Received: 6 July 2019 | Accepted: 19 June 2020

DOI: 10.1111/boer.12274

BULLETIN OF ECONOMIC RESEARCH

ARTICLE

The effect of class size on grades and course evaluations: Evidence from multisection courses

Alexei Karas[1,2]

[1] Social Science Department, University College Roosevelt, Middelburg, The Netherlands

[2] Utrecht School of Economics, Utrecht University, Utrecht, The Netherlands

Correspondence
Alexei Karas, University College Roosevelt, P.O. Box 94, 4330 AB Middelburg, The Netherlands.
Email: a.karas@ucr.nl

Abstract

Using rich administrative data from a small Dutch liberal arts college, I study how the number of students enrolled in a course affects student grades and course evaluations. Exploiting variation across parallel sections of the same course taught by the same instructor, I show that class size has a significant negative effect on student grades in mandatory courses, but not in electives. I show similar results for various components of student course evaluations: perceived overall course quality, perceived amount learned, student participation and engagement. I interpret these findings to be consistent with class size affecting educational outcomes through student engagement.

KEYWORDS
class size, course evaluations, grades, student engagement

JEL CLASSIFICATION
I21, I23

1 | INTRODUCTION

In this paper, I address the question of why college class size matters for educational outcomes. Using rich administrative data from a small Dutch liberal arts college, I study how the number of students enrolled in a course affects student grades and course evaluations. It is already well

established that, on average, bigger classes lead to lower grades (Bandiera, Larcinese, & Rasul, 2010; De Giorgi, Pellizzari, & Woolston, 2012; Kokkelenberg, Dillon, & Christy, 2008) and lower evaluations (Bedard & Kuhn, 2008; Mandel & Sussmuth, 2011; Sapelli & Illanes, 2016). What I seek to learn here is whether this effect systematically varies across courses with different characteristics.

Understanding this variation is important. Knowing which courses derive the most benefit from smaller classes allows institutions to more optimally allocate scarce resources. As shown by, for example, Bound and Turner (2007) and Bound, Lovenheim, and Turner (2010), these resources represent an important constraint on student success in college.

Specifically, I explore whether class size matters more in mandatory courses compared to electives. I find that the answer is "yes." Exploiting variation in class size across parallel sections of the same course taught by the same instructor, I show that class size has a significant negative effect on student grades in mandatory courses, but not in electives. I show similar results for various components of student course evaluations: perceived overall course quality, perceived amount learned, student participation and engagement.

I motivate this empirical exercise as a test of a specific channel through which class size affects educational outcomes. I refer to it as "engagement channel." It works through student behavior: in smaller classes, students become more visible to the teacher and thus experience a stronger pressure to participate; alternatively, students enjoy a stronger sense of belonging to the group; both visibility and sense of belonging drive their academic and/or social engagement up. This increased engagement, in turn, leads to better educational outcomes (Finn, Pannozzo, & Achilles, 2003).

To test this engagement channel, I compare class size effects in mandatory courses and electives. My rationale is as follows. Students enroll in electives by choice; engagement in those courses must therefore be high regardless of class size.[1] In contrast, students often attend mandatory courses against their will, and instructors have to work hard to engage them. If class size affects student achievement through engagement, its effects should be particularly pronounced in mandatory courses: in those courses engagement starts low, and smaller class size helps instructors to build it.

To identify the causal effect of class size, I use a fixed effects regression. Specifically, I regress final grades and student evaluation scores on the number of students enrolled in a section of a course controlling for time-course-instructor fixed effects. Including these fixed effects amounts to only exploiting variation in class size across parallel sections of the same course-instructor, and thus allows me to rule out many alternative interpretations of the observed correlation. In addition, I rerun the regression for student grades on a subsample of the first-semester students. Because first semesters do not get a chance to choose a specific section within a course, I argue that this regression allows for a particularly credible identification of the class size effect.

This paper contributes to the long-standing debate on the effects of class size on educational outcomes (Hattie, 2005; Mishel, Rothstein, Krueger, Hanushek, & Rice, 2002). It is most closely related to recent econometric studies in higher education: Kokkelenberg et al. (2008), Bandiera et al. (2010), De Giorgi et al. (2012), and De Paola, Ponzo, and Scoppa (2013) estimate a negative

[1] For evidence that the provision of choice raises student engagement, see, for example, a review by Stroet, Opdenakker, and Minnaert (2013). The explanation is offered by the self-determination theory (Deci & Ryan, 2000): provision of choice helps to satisfy one of the students' basic psychological needs, the need for autonomy, and thereby, raises motivation and engagement.

effect of class size on student performance; Bedard and Kuhn (2008), Westerlund (2008), Cheng (2011), Mandel and Sussmuth (2011), Monks and Schmidt (2011), and Sapelli and Illanes (2016) do the same for student course evaluations. My main added value to this literature is the new source of variation I use for identification—variation in class size across parallel sections of the same course-instructor. This identification appears particularly credible for the subsamples of first-semester students because those students have no say in their assignment to a specific section within a course.

My second contribution is to the literature on "why class size matters." I show that the heterogeneity of class size effects across courses is consistent with class size affecting educational outcomes through student engagement. The engagement hypothesis is proposed by Finn et al. (2003); supporting evidence is presented in, for example, Babcock and Betts (2009) and Blatchford, Bassett, and Brown (2011). This literature, however, focuses on pre-university education. I believe mine is the first paper to show that the same mechanisms that explain class size effects at school operate at the university.

My findings are also consistent with Lazear (2001). Lazear builds a theoretical model of educational production with a focus on student disruptive behavior. The model illustrates that schools enrolling students with a higher propensity to disrupt realize a larger benefit from small classes through a greater reduction in the time lost to disruptive behavior. Viewing disruptive behavior as an extreme form of disengagement, the model shows that optimal class size varies with student engagement. If engagement systematically varies across courses (i.e., mandatory vs. electives), so does optimal class size. The heterogeneity of class size effects I find across courses is consistent with this proposition.

