0.1 percent level	
GPA	395	22	5.57%	5	1.27%	1	0.25%
Age	392	24	6.12%	11	2.81%	4	1.02%
Gender	389	18	4.63%	3	0.77%	0	0.00%
ID rank	395	20	5.06%	8	2.03%	2	0.51%

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

7 Empirical Strategy and Results

7.1 Empirical Strategy

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

$$Y_{igt} = \beta_1 GPA_{i,t-1} + \beta_2 \overline{GPA}_{g-i,t-1} + \gamma' Z_{igt} + \varepsilon_{igt}. \qquad (7)$$

The dependent variable Y_{ist} is the grade of student i, in a course-specific section g, at time t. $GPA_{i,t-1}$ is the past GPA of student i, and $\overline{GPA}_{g-i,t-1}$ is the average past GPA of all the students in the section excluding student i. Z_{igt} is a vector of additional controls, and ε_{igt} is an error term. In all specifications, Z_{igt} consists of dummies for day of the week and time of the day of the sessions, German, Dutch, exchange student status, late registration status, and year-course-period fixed effects.[32] In other specifications, we also include other-course fixed effects—fixed effects for the other course taken at the same time—and teacher fixed effects.[33] Note that $GPA_{i\,t-1}$ and $\overline{GPA}_{g-i,t-1}$ might measure own and peer ability with some error.[34] This might bias our peer effects estimates through the mechanisms described in Section 3. Since group assignment is random at the section level and we include year-course-period fixed effects, this will lead to an attenuation of peer effects. Including stratification controls and teacher fixed effects should increase the precision but not affect the size of the estimates. Conceptually, including scheduling controls and other-

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

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

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

course fixed effects should pick up all leftover non-random variation in section assignment that is due to conflicting schedules. To allow for correlations in the outcomes of students within each course, we cluster the standard errors at the course-year-period level. We standardized Y_{igt}, $GPA_{i\,t-1}$ and $\overline{GPA}_{g-i,t-1}$ to have mean of zero and a standard deviation of one over the estimation sample to simplify the interpretation of the coefficients.

7.2 Linear-in-means Results

Before we show estimates of peer effects on grades, we check whether peer GPA is related to course dropouts. The course dropout rate is only 8 percent at the SBE. OLS regressions, which we omit for brevity, show that neither average peer GPA nor the other peer GPA variables we use when estimating heterogeneous effects significantly predict course dropout. We therefore don't worry about selection bias when interpreting peer effects estimates on grades.

Table 4: Baseline Estimates – Linear-in-means

	(1) Std. Grade	(2) Std. Grade	(3) Std. Grade	(4) Std. Grade
Standardized peer GPA	0.0108*	0.0114*	0.0121**	0.0126**
	(0.006)	(0.006)	(0.005)	(0.006)
Standardized GPA	0.5606***	0.5605***	0.5623***	0.5622***
	(0.016)	(0.016)	(0.016)	(0.016)
Observations	39,813	39,813	39,813	39,813
R-squared	0.432	0.441	0.448	0.457
Course FE	YES	YES	YES	YES
Staff FE	NO	YES	NO	YES
Other course FE	NO	NO	YES	YES

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

Table 4 shows the results of OLS regressions with the standardized grade as the dependent variable. The table shows that being assigned to section peers with a higher GPA causes higher course grades. The coefficient of standardized peer GPA is small but statistically significant in all models. The inclusion of teacher fixed effects and other-course fixed effects hardly change the effect size or its standard errors. The reported estimate in the most complete specification in Column (4) shows that being assigned to peers with a one standard deviation higher GPA increases the student's grade by, on average, 1.26 percent of a standard deviation. The results are very similar when we define own and peer GPA solely based on first year grades (see Table A2 in the Appendix). In terms of the Dutch grading scale, this estimate means that, for example, an increase of peer GPA from 6.5 to 7.0 is associated with a grade increase from 6.50 to 6.523, a small and economically insignificant effect. It follows from our discussion in Section 3 that measurement error leads to attenuation of our estimator, and this attenuation is proportional to the test reliability of ability. To get a rough idea of the unattenuated coefficient, we can divide the coefficient by the split-half correlation of GPA, 0.72, an estimator of the test reliability.[35] This increases the estimate to 1.75 percent of a standard deviation.

To explore heterogeneous effects we extend the baseline analysis by additionally including interaction terms of peer GPA with dummies of course type or student gender. Table 5 shows the results of this analysis, where we include the baseline results of Table 4 Column (4) in Column (1) for comparison. Column (2) shows that peer effects estimates are larger for master's compared to bachelor's courses, although the difference is not statistically significant. Previous studies have

[35] In order to calculate the split-half correlation we randomly assigned all past grades of a student into two groups and constructed two GPAs based on these subgroups. The split-half correlation is the correlation of these ability measures.

found that peer effects differ between technical and non-technical subjects. While Brunello et. al. (2010) and Carrel et al. (2009) find larger peer effects in technical subjects, Arcidiacono, Foster, Goodpaster and Kinsler (2012) find larger peer effects in non-technical subjects.[36] To test whether peer effects in our settings differ by course technicality, we classified a course as "Technical" if at least one of the following words appeared in the course description: "math, mathematics, mathematical, statistics, statistical, theory focused." Doing this, we categorized 31 percent of the courses as "Technical." Column (3) shows that peer effects estimates are a little bit smaller in

Table 5: Linear-in-means Estimates with Course Type and Gender Interactions

	(1) Std. Grade	(2) Std. Grade	(3) Std. Grade	(4) Std. Grade
Standardized peer GPA	0.0126**	0.0303*	0.0153**	0.0164**
	(0.006)	(0.017)	(0.007)	(0.007)
Standardized peer GPA * bachelor course		-0.0207		
		(0.018)		
Standardized peer GPA * Technical course			-0.0077	
			(0.012)	
Standardized peer GPA * Female				-0.0080
				(0.011)
Standardized GPA	0.5622***	0.5622***	0.5622***	0.5841***
	(0.016)	(0.016)	(0.016)	(0.017)
Female				0.0401***
				(0.012)
Observations	39,813	39,813	39,813	37,210
R-squared	0.457	0.457	0.457	0.475

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

[36] Brunello et al. (2010) compare peer effects estimates in hard science and social science majors. Carrell et. al. (2009) compare peer effects in math and science with humanities and social science courses. Arcidiacono et al. (2012) compare peer effects in humanities, social sciences, hard sciences and mathematics and find larger effects for humanities and social sciences.

technical courses, but this difference is not statistically significant. Column (4) shows that estimated peer effects are somewhat larger for males, but again this difference is not statistically significant.