2 | INSTITUTIONAL BACKGROUND

I use administrative data from a small liberal arts and sciences college in the Netherlands. The college offers a broad English-taught bachelor program with majors in science, social science, and arts and humanities. The program takes six semesters to complete with students taking four courses per semester. The maximum capacity of the college is 600 students; yearly intake is around 200 students; more than a third come from outside of the Netherlands. The data span 28 semesters from fall 2004 to spring 2018.

Several features of the college play an important role in the empirical analysis that follows. First, attendance in all courses is compulsory. This rule assures that the number of students enrolled in a section of a course, which I use to measure class size, closely approximates the number of students effectively present in the same class.

Second, graduation requirements represent a mixture of mandatory courses and electives. Specifically, to get a bachelor degree students need to

1. complete 24 courses;
2. achieve depth in at least two disciplines;
3. take at least one course in each department (science, social science, and arts and humanities);
4. take a number of mandatory courses.

An overview of mandatory courses is provided in Table 1. Some of these courses have to be taken by all students; others by specific majors. Some (e.g., foreign language) allow for student choice; others do not. The key point here is that students have to take mandatory courses even

TABLE 1 Overview of mandatory courses

Mandatory for	Course
All students	Methods and Statistics I
All students	Academic Writing and Presenting
All students	Writing Across the Disciplines
All students	Foreign language: Dutch, French, German, or Spanish
Arts and humanities majors	Introduction to Rhetoric and Argumentation
Science majors	Mathematical Ideas and Methods in Context
Social science majors	Methods and Statistics II or Qualitative Methods
Students with insufficient English	English for Academic Purposes (abolished in 2015)

if they would prefer not to. I assume that this compulsory nature drives average engagement in mandatory courses below that in electives. If this assumption is true, and holding all else constant, I can attribute differences in class size effect between these two groups of courses to varying student engagement.

Third, students enjoy much freedom in picking courses they would like to take. The procedure for course enrollment works as follows. Around mid-semester, the timetable for the following semester gets published. The timetable reports the list of offered courses, whether the course has multiple parallel sections, who teaches each section, and in which time slot. In a particular semester, a full-time instructor can be assigned to teach at most three course sections; these sections may belong to one course or to different courses. Students select their preferred courses, either on their own or in consultation with their academic advisor. The latter then registers students for the chosen courses in the student information system.

Many factors affect student course choice: preference to work with a certain instructor; interest in the subject matter; need to take a course to qualify for a specific master; desire to spread known-to-be-difficult courses evenly over their study program; desire to spread classes evenly over the week; preference for classes scheduled in certain parts of the day; need to take courses that do not clash, that is, are offered in different time slots. Some of the factors that drive student course choice (and by implication class size) likely correlate with educational outcomes. The standard approach to control for these omitted factors in the literature is to use student, instructor and course fixed effects (Bandiera et al., 2010; De Giorgi et al., 2012; Kokkelenberg et al., 2008). I believe that this approach is not fully satisfactory, particularly not for the college I study.

The reason is that the college has a "teaching first" culture. This culture leads to teaching practices that vary widely within instructors, both over time and across courses. The "teaching first" culture manifests itself in several ways. First, the college uses "excellent teaching" as one of its unique selling points: its website emphasizes small classes, innovative pedagogical approaches, and high level of student–teacher interaction; its history counts top performance on teaching in national rankings and student surveys, labels of excellence from accreditation committees, and prizes for individual teachers. Second, teaching is central in what faculty do: it occupies no less than three quarters of the yearly workload of a full-time faculty member (the rest goes to research and service); it is prominent in annual performance reviews and promotion decisions, in activities organized for faculty development, and even in social activities, such as reading groups, in which faculty read books on teaching together. Third, student feedback is taken seriously: in response, faculty often make substantial changes to their courses. As a result of this "teaching first" culture faculty have strong incentives to innovate and improve. These innovations, in turn,

TABLE 2 Equivalence between letter grades and grade points

Letter	F	D−	D	D+	C−	C	C+	B−	B	B+	A−	A	A+
Grade point	0	0.7	1	1.3	1.7	2	2.3	2.7	3	3.3	3.7	4	4

lead to much variation in teaching practices within instructors. This variation is not adequately captured by the time-invariant fixed effects.[2]

To improve on the existing approach of controlling for omitted factors through time-invariant fixed effects, I exploit two additional features of the institutional environment. First, the college imposes an upper limit on class size. This limit serves to facilitate student-centered teaching, and applies to nearly all courses.[3] It was changed only once in my sample period, from 25 to 26 students in 2017 (with two exceptions allowed to run at 27). To abide by this limit, the college offers courses with large enrollments in multiple parallel sections. These sections differ little: they use the same course outline, the same assignments and assessments, and often are taught by the same instructor. They do however often differ in class size. Exploiting these differences allows me to control for sources of bias not accounted for in prior literature.

Second, the course registration procedure for the first-semester students differs from the one described above. When prospective students apply to college, they indicate their preferred first semester courses on the application form. At that point, applicants only observe the general list of offered courses published on the college's external website; they do not have access to the following semester timetable, which is published on the college's internal website. Before semester starts, the Head of Academic Advising enrolls first semesters in courses trying to accommodate their preferences. If a course has multiple sections, the Head of Advising picks one without consulting the student.

The description above makes clear that student choice plays a lesser role in course enrollments of the first-semester students. Because these students do not observe the following semester timetable, they have no access to information on time slots and instructors. These factors can therefore not drive their course choice. Most importantly, student choice plays no role in her assignment to a specific section within a course. I will use this absence of student choice to further mitigate concerns about omitted variable bias in the estimation of class size effects.

3 | DATA AND VARIABLES

I study the effect of class size on final grades and student course evaluations. Final grades in courses are assigned as letters, from F to $A+$. The Student Office then converts these letters to grade points following Table 2. In the regressions I use grade points, but make one adjustment: to distinguish $A+$ from A I change all grade points that correspond to $A+$ from 4.0 to 4.3.