7.3 Heterogeneity by Own and Peer GPA

The specifications in Tables 4 and 5 are linear-in-mean, which implicitly assumes that all students are linearly affected by the mean GPA of their peers. Previous studies, however, have shown that peer effects are likely heterogeneous with respect to both student and peer achievement (Burke & Sass, 2013; Carrell et al., 2013). We test for these two sources of heterogeneity simultaneously by estimating a two-way interaction model similar to those of Carrell et al. (2013) and Burke and Sass (2013). To do this, we classify students as high, middle and low GPA based on whether their GPA is in the top, middle or bottom third of the course GPA distribution, respectively. We then calculate for each section the fraction of peers with high and low GPA and include interactions of students' own type (high, middle and low GPA) with the fraction of high and low GPA peers in the model we estimate. The coefficient "High GPA * Fraction of High GPA peers," for example, can be interpreted as showing how high GPA students are affected by increasing the fraction of high GPA peers in the section while keeping the fraction of low GPA peers constant. Put differently, the coefficient shows how high GPA students are affected if middle GPA peers (the reference group) are replaced with high GPA peers.

Table 6 shows the coefficients of these six interactions. Overall, the estimated effects are small in magnitude: for example, the largest coefficient, "Low GPA * Fraction of low-GPA peers," suggests that an increase of 20 percent in low GPA peers, which is equivalent to replacing three out of 15 middle GPA peers with low GPA peers, decreases the grade of a low GPA students by

2.63 percent of a standard deviation. The results for high and middle GPA students are in line with the linear-in-mean model: high and middle GPA students are positively affected by high GPA peers and negatively affected by low GPA peers. The results for low GPA students, however, are noticeably different. The point estimate suggests that low GPA students are *negatively* affected by high GPA peers. They are also negatively affected by peers from their own GPA group—low GPA peers. The effect of increasing the fraction of high GPA peers is significantly different for low GPA students compared to high and middle GPA students. To visualize the heterogeneous results, we plot the coefficients of the interactions in Table 6 in Figure 5. It shows that although peer effects seem to increase linearly with peer GPA for high and middle GPA students, the effect first increases and then decreases for low GPA students.

These estimates are qualitatively different from the pre-treatment findings of Carrell et al. (2013), who exploit random assignment, which suggested that in particular low achieving students benefit from high achieving peers. This result, however, was not robust to an intervention in which low achieving students were assigned to squadrons with a large fraction of high achieving peers. Contrary to the predictions of their pre-treatment findings, but in line with our results, low achieving students were actually harmed by this intervention.

To put these findings into a broader perspective, it is useful to think about which of the existing peer effects models are consistent with the patterns we observe in the data. Hoxby and Weingarth (2005) review a number of models that differ in the structure of peer effects they generate. While it not really possible to point to one single model that explains all of our results,

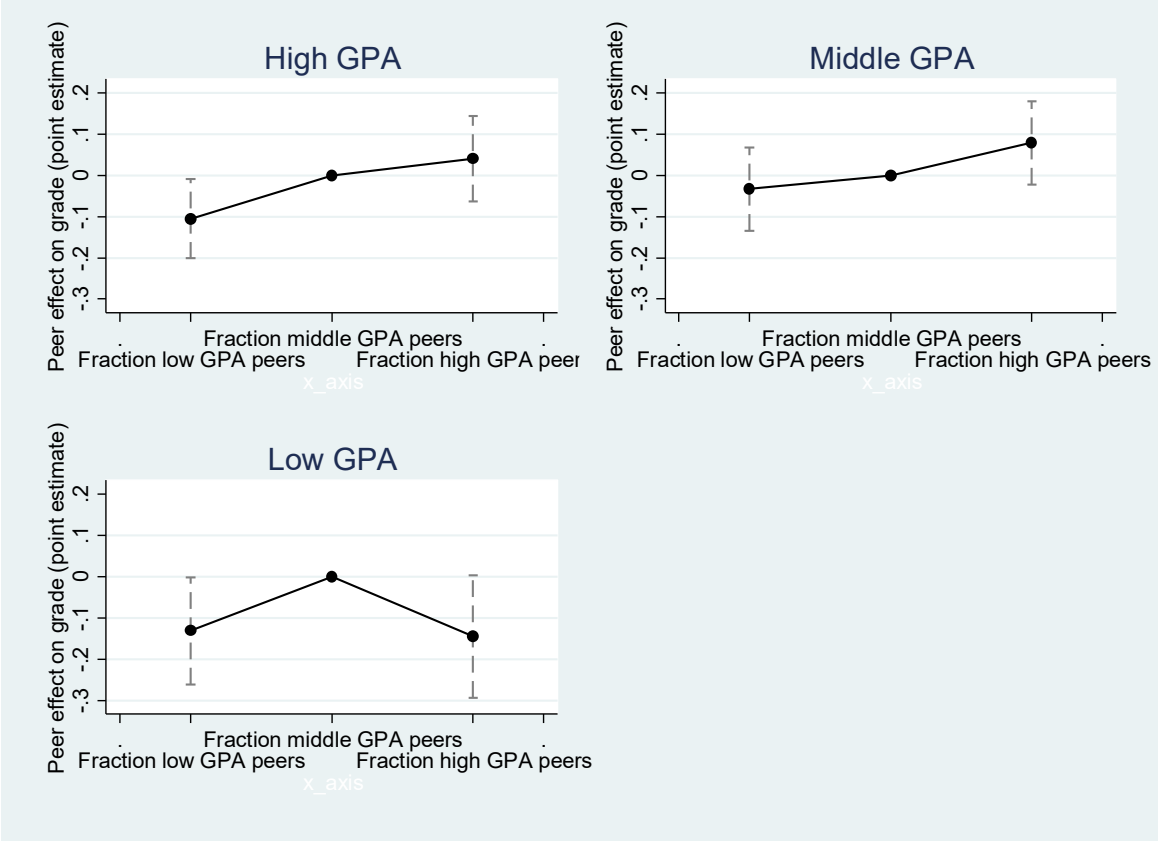
Table 6: Heterogeneous Effects

	(1) Std. Grade
High GPA * Fraction of high-GPA peers	0.0410
	(0.053)
High GPA * Fraction of low-GPA peers	-0.1047**
	(0.049)
Middle GPA * Fraction of high-GPA peers	0.0789
	(0.052)
Middle GPA * Fraction of low-GPA peers	-0.0332
	(0.052)
Low GPA * Fraction of high-GPA peers	-0.1449*
	(0.076)
Low GPA * Fraction of low-GPA peers	-0.1315**
	(0.066)
Observations	39,813
R-squared	0.461
F fraction of high peers [middle vs low]	5.40**
p-value	0.0207
F fraction of high peers [high vs low]	4.38**
p-value	0.0370
F fraction of high peers [high vs middle]	0.26
p-value	0.6114
F fraction of low peers [middle vs low]	1.39
p-value	0.2386
F fraction of low peers [high vs low]	0.10
p-value	0.7462
F fraction of low peers [high vs middle]	1.14
p-value	0.2867

Note: Robust standard errors clustered at the course level are in parentheses. The dependent variable is the standardized course grade. Additional controls include standardized GPA, course fixed effects, other-course fixed effects, teacher fixed effects as well dummies for day of the session, time of the session, German, Dutch, exchange student status and late registration status. *** p<0.01, ** p<0.05, * p<0.

the evidence we present in this paper is largely consistent with the Bad Apple model.[37] This model predicts that an increase in the proportion of bottom-achieving peers has a disproportionate negative effect on student achievement. Consistent with this, Table 6 shows that low GPA peers have a disproportionally large negative effect on high and low GPA students. Their effect on middle GPA students is also negative but small and not statistically different from zero.

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



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

[37] The Shining Light model would predict that a few excellent students would have a disproportionate positive effect on all other students. According to the Boutique model, students benefit most from interacting with students of their own type. Our findings that low achievers are harmed by both a higher proportion of low and high GPA peers are inconsistent with the Boutique model and the Shining Light model.

8 Channels of Peer Effects

8.1 Thinking about the Channels of Peer Effects

The peer effects we estimated in the previous section might be driven by a number of potential channels. Equation (7) describes a simplified version of the education production function that establishes the role of three channels, which we can test empirically.

$$Y_{ig}(x_g) = G_{ig}(x_g) + E_{ig}(x_g) + T_g(x_g) + \omega_{ig}, \qquad (7)$$

Y_{ig} is the grade of student i in section g, which is a function of peer ability x_g that enters through three channels: group functioning $G_{ig}(x_g)$, students' effort $E_{ig}(x_g)$ and teacher functioning $T_g(x_g)$. ω_{ig} captures all other omitted factors, including own ability and the effect of other channels. [38]

Prior to looking at the results, it is helpful to consider how we would expect these channels to matter in our specific setting. As described in more detail in Section 4.1, the Problem Based Learning teaching style practiced at the SBE has a large focus on classroom discussion in which students are supposed to explain the course content to each other. Since higher achieving students are likely more able to contribute in the classroom, we expect peer GPA to have a positive effect on group functioning. The expected direction of the effect of peer GPA on student effort and teacher functioning is less clear. On the one hand, better peers might induce students to work harder, e.g. via peer pressure, higher aspirations or social norms.[39] On the other hand, lower performing students might be demotivated by much better peers and exert less effort. Regarding the teacher

[38] For a theoretical model on how different channels might enter peer effects, see Duflo et al. (2011) and Conley Mehta, Stinebrickner and Stinebrickner (2015).

[39] See Mas and Moretti (2009) and Falk and Ichino (2006) for evidence of peer effects on effort provision. Peers expected performance might also provide a reference point which influences students' effort provision (see Kőszegi and Rabin, 2006 and Abeler, Falk, Goette and Huffman 2011).

behavior, we would expect teachers to adjust to the classroom peer composition by altering the difficulty level of their instructions. While an instruction level closer to the own ability level might be beneficial, larger deviations are likely to be detrimental. Such a mechanism would then imply that students benefit from having more peers of similar abilities. [40]

8.2 Empirical Evidence on the Channels

Table 7 shows the wording and answering scales of the items regarding group interaction, self-study hours and teacher evaluation.[41] In our estimation sample, 38 percent of the students start filling out the questionnaire. The last column in Table 7 shows that once students started the questionnaire, they answered almost all of items. Answering the course evaluation questionnaire is selective. We observe, for example, that students with higher GPAs are more likely to take part in the evaluation, and our results should be interpreted in light of this finding. We have nevertheless chosen to analyze students' course evaluation for two reasons. First, the survey response is not significantly related to peer quality as measured by mean peer GPA or to the peer GPA variables used in Section 7.3 (see Table A3 in the Appendix). Second, the student evaluation data gives us a unique insight into potential mechanisms in a way that is not available in most other studies. All results, however, should be interpreted with caution, and we interpret them as providing suggestive evidence.

For our analysis, we aggregate items in the domains of group functioning (two items) and teacher functioning (six items) by first summing all the standardized item answers and then

[40] It is also possible that both students and teachers compensate for having worse peers by working harder. Such compensation behavior could explain why the peer effects on grades are overall small in magnitude.
[41] Standard items on the course evaluation questionnaire also include items about learning material and general course evaluation. For a complete list of all the standard evaluation items and how each item relates to mean peer GPA, see Table A4 in the Appendix.

standardizing the sum of these values in each domain so that the aggregated categories have a mean of zero and a standard deviation of one. Answers to the question about self-study hours, which we use as a measure of student effort, are left in their natural unit. We impute missing values for items that students who started filling out the questionnaire did not answer either because the question was not in their course specific questionnaire or because they chose not to answer the question.[42]

Table 7: Evaluation Item, Answering Scales and Response Rates

Nr	Item Domain	Item Wording	Answer Scale	Response Rate Conditional on Participation
1	Group interaction	My tutorial group has functioned well.	1 - 5	94.1%
2	Group interaction	Working in tutorial groups with my fellow-students helped me to better understand the subject matters of this course.	1 - 5	93.5%
3	Self-study hours	How many hours per week on the average (excluding contact hours) did you spend on self-study (presentations, cases, assignments, studying literature, etc)?	0 - 80	92.8%
4	Teacher functioning	Evaluate the overall functioning of your tutor in this course with a grade	1 - 10	93.5%
5	Teacher functioning	The tutor sufficiently mastered the course content.	1 - 5	93.6%
6	Teacher functioning	The tutor stimulated the transfer of what I learned in this course to other contexts.	1 - 5	93.4%
7	Teacher functioning	The tutor encouraged all students to participate in the (tutorial) group discussions.	1 - 5	93.0%
8	Teacher functioning	The tutor was enthusiastic in guiding our group.	1 - 5	93.5%
9	Teacher functioning	The tutor initiated evaluation of the group functioning.	1 - 5	91.5%

Note: At Maastricht University, the teaching staff member present in the classroom is referred to as "tutor." Sections are commonly called "tutorial groups."

[42] Conditional on answering at least one of these questions the percentage of missing answers is between 5.9 and 9.2 percent depending on the item (see Table 7). We apply multiple imputations by chained equations (MICE) with 10 cycles. Note that imputing missing values might bias estimates if the missing at random assumption does not hold. We therefore also report ranges of point estimates using bounding methods in Table A5 in the Appendix, where we assume extreme values for missing answers. We also report results without any imputations in Table A6 in the Appendix. Results are very similar across these different models.