Throughout the course the instructor assesses students several (typically 4–10) times. This practice is known as "continuous assessment"; it aims to stimulate students' continuous development. Assessment forms vary across courses and include exams, papers, homework assignments,

[2] The effects I describe here are in line with recent evidence that teacher effectiveness varies within instructors. In particular, Harris and Sass (2011) and Papay and Kraft (2015) show that teachers improve with experience, and that these improvements vary across disciplines.

[3] Courses deviating from this general rule are excluded from the analysis. These are performance courses, in which teaching happens one on one, as well as labs and capstone courses, in which the upper limit is set at 15.

presentations, portfolios, class participation, research projects, and more. The methods, timing, and contribution of each assessment to the final grade are described in the course outline and shared with students. No single assessment can count for more than 40% of the final grade.

Final grades are not curved; and there are no grade targets. The first fact assures grades reflect each student's absolute performance on the course. The second fact allows for grades to vary substantially across courses. On one end of the distribution one can find courses that only give *A*s. On the other end some courses give *C*− as an average grade.

Student course evaluations are collected as follows. At the end of every semester, in class, all (attending) students anonymously fill in a questionnaire with several multiple choice and open questions. A student volunteer then brings all questionnaires to the Student Office. The forms are scanned and the responses are digitized.

Course evaluation data are available for 18 semesters from fall 2009 to spring 2018. Throughout this period, the college used two versions of the questionnaire. The old form (see Table A.4 in Supporting Information Appendix A) was used over six semesters (from fall 2009 to spring 2012). The new form (see Table A.5 in Supporting Information Appendix A) was used over 12 semesters (from fall 2012 to spring 2018). It is not possible to match all items from the new form to items in the old form. For this reason, for the main part of my empirical analysis I only use data from the new form. In Section 6, however, I check whether the results are robust to combining some items from both forms.

I analyze student responses to the following multiple choice questions (response options are in parentheses):

1. Q1. I learned a great deal (Strongly disagree, Disagree, Neither agree nor disagree, Agree, Strongly agree).
2. Q2. Active student participation was encouraged (Same as above).
3. Q3. My critical thinking was stimulated (Same as above).
4. Q4. The requirements were clear (Same as above).
5. Q5. The grading criteria were clear (Same as above).
6. Q6. I was provided with feedback on my individual work (Same as above).
7. Q7. The instructor was available for individual guidance (Same as above).
8. Q8. How actively engaged were you in this course? (Very little, Little, Average, Much, Very much)
9. Q9. Approx. how many hours per week did you devote to this class (NOT including class time)? (<5 hr, 5–8 hr, 8–12 hr, 12–15 hr, >15 hr)
10. Q11. In your opinion the overall quality of this course is (Very bad, Bad, Neutral, Good, Very good).

All these questions have five ordered response options. For the empirical analysis I convert these responses into numerical scores from 1 to 5.

To start, I analyze two questions that describe students' general experience with the course. These are questions on overall course quality (Q11) and perceived amount learned (Q1). The former reflects students' general satisfaction with the course, and can be viewed as an educational outcome interesting in its own right; it has been studied by, for example, Bedard and Kuhn (2008) and Cheng (2011). The latter can be viewed as an alternative measure of study achievement, complementary to grades.

I then continue to analyze student responses to all remaining questions (except Q10 on the expected final grade). I split these questions into two groups. The first relates to teacher

TABLE 3 Summary statistics

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Electives					Mandatory				
Variables	***N***	**Mean**	**SD**	**Min**	**Max**	**I**	**Mean**	**SD**	**Min**	**Max**
Grade point	33,342	3.28	0.90	0	4.30	14,721	3.12	0.95	0	4.30
Score on Q1	15,068	4.18	0.81	1	5	6371	3.77	0.91	1	5
Score on Q2	15,028	4.05	0.86	1	5	6364	4.04	0.81	1	5
Score on Q3	15,014	4.02	0.89	1	5	6254	3.45	0.97	1	5
Score on Q4	15,007	3.86	0.97	1	5	6365	3.79	0.97	1	5
Score on Q5	14,916	3.76	0.99	1	5	6313	3.79	0.96	1	5
Score on Q6	14,889	3.97	0.91	1	5	6342	4.06	0.89	1	5
Score on Q7	14,394	4.19	0.79	1	5	6158	4.17	0.78	1	5
Score on Q8	15,064	3.57	0.89	1	5	6377	3.42	0.89	1	5
Score on Q9	15,063	2.35	0.91	1	5	6382	2.14	0.94	1	5
Score on Q11	15,045	4.09	0.82	1	5	6360	3.80	0.84	1	5
Class size	33,782	20.9	4.71	1	27	14,943	21.5	4.01	1	26
% Female	33,782	65.2	16.3	0	100	14,943	65.2	11.6	0	100
% Dutch	33,782	69.6	13.2	0	100	14,943	67.3	23.4	0	100
Mean of age	33,782	20.5	0.79	18.3	22.9	14,943	20.0	0.89	18.3	23.2
SD of age	33,780	1.39	0.48	0	5.69	14,930	1.36	0.55	0.50	7.81

behavior: these are questions on the clarity of requirements (Q4) and grading criteria (Q5), provided feedback (Q6), and instructor availability (Q7). The second relates to student behavior: these are questions on participation (Q2), engagement(Q8), critical thinking (Q3), and time on task (Q9). By analyzing these questions I hope to unpack whose behavior is more likely to drive my main results.

In addition to class size, grades, and course evaluations, I use information on student personal characteristics. Specifically, for every course section I compute the percentage of female students, $\%Female$; the percentage of Dutch students, $\%Dutch$; average age; and its standard deviation.

Table 3 reports summary statistics for electives and for mandatory courses. The sample includes 2672 students, 173 instructors, 313 courses, and 2448 course sections, and covers 28 semesters from fall 2004 to spring 2018. Course evaluation data cover 12 semesters (from fall 2012 to spring 2018). Missing grades represent cases when students withdraw from the course for medical reasons. Missing evaluation scores represent cases when students don't answer a particular question, or when they fill in their answer in a manner that a machine cannot read.