Panel A of Table 8 shows that the average peer GPA affects the evaluation of the group interaction positively. A one standard deviation increase in peer GPA leads to a 0.056 standard deviation increase in evaluation of group interaction. When redoing the analysis with each of the two items separately, we find that this result is only driven by the first item (see Table A4 in the Appendix). This suggests that the students notice the better group functioning but do not perceive higher benefits from it—a result that is not surprising given the small magnitude of the estimated peer effects on grades. Hours worked and teacher functioning are not significantly affected by peer GPA.

Panel B shows the results using the same specification for identifying peer effect heterogeneity as in Section 7. This model allows us to investigate if the effect of peer GPA on course evaluations is heterogeneous in terms of student and peer achievement. When comparing the different evaluation domains, we see that the peer variables are jointly significant in explaining the evaluation of the group interaction and not jointly significant in explaining the evaluation of the teacher or the self-study hours. The results for group interaction suggest that in particular the presence of high GPA peers matters. The point estimate suggests that a 20 percent increase in high GPA peers increases high GPA students' evaluation of the group interaction by about 0.105 standard deviations. The estimated effects of increasing high GPA peers are also positive and about half the size for medium and low GPA students, although the effect for medium GPA students is not quite statistically significant (p-value: 0.128).

All in all, our results suggest that, in our setting, group interaction is the most important of the three discussed channels. Interestingly, the effect of peer quality on group interaction appears to be linear. This implies that the inverse u-shaped pattern for low GPA students we found in

Section 7 is driven by other unobserved factors. We do not find evidence for adjustment of teacher behavior or student effort.

Table 8: The Effect of Peer Composition on Student Evaluations

Panel A

	(1) Std. Group interaction	(2) Self-study hours	(3) Std. Teacher evaluation
Standardized peer GPA	0.0561***	0.0339	0.0017
	(0.015)	(0.092)	(0.013)
Standardized GPA	-0.0348***	0.1353	-0.0426***
	(0.010)	(0.083)	(0.009)
Observations	15,441	15,441	15,441

Panel B

	(2)	(1)	(3)
High GPA * Fraction of high-GPA peers	0.5267***	-0.4390	0.0839
	(0.131)	(0.858)	(0.096)
High GPA * Fraction of low-GPA peers	-0.1132	-0.8487	-0.0893
	(0.121)	(0.852)	(0.101)
Middle GPA * Fraction of high-GPA peers	0.2265	-0.0499	-0.0462
	(0.139)	(0.992)	(0.124)
Middle GPA * Fraction of low-GPA peers	0.0202	1.8395*	0.0055
	(0.110)	(0.962)	(0.114)
Low GPA * Fraction of high-GPA peers	0.2659**	-0.3977	0.0650
	(0.119)	(0.919)	(0.114)
Low GPA * Fraction of low-GPA peers	0.0369	-1.6370*	-0.0874
	(0.120)	(0.956)	(0.114)
Observations	15,441	15,441	15,441
F joint significance of peer variables	5.60***	1.34	0.88
Prob > F	0.0000	0.2391	0.5080

Note: Robust standard errors clustered at the course-year-period level are in parentheses. The dependent variables are standardized Group interaction in Column (1), self-study hours in Column (2) and standardized teacher evaluation in Column (3). All specifications include course fixed effects, other-course fixed effects, teacher fixed effects as well dummies for day of the session, time of the session, German, Dutch, exchange student status and late registration status. All regressions reported in Panel B also include standardized GPA. We imputed missing values as explained in Footnote 40. *** p<0.01, ** p<0.05, * p<0.1.

Our results are consistent with findings of Booij et al. (2015), who study peer effects in settings similar to ours, where the peer group is defined at the section level at a Dutch University. They also find evidence for peer effects on group functioning and no evidence of peer effects on teacher functioning. Lavy et al. (2012) study the effect of the proportion of repeaters on student outcomes in secondary schools. Using student surveys they identify changes in teachers' pedagogical practices and increases in violence and classroom disruptions as important channels. Duflo et al. (2011) find in the context of Kenyan primary schools that teachers provide more effort, as measured by teacher absenteeism, when they are randomly assigned to a class of high compared to low achieving students. Taken together, these findings confirm the notion that the channels, which depend on the specific contexts, can create very heterogeneous peer effects.

9 Conclusion

This article adds to the discussion about threats to the identification of peer effects and provides empirical evidence of peer effects in higher education. In the analytical part of this paper, we have shown that measurement error can lead to substantial overestimation of peer effects in settings where peer group assignment is systematic. In settings where peer group assignment is random or based on an observable variable, however, measurement error will only lead to attenuation bias. These findings are good news for past and future peer effects studies that rely on natural random variation or exploit a perfectly observable assignment mechanism. Peer effects estimates obtained from studies with non-observable peer group assignment mechanism have to be interpreted with particular caution since they are prone to potentially severe upward bias due to measurement error. This bias is not the same as, and may occur on top of, any potential selection bias.

In the empirical part of this paper, we have estimated peer effects in a sample where university students are randomly assigned to sections. Consistent with previous research, we find effects of average peer quality on student grades that are small in size but statistically significant. These average effects hide important heterogeneity however. While high and middle ability students benefit from better peers, low ability students are harmed by their high ability peers. Evidence from students' course evaluations suggests that peer effects are driven mainly by changes in group interaction and not by adjustments in teachers' behavior or students' effort.

Our non-linear estimates suggest that it would be possible to achieve small overall gains in student performance by reorganizing peer groups. Without knowing the process that generates the observed peer effects, however, it is not clear whether this would be welfare enhancing. In principle, increased student performance can be a result of an increase in efficiency or an increase in students' or teachers' effort. An increase in student or teacher effort implies costs which should be weighed against the benefits from increased student performance. If, however, as our results suggest, the increase in students' performance is driven by better group interaction, reorganization of peer groups can lead to higher efficiency, and welfare gains could therefore be possible.

References

Acemoglu, D., & Angrist, J. (2001). How Large are Human-capital Externalities? Evidence From Compulsory-schooling Laws *NBER Macroeconomics Annual 2000, Volume 15* (pp. 9-74): MIT Press.

Ammermueller, A., & 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. doi: 10.1086/603650

Angrist, J. D. (2014). The Perils of Peer Effects. *Labour Economics, 30*, 98-108.

Arcidiacono, P., Foster, G., Goodpaster, N., & Kinsler, J. (2012). Estimating Spillovers Using Panel Data, with an Application to the Classroom. *Quantitative Economics, 3*(3), 421-470. doi: 10.3982/qe145

Board, T. C. (2014). Test Characteristics of the SAT

Booij, A. S., Leuven, E., & Oosterbeek, H. (2015). Ability Peer Effects in University: Evidence from a Randomized Experiment. *IZA Discussion Papers*(No. 8769).