On average, electives report higher final grades compared to mandatory courses. Similarly, electives mostly report higher scores on course evaluations. Some of these differences are substantial: for example, average scores on Q1 "I learned a great deal" are 4.18 versus 3.77. Other differences are trivial: for example, average scores on Q2 "Active student participation was encouraged" are 4.05 versus 4.04. The average class size is a little higher in mandatory courses: 21.5 versus 20.9. Controls for gender and age distribution are, on average, very similar. The somewhat lower percentage of Dutch students in mandatory courses has to do with non-Dutch students being obliged to take Dutch language as an additional required course.

Figure 1 depicts class size distribution across all course sections offered in the sample period. Mandatory courses and electives are shown separately. The two distributions are fairly similar.

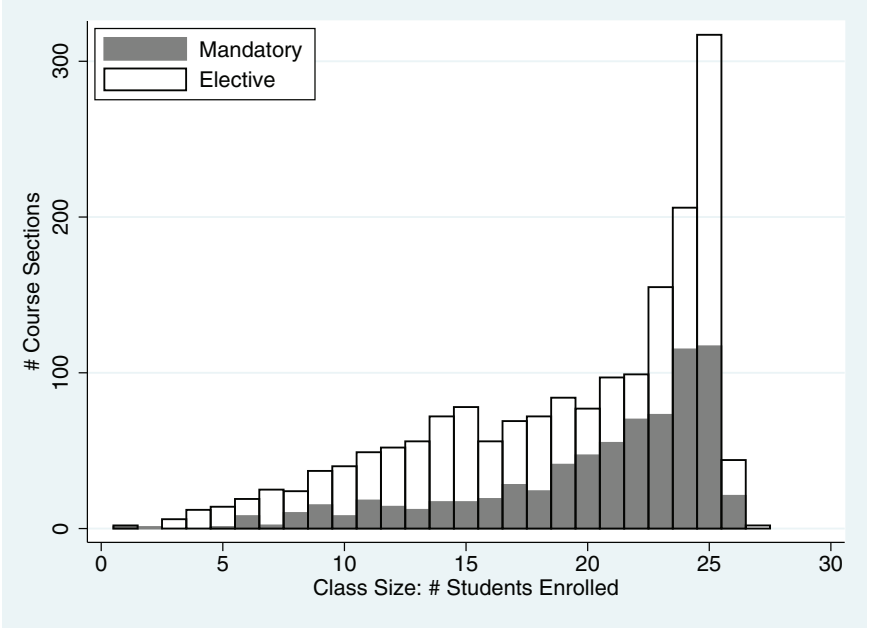

FIGURE 1 Distribution of class size [Colour figure can be viewed at wileyonlinelibrary.com]

Both demonstrate substantial variation in class size: while most sections run at 20–26 students, about a third run at 10–19, and some have fewer than 10 students.

Most of my empirical analyses exploit variation in class size across parallel sections of the same course-instructor. This variation is shown in Figure 2. For each combination of semester t course c and instructor i with multiple parallel sections (each denoted by g) I compute average class size, $\overline{N_{tci}}$. I then subtract average class size $\overline{N_{tci}}$ from the original class size N_{tcig}. Figure 2 plots the distribution of this demeaned class size, $N_{tcig} - \overline{N_{tci}}$. Over the sample period, there have been 249 parallel sections of mandatory courses, and 102 of electives.[4] The number of such sections per course-instructor varies between two and three. We can see that the class size differences between parallel sections of the same course are often small to moderate, ranging between zero and four students. An example would be two sections running at 22 and 26 students, which translates in -2 and $+2$ students in Figure 2. But sometimes the differences can be as high as six students or more.[5]

The modest within-semester-course-instructor variation in class size shown in Figure 2 may make it hard to precisely estimate the effect of interest. However, the range of class sizes I study (see Figure 1) likely helps: it is precisely this range that has been found to matter in pre-university literature. For example, the famous Project STAR compared small classes of 13–17 students with

[4] The frequency distribution of these sections across courses is presented in Tables A.1 and A.2 in Supporting Information Appendix A.

[5] I check whether student body characteristics differ between parallel sections with a relatively small number of students (i.e., those on the left of zero in Figure 2) and those with a relatively large number of students (i.e., those on the right of zero in Figure 2). Specifically, I perform a mean comparison test for $\%Female$, $\%Dutch$, average age, and its standard deviation. I perform this test for parallel sections of all courses, as well as for each course type separately (mandatory vs. electives). In all cases I find small, statistically insignificant differences.

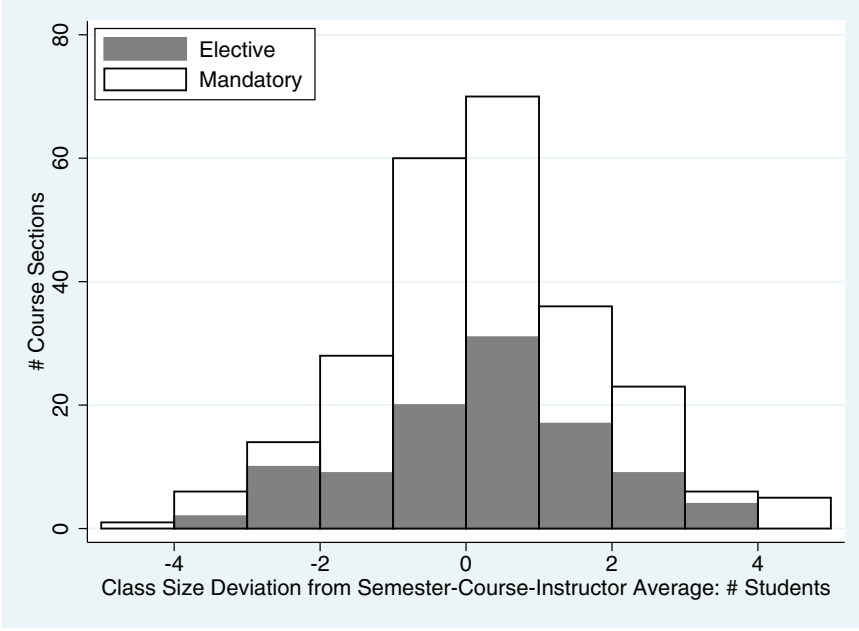

FIGURE 2 Within-semester-course-instructor variation in class size [Colour figure can be viewed at wileyonlinelibrary.com]

regular classes of 22–26 students, and found substantial effects (Finn & Achilles, 1999; Krueger & Whitmore, 2001). Similarly, Blatchford, Bassett, Goldstein, and Martin (2003) mention the number 25 as a threshold below which class size effects become particularly visible.