Brock, W. A., & Durlauf, S. N. (2001). Discrete Choice with Social Interactions. *The Review of Economic Studies, 68*(2), 235-260.

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

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

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

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

Conley, T., Mehta, N., Stinebrickner, R., & Stinebrickner, T. (2015). *Social Interactions, Mechanisms, and Equilibrium: Evidence from a Model of Study Time and Academic Achievement*. NBER working paper. Cambridge, MA.

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

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

Epple, D., & Romano, R. (2011). Peer Effects in Education: A Survey of the Theory and Evidence. *Handbook of Social Economics, 1*(11), 1053-1163.

Feld, J., Salamancam, N., & Hamermesh, D. S. (2015). Endophilia or Exophobia: Beyond Discrimination. *The Economic Journal*, n/a-n/a. doi: 10.1111/ecoj.12289

Figlio, D. N. (2007). Boys Named Sue: Disruptive Children and Their Peers. *Education Finance and Policy, 2*(4), 376-394.

Frisch, R., & Waugh, F. V. (1933). Partial Time Regressions as Compared with Individual Trends. *Econometrica, 1*(4), 387-401.

Hoxby, C. (2000). *Peer Effects in the Classroom: Learning from Gender and Race Variation*. NBER Working Paper. National Bureau of Economic Research. Cambridge, MA.

Hoxby, C. M., & Weingarth, G. (2005). Taking Race out of the Equation: School Reassignment and the Structure of Peer Effects: Working paper.

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

Lazear, E. P. (2001). Educational Production. *The Quarterly Journal of Economics, 116*(3), 777-803.

Lin, X. (2010). Identifying Peer effects in Student Academic Achievement by Spatial Autoregressive Models with Group Unobservables. *Journal of Labor Economics, 28*(4), 825-860.

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

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

Moffitt, R. A. (2001). Policy Interventions, Low-level Equilibria, and Social Interactions. In H. P. Y. Steven N. Durlauf (Ed.), *Social Dynamics* (Vol. 4, pp. 45-82): MIT Press.

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

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

Sacerdote, B. (2011). Peer Effects in Education: How Might They Work, How Big are They and How much do we Know Thus Far? *Handbook of the Economics of Education, 3*, 249-277.

Sacerdote, B. (2014). Experimental and Quasi-Experimental Analysis of Peer Effects: Two Steps Forward? *Annual Review of Economics, 6*(1), 253-272. doi: doi:10.1146/annurev-economics-071813-104217

Stinebrickner, R., & Stinebrickner, T. R. (2006). What can be Learned about Peer Effects using College Roommates? Evidence from new Survey Data and Students from Disadvantaged Backgrounds. *Journal of Public Economics, 90*(8), 1435-1454.

Vigdor, J., & Nechyba, T. (2007). Peer Effects in North Carolina Public Schools. In P. E. P. Ludger Woessmann (Ed.), *Schools and the Equal Opportunity Problem*: Schools and the Equal Opportunity Problem', MIT Press.

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

APPENDIX

A1 Classical Measurement Error and the Estimation of Peer Effects

A1.1 Decomposing π_1

We rewrite Equation (2) as

$$\pi_1 = \frac{\psi_{IV} - \psi_{OLS}}{1 - R^2} \tag{A1}$$

and take this as starting point to further decompose the peer effects estimator based on the data generating process and variables defined in Section 3. ψ_{IV} is a two-stage least square IV estimator. The first stage of this IV regression uses group dummies as instruments for x_i, and the predicted values of this first stage are thus group averages of x_i. ψ_{IV} is therefore equal to the coefficient from a bivariate regression of y_{ig} on $\bar{x}_g$:

$$\psi_{IV} = \frac{Cov(\bar{x}_g, y_{ig})}{Var(\bar{x}_g)}. \tag{A2}$$

By substituting y_{ig} with Equation (3) and rearranging, we get:[43]

$$\psi_{IV} = \frac{Cov(\bar{x}_g, (\delta + \beta_0 x_i^* + \beta_1 \bar{x}_g^* + u_i))}{Var(\bar{x}_g)} \tag{A3}$$

$$\psi_{IV} = \beta_0 \frac{Var(\bar{x}_g^*)}{Var(\bar{x}_g)} + \beta_1 \frac{Var(\bar{x}_g^*)}{Var(\bar{x}_g)}.$$

Analogously, we can express ψ_{OLS}:

$$\psi_{OLS} = \frac{Cov(x_i, y_{ig})}{Var(x_i)}. \tag{A4}$$

By substituting y_{ig} with Equation (3) and rearranging, we get

$$\psi_{OLS} = \frac{Cov(x_i, (\delta + \beta_0 x_i^* + \beta_1 \bar{x}_g^* + u_i))}{Var(x_i)} \tag{A5}$$

[43] Note that $Cov(\bar{x}_g, x_i^*) = Var(\bar{x}_g^*)$.

$$\psi_{OLS} = \beta_0 \frac{Var(x_i^*)}{Var(x_i)} + \beta_1 \frac{Var(\bar{x}_g^*)}{Var(x_i)}.$$

The R^2 from first stage of the above IV estimation is equal to

$$R^2 = \frac{Var(\bar{x}_g)}{Var(x_i)}. \tag{A6}$$

Combining all the parts and substituting them into Equation (A1) we get:

$$\pi_1 = \frac{(\beta_0 \frac{Var(\bar{x}_g^*)}{Var(\bar{x}_g)} + \beta_1 \frac{Var(\bar{x}_g^*)}{Var(\bar{x}_g)}) - (\beta_0 \frac{Var(x_i^*)}{Var(x_i)} + \beta_1 \frac{Var(\bar{x}_g^*)}{Var(x_i)})}{1 - \frac{Var(\bar{x}_g)}{Var(x_i)}} \tag{A7}$$

$$\pi_1 = \beta_1 \phi (\frac{Var(\bar{x}_g^*)}{Var(\bar{x}_g)} - \frac{Var(\bar{x}_g^*)}{Var(x_i)}) + \beta_0 \phi (\frac{Var(\bar{x}_g^*)}{Var(\bar{x}_g)} - \frac{Var(x_i^*)}{Var(x_i)})$$

$$\pi_1 = \beta_1 \phi W + \beta_0 \phi Q,$$

where $\phi = \frac{1}{1 - \frac{Var(\bar{x}_g)}{Var(x_i)}}$.

A1.2 Understanding the Direction of the Overall Bias

Here we show how measurement error affects ϕW and ϕQ.