4 | METHODOLOGY

To study the effect of class size on student grades I estimate the following equation:

$$Y_{tcigs} = \beta_1^E D^E N_{tcig} + \beta_1^M D^M N_{tcig} + FE + Controls_{tcig} + e_{tcigs}. \quad (1)$$

Y_{tcigs} is the final grade of student s enrolled in section g of course c in semester t taught by instructor i. N_{tcig} is the number of students enrolled in section g of course c in semester t taught by instructor i. Dummy D^E equals 1 for electives. Dummy D^M equals 1 for mandatory courses. FE stands for various levels of fixed effects. All estimations are performed using Stata routine *reghdfe* (Correia, 2016). Standard errors are three-way clustered at the level of instructor, course and student.

The coefficients of interest are β_1^E and β_1^M. They measure the ceteris paribus effect of a one-student increase in class size on the final grade in, respectively, electives and mandatory courses. In line with most of the empirical literature, I expect both β_1s to be negative. That is, smaller classes are expected to result in better student performance. My main hypothesis of interest relates to the difference between β_1^M and β_1^E. Specifically, I expect the class size effect to be more pronounced in mandatory courses, that is, $\beta_1^M < \beta_1^E$. This finding would support the engagement channel.

Fixed effects control for unobserved factors that may contaminate the estimates of class size effects. These factors affect student performance as well as their course choice (and by implication class size). I discuss the specific rationale for included fixed effects below.

Student fixed effects, α_s, control for time-invariant student characteristics, for example, ability. More able students may deliberately seek out smaller classes in order to enjoy more quality time one-on-one with the instructor. This mechanism would introduce a negative correlation between student performance and class size, and thus bias β_1s downward.

Semester-course-instructor fixed effects, λ_{tci}, control for unobserved heterogeneity on several levels. To start, they control for time-invariant instructor and course characteristics as well as for college-wide trends and seasonality (e.g., grade inflation). These sources of heterogeneity may introduce a correlation between grades and class size, and thus bias β_1s. For example, instructors or courses may be popular because students get easy *A*s. This mechanism introduces a positive correlation between grades and class size, and thus biases β_1s upward.

Semester-course-instructor fixed effects control for more than the three sources of heterogeneity discussed above. Including these fixed effects amounts to only exploiting variation across parallel sections within a course-instructor. Any change that applies to all such parallel sections is absorbed by the fixed effect. One example is instructor experience. If longer experience leads to better teaching, the same course-instructor may enjoy growing popularity and improving student performance over time. This mechanism would introduce a positive correlation between grades and class size, and thus bias β_1s upward. Other examples include course readings, exam difficulty, teaching approach, and instructor's outfit. So long that these changes apply to all parallel sections of the same course-instructor, they are absorbed by the fixed effect.

In Section 2 I argued that, because of the college's "teaching first" culture, teaching practices are likely to vary substantially within instructors, both over time and across courses. This argument provides a rationale for including semester-course-instructor fixed effects, on top of the standard in the literature separate sets of course, instructor, and time fixed effects. I now show supporting evidence. I test the joint significance of the interacted fixed effects in the presence of separate sets of course, instructor, and time fixed effects. Specifically, in the regression for student grade I include separate sets of time, course, and instructor fixed effects. I then add a set of dummies for all cases of parallel course sections: for each case of two parallel sections I add a dummy for the second section; for each case of three parallel sections I add one dummy for the second section and one for the third. I then use an *F*-test to test the joint significance of these added dummies. The resulting *p*-value is essentially zero; that is, the added dummies are strongly statistically significant. I perform a similar exercise with the regression for overall course quality, and find the same result. That is, *F*-tests support the inclusion of semester, course, and instructor fixed effects interactively, on top of additively.

Two main forces drive variation in class size within semester-course-instructors. First, parallel sections of the same course-instructor are offered in different time slots, some potentially more favored by students than others. For example, students tend to avoid classes in the early morning, as well as on Friday afternoon. If these time slot preferences correlate with student performance, the estimated β_1s would be biased. To alleviate this concern, I add time slot fixed effects, γ_h. The list of time slots is provided in Supporting Information Table A.3.

Second, the college-wide timetable configuration may, accidentally, favor one section over another. For example, a section may be scheduled in the same time slot as some popular courses, and for that reason be avoided by a large number of students. Alternatively, a section may be scheduled on a day that helps a large number of students to achieve a more even distribution of classes over the week, and for that reason be popular. Because the timetable configuration represents an

outcome of dozens of random factors (e.g., availability and time preferences of instructors, year-on-year variation in the number of students, strategic changes in course offerings introduced by the administration, etc.), I argue that this source of variation in class size is as good as exogenous. It is this variation that I use to identify the causal effect of interest.

A final consideration is student body characteristics. These too may differ across course sections in a manner that contaminates the estimation of class size effects. For example, if bigger classes enjoy a more diverse student body, and if diversity affects student performance, the estimated β_1s would be biased. To control for these differences I add to the regressions the following student body characteristics: the percentage of female students, $\%Female$; the percentage of Dutch students, $\%Dutch$; average age; and its standard deviation. These are standard controls employed in prior literature.

I estimate Equation (1) for all students and for the first-semester students only. As argued in Section 2, student choice plays no role in the assignment of first semesters to a specific section within a course. For that reason, restricting the sample to the first-semester students should further alleviate concerns about omitted variable bias.

To study the effect of class size on student course evaluations, I re-estimate Equation (1) replacing grades on the left-hand side with each of the numerical scores reported in Table 3. Because course evaluations are anonymous, these regressions do not include student fixed effects nor clustering at the student level. For the same reason, I cannot run these regressions for first semesters only.

5 | RESULTS

Table 4 reports the main estimation results. The dependent variable varies across specifications and is reported in column headings. Included fixed effects are reported per column. In column 2, the sample is restricted to the first-semester students.

In columns 1 and 2, the dependent variable is student grade. In both specifications, β_1^M is negative and significant (at least at 5%). The negative sign is in line with expectations: bigger classes harm student performance. The average magnitude of β_1^M across the two columns is -0.021. It implies that a one-student increase in class size reduces the average grade in a mandatory course by about 0.02. To compute the corresponding effect size note that the standard deviation of class size in mandatory courses is four students while the standard deviation of grade point is 0.95 (see Table 3). Therefore, a 1 standard deviation increase in class size reduces the average grade in a mandatory course by $0.021 \times 4/0.95 \approx 0.09$ standard deviations. This effect size is very close to the ones reported by Bandiera et al. (2010) and De Giorgi et al. (2012), and is somewhat lower than the average effect size reported in a review by Hattie (2005). As argued in Section 4, the estimate in column 2, which is based on first semesters only, is particularly unlikely to suffer from omitted variable bias. Using this value would increase the estimated effect size to $0.026 \times 4/0.95 \approx 0.11$.

The estimates of β_1^E in columns 1 and 2 are positive; one is significant ($p < 0.05$). This finding is unexpected: it suggests that bigger classes raise student performance in electives. One explanation of the positive class size effect builds on social cognitive learning theory (sometimes referred to as peer effects): a bigger class benefits learning of a particular student by raising the (expected) number of classmates from whom that student can learn, that is, classmates with a similar level of competence. Estimates of positive class size effects are not new in the literature; they have been reported by, for example, Dobbelsteen, Levin, and Oosterbeek (2002) and Denny and Oppedisano (2013).

TABLE 4 Main results

Variables	(1) Grade point	(2) Grade point	(3) Score on Q11	(4) Score on Q1
β_1^E	0.021**	0.017	0.038**	0.031**
	(0.010)	(0.014)	(0.016)	(0.016)
β_1^M	−0.017***	−0.026**	−0.024**	−0.023**
	(0.005)	(0.010)	(0.010)	(0.010)
%Female	−0.001	0.000	−0.001	0.003*
	(0.001)	(0.002)	(0.002)	(0.002)
%Dutch	0.000	−0.001	−0.004*	−0.000
	(0.001)	(0.002)	(0.002)	(0.001)
Mean age	0.026	0.022	−0.052	0.012
	(0.032)	(0.083)	(0.091)	(0.090)
SD of age	−0.002	0.057	0.065	0.030
	(0.028)	(0.049)	(0.063)	(0.065)
Student FE, α_s	Yes	Yes	No	No
Semester-course-instructor FE, λ_{tci}	Yes	Yes	Yes	Yes
Time slot FE, γ_h	Yes	Yes	Yes	Yes
Observations	48,042	9959	21,405	21,439
R^2	0.625	0.765	0.341	0.282
p-value: $\beta_1^E = \beta_1^M$	0.002	0.026	0.000	0.001

Note: $Y_{tcigs} = \beta_1^E D^E N_{tcig} + \beta_1^M D^M N_{tcig} + FE + Controls_{tcig} + e_{tcigs}$; $tcigs$ identifies student s enrolled in section g of course c in semester t taught by instructor i. Y_{tcigs} is reported in column headings . N_{tcig} is class size. Dummy D^E (D^M) equals 1 for electives (respectively, mandatory courses). Fixed effects, FE, are reported per column. Clustered standard errors are in parentheses. The last row reports the p-value of an F-test of the null hypothesis: $\beta_1^E = \beta_1^M$. ***$p < 0.01$; **$p < 0.05$; *$p < 0.1$.

Controls for student characteristics are not significant. The last row reports the p-value of an F-test of the hypothesis of interest: $\beta_1^E = \beta_1^M$. The equality $\beta_1^E = \beta_1^M$ is rejected in both columns ($p < 0.05$). That is, comparing parallel sections of the same course-instructor I find that bigger classes harm student performance in mandatory courses but not in electives. These findings support the engagement channel.

The results reported for all students (column 1) are fairly similar to those for first semesters only (column 2). Recall that restricting the sample to first semesters addresses concerns about unobserved drivers of student course choice that may contaminate the estimates of β_1s in column 1. Because the first-semester students do not get a chance to choose a specific section within a course, such concerns do not apply to the regression reported in column 2. Arguably, this regression provides a particularly credible identification of the class size effect. Importantly, the results support the engagement channel.

Columns 3 and 4 report the results for course evaluation questions on overall course quality (Q11) and amount learned (Q1). The pattern of findings is similar to that reported for final grades in column 1: β_1^M is negative and significant ($p < 0.05$); β_1^E is positive and significant ($p < 0.05$); the equality $\beta_1^E = \beta_1^M$ is rejected ($p < 0.01$). The estimate of β_1^M in column 3 implies that a one-student increase in class size reduces the overall quality score in a mandatory course by 0.024; effect size ≈0.11. This estimate is very close to that reported in prior literature (Cheng, 2011).