We start by rewriting ϕW

$$\phi \text{W} = \frac{1}{1 - \frac{Var(\bar{x}_g)}{Var(x_i)}} * (\frac{Var(\bar{x}_g^*)}{Var(\bar{x}_g)} - \frac{Var(\bar{x}_g^*)}{Var(x_i)}) = \frac{\frac{Var(\bar{x}_g^*)}{Var(\bar{x}_g)} - \frac{Var(\bar{x}_g^*)}{Var(x_i)}}{1 - \frac{Var(\bar{x}_g)}{Var(x_i)}}. \tag{A8}$$

Because $1 = \frac{Var(\bar{x}_g^*)}{Var(\bar{x}_g)} + \frac{Var(\bar{\varepsilon}_g)}{Var(\bar{x}_g)}$ and $\frac{Var(\bar{x}_g)}{Var(x_i)} = \frac{Var(\bar{x}_g^*)}{Var(x_i)} + \frac{Var(\bar{\varepsilon}_g)}{Var(x_i)}$, we can further rewrite ϕW as follows:

$$\phi W = \frac{\left[\frac{Var(\bar{x}_g^*)}{Var(\bar{x}_g)} - \frac{Var(\bar{x}_g^*)}{Var(x_i)}\right]}{\left[\frac{Var(\bar{x}_g^*)}{Var(\bar{x}_g)} - \frac{Var(\bar{x}_g^*)}{Var(x_i)}\right] + \langle \frac{Var(\bar{\varepsilon}_g)}{Var(\bar{x}_g)} - \frac{Var(\bar{\varepsilon}_g)}{Var(x_i)} \rangle}. \tag{A9}$$

Note that the terms in […] are identical. In the absence of measurement error

(i.e., if $Var(\varepsilon_i) = 0$), the terms in $\langle \ldots \rangle$ are equal to zero so that ϕW is equal to 1. In the presence of measurement error the denominator is larger than the numerator because $\frac{Var(\bar{\varepsilon}_g)}{Var(\bar{x}_g)} > \frac{Var(\bar{\varepsilon}_g)}{Var(x_i)}$, and therefore $0 < \phi W < 1$. Note that when students are randomly assigned, ϕW is equal to the test reliability of ability $\frac{Var(x_i^*)}{Var(x_i)}$.[44]

Now let's have a look at ϕQ. If all ability measures have the same variance, we can use the formula for the variance of the mean of correlated variables to rewrite $Var(\bar{x}_g^*) = \frac{Var(x_i^*)}{N_g} + \frac{N_g - 1}{N_g}\rho Var(x_i^*)$ and $Var(\bar{x}_g) = \frac{Var(x_i^*)}{N_g} + \frac{N_g - 1}{N_g}\rho Var(x_i^*) + \frac{Var(\varepsilon_i)}{N_g}$, where N_g is the number of students in group g, and ρ is the average correlation of the distinct student abilities in group g.[45] After canceling out N_g, we can rewrite $\frac{Var(\bar{x}_g^*)}{Var(\bar{x}_g)}$ as follows:

$$\frac{Var(x_i^*) + (N_g - 1)\rho Var(x_i^*)}{Var(x_i^*) + (N_g - 1)\rho V \quad (x_i^*) + Va \ (\varepsilon_i)} . \tag{A10}$$

We can now rewrite Q as follows:

$$Q = \frac{Var(x_i^*) + (N_g - 1)\rho Var(x_i^*)}{Var(x_i^*) + (N_g - 1)\rho Var(x_i^*) + Var(\varepsilon_i)} - \frac{Var(x_i^*)}{Var(x_i^*) + Var(\varepsilon_i)} \tag{A11}$$

$$Q = \frac{(1 + (N_g - 1)\rho) Var(x_i^*)}{(1 + (N_g - 1)\rho) Var(x_i^*) + Var(\varepsilon_i)} - \frac{Var(x_i^*)}{Var(x_i^*) + Var(\varepsilon_i)}$$

$$Q = \frac{Var(x_i^*)}{Var(x_i^*) + \frac{Var(\varepsilon_i)}{(1 + (N_g - 1)\rho)}} - \frac{Var(x_i^*)}{Var(x_i^*) + Var(\varepsilon_i)}.$$

[44] Note that under random assignment, $Var(\bar{x}_g^*) = \frac{Var(x_i^*)}{N_g}$ and $Var(\bar{x}_g) = \frac{Var(x_i^*)}{N_g} + \frac{Var(\varepsilon_i)}{N_g}$. Plugging these in to Equation (A8) and rearranging, you can see that $\phi \mathrm{W} = \frac{Var(x_i^*)}{Var(x_i)}$.

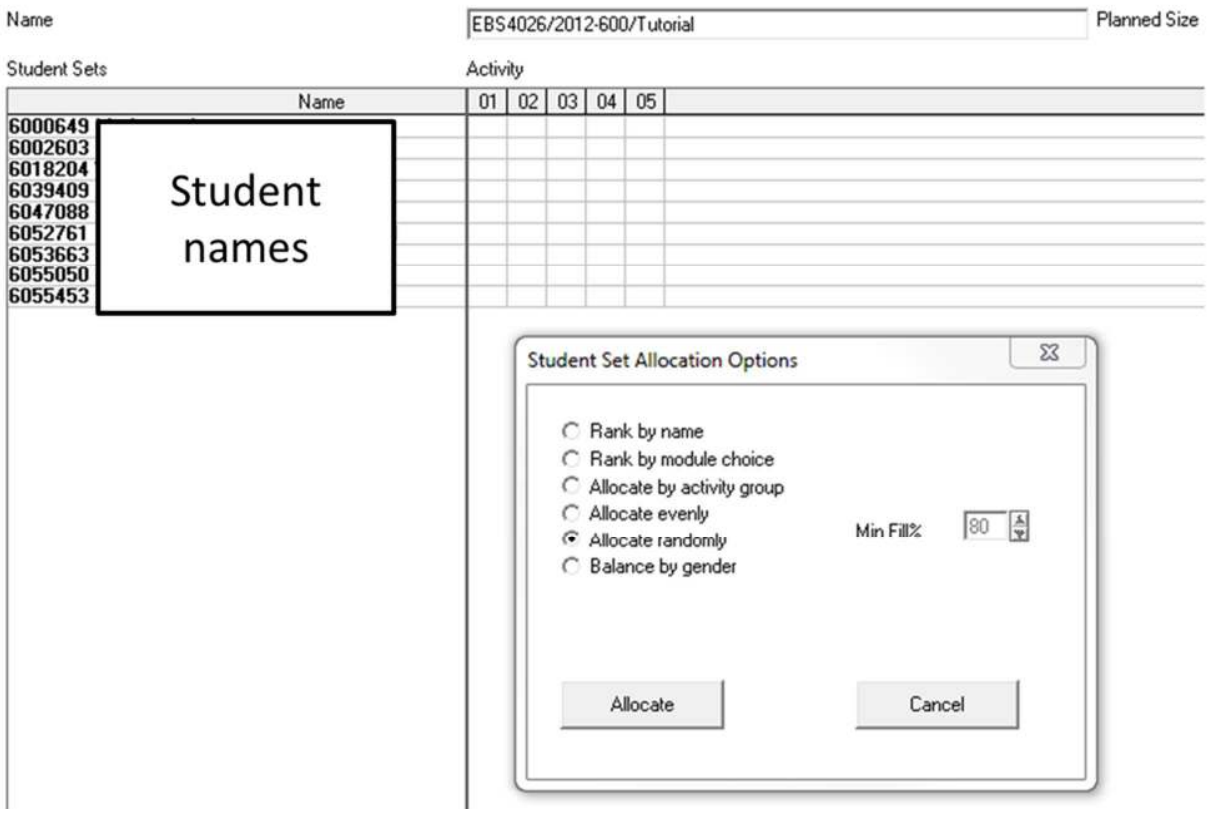
[45] $\rho = \frac{1}{N_g(N_g - 1)} \sum_{i \neq j}^{N_g} \frac{Cov(x_i^*, x_j^*)}{Var(x_i^*)}$.