TABLE 5 Results: teacher behavior

Variables	(1) Score on Q4	(2) Score on Q5	(3) Score on Q6	(4) Score on Q7
β_1^E	0.018	−0.010	−0.001	−0.001
	(0.018)	(0.019)	(0.019)	(0.026)
β_1^M	−0.008	−0.007	0.006	−0.005
	(0.020)	(0.011)	(0.017)	(0.014)
Student FE, α_s	No	No	No	No
Semester-course-instructor FE, λ_{tci}	Yes	Yes	Yes	Yes
Time slot FE, γ_h	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes
Observations	21,372	21,229	21,231	20,552
R^2	0.282	0.248	0.247	0.213
p-value: $\beta_1^E = \beta_1^M$	0.251	0.864	0.749	0.875

Note: $Y_{tcigs} = \beta_1^E D^E N_{tcig} + \beta_1^M D^M N_{tcig} + FE + Controls_{tcig} + e_{tcigs}$; $tcigs$ identifies student s enrolled in section g of course c in semester t taught by instructor i; Y_{tcigs} is reported in column headings. N_{tcig} is class size. Dummy D^E (D^M) equals 1 for electives (respectively, mandatory courses). Fixed effects, FE, are reported per column. Clustered standard errors are in parentheses. The last row reports the p-value of an F-test of the null hypothesis: $\beta_1^E = \beta_1^M$. ***$p < 0.01$; **$p < 0.05$; *$p < 0.1$.

Two controls for student body characteristics are marginally significant ($p < 0.1$). First, in column 3 a higher proportion of Dutch students leads to a lower score on perceived overall quality (Q11); effect size ≈0.08. Second, in column 4 a higher proportion of female students leads to a higher score on perceived amount learned (Q1); effect size ≈0.05. Other controls are not significant.

Let me summarize and interpret this first set of results from Table 4. In mandatory courses bigger classes harm student performance; in addition, they reduce perceived overall course quality and perceived amount learned. The same is not observed in electives. I interpret these findings to be consistent with student engagement being the channel through which class size affects educational outcomes. Further, the opposite signs of β_1^M and β_1^E suggest optimal class size is smaller in mandatory courses compared to electives. These findings are in line with the theoretical model of Lazear (2001), in which optimal class size is shown to vary across courses as a function of student engagement.[6]

I now turn to the analysis of other items on the course evaluations. Table 5 reports the results for questions that relate to teacher behavior: "Q4. The requirements were clear" in column 1, "Q5. The grading criteria were clear" in column 2, "Q6. I was provided with feedback on my individual work" in column 3, and "Q7. The instructor was available for individual guidance" in column 4. In all columns, I find small and highly insignificant β_1s; the equality $\beta_1^E = \beta_1^M$ is never rejected ($p > 0.1$). These findings suggest that class size does not affect the clarity with which teachers explain requirements and grading criteria. Similarly, class size does not affect the amount of feedback teachers provide, nor their availability for individual guidance. I interpret this latter (somewhat counterintuitive) finding as follows. Instructors who teach multiple parallel sections do not divide their total time available for student guidance equally across *sections*. Instead, they divide this time

[6] Lazear (2001) focuses on disruptive behavior, which can be viewed as an extreme form of disengagement (Finn et al., 2003). He defines optimal class size as class size that equalizes marginal costs and benefits.

TABLE 6 Results: student behavior

Variables	(1) Score on Q2	(2) Score on Q8	(3) Score on Q3	(4) Score on Q9
β_1^E	0.037**	−0.006	0.016	−0.007
	(0.015)	(0.019)	(0.011)	(0.017)
β_1^M	−0.027***	−0.018***	−0.014	−0.007
	(0.008)	(0.005)	(0.014)	(0.009)
Student FE, α_s	No	No	No	No
Semester-course-instructor FE, λ_{tci}	Yes	Yes	Yes	Yes
Time slot FE, γ_h	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes
Observations	21,392	21,441	21,268	21,445
R^2	0.291	0.135	0.325	0.198
p-value: $\beta_1^E = \beta_1^M$	0.000	0.522	0.067	0.971

Note: $Y_{tcigs} = \beta_1^E D^E N_{tcig} + \beta_1^M D^M N_{tcig} + FE + Controls_{tcig} + e_{tcigs}$; $tcigs$ identifies student s enrolled in section g of course c in semester t taught by instructor i. Y_{tcigs} is reported in column headings. N_{tcig} is class size. Dummy D^E (D^M) equals 1 for electives (respectively, mandatory courses). Fixed effects, FE, are reported per column. Clustered standard errors are in parentheses. The last row reports the p-value of an F-test of the null hypothesis: $\beta_1^E = \beta_1^M$. ***$p < 0.01$; **$p < 0.05$; *$p < 0.1$.

equally across *students* from these sections. As a result, a student from a small section does not get more time; a student from a big section does not get less.

To sum up, across parallel sections of the same course-instructor I find no evidence that class size affects teacher behavior, neither in mandatory courses nor in electives. These findings are not in line with the literature. For example, Blatchford, Moriarty, Edmonds, and Martin (2002) and Blatchford, Bassett, and Brown (2005) report evidence that smaller classes do lead to changes in teacher behavior.

Table 6 reports the results for questions that relate to student behavior. Columns 1 and 2 report the results for questions on, respectively, student participation (Q2) and engagement(Q8). In both columns β_1^M is negative, significant, and smaller than β_1^E. β_1^E is significantly positive for student participation (column 1) and insignificant for engagement (column 2). The equality $\beta_1^E = \beta_1^M$ is rejected at 1% in column 1, not in column 2. Columns 3 and 4 report the results for questions on, respectively, critical thinking (Q3), and time on task (Q9). In both columns I find insignificant β_1s. The equality $\beta_1^E = \beta_1^M$ is marginally rejected ($p < 0.1$) in column 3, not in column 4.

To sum up, in mandatory courses bigger classes result in students feeling less encouraged to participate and less engaged. The same is not observed in electives. Effects on the amount of time students devote to the course and on their critical thinking are insignificant. I take these findings to provide further support that student behavior (i.e., participation and engagement) is the channel through which class size affects educational outcomes.

6 | ROBUSTNESS CHECKS

In this section, I describe a number of robustness checks.