Without measurement error, the first and the second term in Q are equal, and thus Q (and ϕQ) is equal to zero. With measurement error, the magnitude Q of depends on average correlation of the distinct student abilities ρ: under random assignment ρ will be equal to zero, both terms in Equation (A11) will be the same and Q will be zero. If students tend to be grouped according to their ability, ρ will be positive; the first term will be larger than the second term in Equation (A11), and Q (and ϕQ) will be positive. Given that students are systematically assigned to groups, the size of Q increases with ρ, N_g and β_0.

To conclude, in the absence of measurement error, ϕW is equal to one and ϕQ is equal to zero so that π_1 is equal to β_1. With measurement error, the sign of the overall bias depends on $\beta_1 \phi W$ and $\beta_0 \phi Q$. With random assignment ($\rho = 0$), ϕQ is equal to zero and π_1 is only attenuated. With systematic assignment ($\rho > 0$), ϕQ is positive. The overall size of the bias then depends on whether the upward bias caused by $\beta_0 \phi Q$ is larger than the downward bias—assuming that peer effects are positive—caused by $\beta_1 \phi W$.

A2 Additional Figure

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



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

A3 Randomization Check

We use the following empirical specification for our tests. Take y_i as a $1xN_i$ vector of the pre-treatment characteristics of students in course *i*. The pre-treatment characteristics we consider are GPA, age, gender, and student ID rank. $\boldsymbol{T} = (t_1, \ldots t_n)$ is a matrix of section dummies. **Z** is a matrix that includes dummies for other course taken at the same time, day and time of the sessions, German, Dutch, exchange student status and late registration status. ε_i is a vector of zero-mean independent error terms.

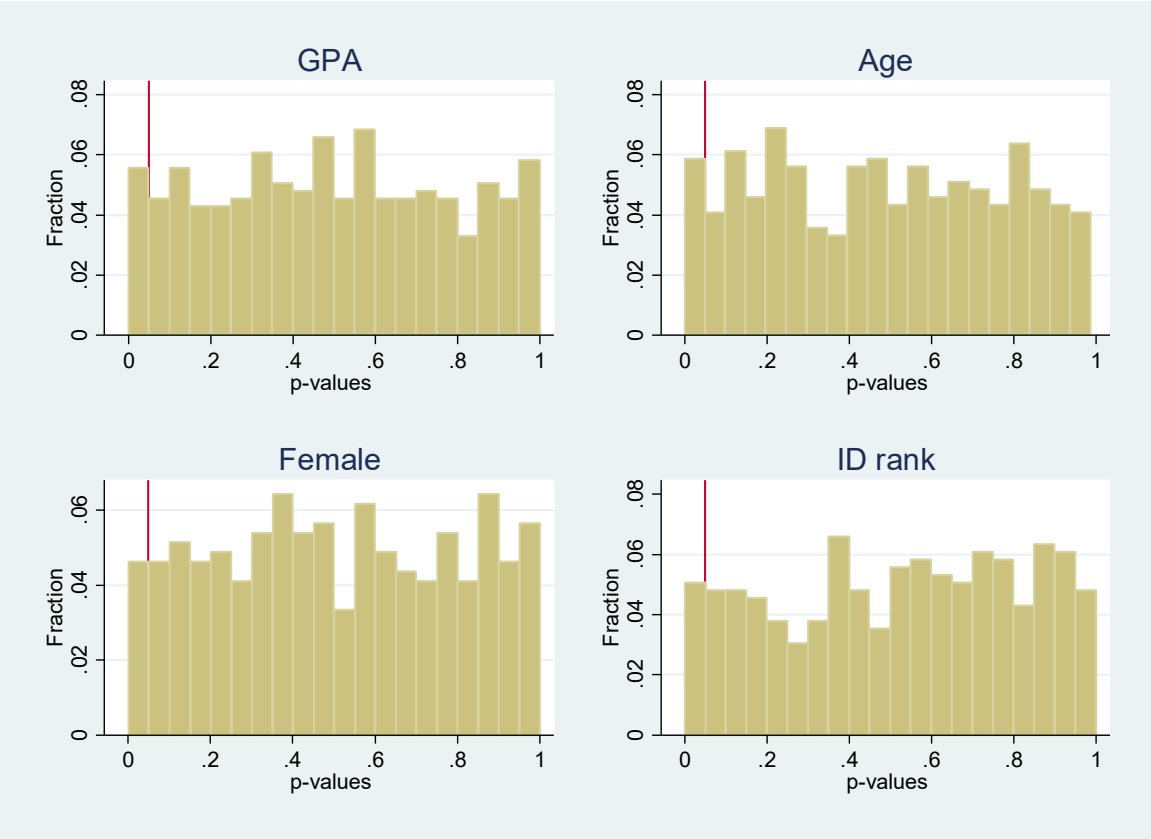
Our randomization tests consist of running, for each course, the following regression:

$$y_i = \alpha + \boldsymbol{T\beta} + \boldsymbol{Z\gamma} + \varepsilon_i \qquad \text{(A12)}$$

Under the null-hypothesis of (conditionally) random assignment to sections within each course, $\boldsymbol{\beta} = \boldsymbol{0}$, which means that the section assignment does not systematically relate to students' pre-treatment characteristics, holding constant scheduling and stratification indicators. Therefore, we expect the F-test to be significant at the 5 percent level in approximately 5 percent of the cases, at the 1 percent level in approximately 1 percent of the cases, and at the 0.1 percent level in approximately 0.1 percent of the cases. Table 3 in Section 6 shows that the actual rejection rates are close to the rejection rates expected under random assignment.

To investigate this issue more closely, we also consider the distribution of p-values. Under the null hypothesis of conditionally random assignment, we would expect the p-values of all the regressions to closely fit a U[0,1] uniform distribution with a mean of 0.5 (Murdoch et al., 2008). Figure A2 shows histograms of the p-values of all four specifications, all of which are roughly uniformly distributed. Column (2) of Table A1 shows the mean of the p-values over all regressions reported in Table 3. The mean of the p-values ranges from 0.48 to 0.52.

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



Note: These are histograms with p-values from all the regressions reported in Table 3. The vertical line in each histogram shows the 0.05 significance level.

Table A1: Randomization Check: Mean p-values

Dependent Variable:	(1) Total Number of Courses	(2) Mean of p-value
GPA	395	0.493
Age	392	0.481
Gender	389	0.508
ID rank	395	0.523

Note: This table is based on the regressions reported in Table 3. Column (2) shows the means of the p-values.

Table A2: Using first year GPA as Measure of Own and Peer Ability

	(1)
	Std. Grade
Standardized first year peer GPA	0.0198**
	(0.008)
Standardized first year GPA	0.6506***
	(0.027)
Observations	12,046
R-squared	0.569
Course FE	YES
Staff FE	YES
Other course FE	YES

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

Table A3: Determinants of Survey Response (OLS)

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

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

Table A4: Separate Regressions for Each Course Evaluation Question

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

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

Table A5: The Effect of Peer Composition on Student Evaluations with Lower and Upper Missing Values Bounds

Panel A:

	(1)	(2)	(3)	(4)	(5)	(6)
	Std. Group interaction		**Self-study hours**		**Std. Teacher evaluation**	
	lower bound	upper bound	lower bound	upper bound	lower bound	upper bound
Standardized peer GPA	0.0544***	0.0562***	-0.0804	0.0665	-0.0003	0.0004
	(0.015)	(0.014)	(0.287)	(0.101)	(0.013)	(0.013)
Standardized GPA	-0.0353***	-0.0340***	-1.8081***	0.4489***	-0.0449***	-0.0400***
	(0.010)	(0.010)	(0.231)	(0.083)	(0.009)	(0.008)
Observations	15,441	15,441	15,441	15,441	15,441	15,441
R-squared	0.222	0.242	0.122	0.226	0.384	0.411

Panel B:

	(1)	(2)	(3)	(4)	(5)	(6)
High GPA * Fraction of high-GPA peers	0.5119***	0.5247***	-0.4185	-0.7356	0.0873	0.0913
	(0.129)	(0.131)	(0.907)	(2.292)	(0.094)	(0.096)
High GPA * Fraction of low-GPA peers	-0.1314	-0.1059	-1.4862*	1.3448	-0.0863	-0.0728
	(0.116)	(0.121)	(0.873)	(2.198)	(0.094)	(0.100)
Middle GPA * Fraction of high-GPA peers	0.2066	0.2153	-0.7174	3.9182	-0.0445	-0.0593
	(0.136)	(0.139)	(1.076)	(2.569)	(0.120)	(0.125)
Middle GPA * Fraction of low-GPA peers	0.0208	0.0215	1.4074	5.3847**	0.0306	0.0104
	(0.110)	(0.110)	(1.079)	(2.368)	(0.111)	(0.114)
Low GPA * Fraction of high-GPA peers	0.2688**	0.2681**	-0.4423	-0.1719	0.0315	0.0754
	(0.116)	(0.118)	(1.013)	(3.282)	(0.116)	(0.113)
Low GPA * Fraction of low-GPA peers	0.0366	0.0461	-1.1360	-4.2525	-0.0825	-0.0775
	(0.118)	(0.120)	(1.056)	(2.989)	(0.114)	(0.115)
Observations	15,441	15,441	15,441	15,441	15,441	15,441
R-squared	0.244	0.224	0.227	0.126	0.411	0.385

Note: Robust standard errors clustered at the course-year-period level are in parentheses. The dependent variables are standardized Group interaction in Columns (1) and (2), self-study hours in Columns (3) and (4), and standardized teacher evaluation in Columns (5) and (6). We assume extreme values for missing answers, which means that for all items we assign the lowest possible value of the used answering scale for our lower bound estimates and the highest possible answer on the answering scale for the upper bound estimates. All specifications include course fixed effects, other-course fixed effects, teacher fixed effects as well dummies for day of the session, time of the session, German, Dutch, exchange student status and late registration status. All regressions reported in Panel B also include Std. GPA. For a list of the exact question wording, see Table 2 in Section 3.2. *** p<0.01, ** p<0.05, * p<0.1.

Table A6: The Effect of Peer Composition on Student Evaluations (Without Bounding or Imputations)

Panel A

	(1) Std. Group functioning	(2) Self-study hours	(3) Std. Teacher functioning
Standardized peer GPA	0.0550***	0.0433	0.0010
	(0.015)	(0.092)	(0.013)
Standardized GPA	-0.0358***	0.0530	-0.0466***
	(0.010)	(0.088)	(0.009)
Observations	15,285	15,232	14,654
	0.222	0.267	0.387

Panel B

	(1)	(2)	(3)
High GPA * Fraction of high-GPA peers	0.5297***	-0.7109	0.0820
	(0.131)	(0.799)	(0.099)
High GPA * Fraction of low-GPA peers	-0.1180	-0.8725	-0.0830
	(0.120)	(0.766)	(0.101)
Middle GPA * Fraction of high-GPA peers	0.1881	0.0058	-0.0492
	(0.141)	(0.991)	(0.127)
Middle GPA * Fraction of low-GPA peers	0.0230	1.7851*	0.0410
	(0.112)	(0.946)	(0.118)
Low GPA * Fraction of high-GPA peers	0.2727**	-0.2952	0.0609
	(0.119)	(0.978)	(0.119)
Low GPA * Fraction of low-GPA peers	0.0454	-1.4976	-0.0665
	(0.121)	(0.967)	(0.120)
Observations	15,285	15,232	14,654
F joined significance of peer variables	0.224	0.268	0.388
Prob > F	<.0001	.2169	0.5226

Note: All regressions include fixed effects for the course, fixed effects for the other course taken at the same time and teacher fixed effects. Robust standard errors clustered at the course-year-period level are in parentheses. The dependent variables are standardized Group interaction in Column (1), self-study hours in Column (2) and standardized teacher evaluation in Column (3). All specifications include dummies for day of the session, time of the session, German, Dutch, exchange student status and late registration status. *** p<0.01, ** p<0.05, * p<0.1.