First, I rerun the regressions for course evaluations on a longer sample, which combines the new and the old version of the questionnaire. Recall that in the main part of the paper I limited the analysis of course evaluations to the new version of the questionnaire, the one operating

TABLE 7 Robustness: merging course evaluation questionnaires

Variables	(1) Score on Q11	(2) Score on Q1	(3) Score on Q2	(4) Score on Q8
β_1^E	0.008	0.014	0.025**	−0.010
	(0.011)	(0.010)	(0.012)	(0.012)
β_1^M	−0.023***	−0.022*	−0.022***	−0.025***
	(0.008)	(0.012)	(0.005)	(0.005)
Student FE, α_s	No	No	No	No
Semester-course-instructor FE, λ_{tci}	Yes	Yes	Yes	Yes
Time slot FE, γ_h	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes
Observations	31,545	31,614	31,566	31,599
R^2	0.338	0.274	0.276	0.125
p-value: $\beta_1^E = \beta_1^M$	0.015	0.027	0.000	0.172

Note: $Y_{tcigs} = \beta_1^E D^E N_{tcig} + \beta_1^M D^M N_{tcig} + FE + Controls_{tcig} + e_{tcigs}$; $tcigs$ identifies student s enrolled in section g of course c in semester t taught by instructor i. Y_{tcigs} is reported in column headings. N_{tcig} is class size. Dummy D^E (D^M) equals 1 for electives (respectively, mandatory courses). Fixed effects, FE, are reported per column. Clustered standard errors are in parentheses. The last row reports the p-value of an F-test of the null hypothesis: $\beta_1^E = \beta_1^M$. ***$p < 0.01$; **$p < 0.05$; *$p < 0.1$.

TABLE 8 Robustness: dropping electives

Variables	(1) Grade point	(2) Q11	(3) Grade point	(4) Q11	(5) Grade point	(6) Q11
β_1^E	0.014	0.040*	0.021	0.029	0.017*	0.041**
	(0.011)	(0.021)	(0.014)	(0.028)	(0.010)	(0.017)
β_1^M	−0.016***	−0.026***	−0.017***	−0.022**	−0.017***	−0.024**
	(0.005)	(0.009)	(0.005)	(0.009)	(0.005)	(0.010)
Student FE, α_s	Yes	No	Yes	No	Yes	No
S-C-I FE, λ_{tci}	Yes	Yes	Yes	Yes	Yes	Yes
Time slot FE, γ_h	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations	46,934	20,942	47,119	21,001	47,117	21,014
R^2	0.624	0.342	0.624	0.345	0.624	0.339
p-value: $\beta_1^E = \beta_1^M$	0.018	0.001	0.019	0.089	0.004	0.000

Note: $Y_{tcigs} = \beta_1^E D^E N_{tcig} + \beta_1^M D^M N_{tcig} + FE + Controls_{tcig} + e_{tcigs}$; $tcigs$ identifies student s enrolled in section g of course c in semester t taught by instructor i. Y_{tcigs} is reported in column headings. N_{tcig} is class size. Dummy D^E (D^M) equals 1 for electives (respectively, mandatory courses). Fixed effects, FE, are reported per column. Clustered standard errors are in parentheses. The last row reports the p-value of an F-test of the null hypothesis: $\beta_1^E = \beta_1^M$. ***$p < 0.01$; **$p < 0.05$; *$p < 0.1$.

since fall 2012. Some of the items in this new form, however, have a very close match in the old form that operated before fall 2012. Specifically, questions on perceived overall course quality, perceived amount learned, student participation, and engagement are available, with a slight change of language, in both forms (see items 21, 2, 14, 19 in Supporting Information Table A.4 and items 11, 1, 2, 8 in Supporting Information Table A.5). I merge each pair of these matched items into a variable and use the resulting variable as a regressand. The sample now covers 18 semesters from fall 2009 to spring 2018. Table 7 reports the results. Across columns β_1^M is negative, significant

TABLE 9 Robustness: controlling for average ability of peers

Variables	(1) Grade point	(2) Grade point	(3) Score on Q11	(4) Score on Q1
β_1^E	0.025**	0.018	0.035**	0.031*
	(0.011)	(0.017)	(0.015)	(0.016)
β_1^M	−0.016***	−0.034**	−0.024**	−0.024***
	(0.005)	(0.014)	(0.009)	(0.009)
%Female	−0.000	0.001	−0.000	0.002
	(0.001)	(0.002)	(0.002)	(0.002)
%Dutch	0.001	0.002	−0.004*	−0.000
	(0.001)	(0.002)	(0.002)	(0.001)
Mean age	0.040	0.111	−0.047	0.012
	(0.044)	(0.098)	(0.088)	(0.089)
SD of age	−0.004	0.054	0.063	0.022
	(0.038)	(0.047)	(0.062)	(0.062)
Mean success chance	−0.043	0.086	−0.049	0.013
	(0.061)	(0.106)	(0.125)	(0.097)
Student FE, α_s	Yes	Yes	No	No
Semester-course-instructor FE, λ_{tci}	Yes	Yes	Yes	Yes
Time slot FE, γ_h	Yes	Yes	Yes	Yes
Observations	31,948	6,749	20,958	20,991
R^2	0.622	0.766	0.340	0.282
p-value: $\beta_1^E = \beta_1^M$	0.001	0.027	0.000	0.001

Note: $Y_{tcigs} = \beta_1^E D^E N_{tcig} + \beta_1^M D^M N_{tcig} + FE + Controls_{tcig} + e_{tcigs}$; $tcigs$ identifies student s enrolled in section g of course c in semester t taught by instructor i. Y_{tcigs} is reported in column headings. N_{tcig} is class size. Dummy D^E (D^M) equals 1 for electives (respectively, mandatory courses). Fixed effects, FE, are reported per column. Clustered standard errors are in parentheses. The last row reports the p-value of an F-test of the null hypothesis: $\beta_1^E = \beta_1^M$. ***$p < 0.01$; **$p < 0.05$; *$p < 0.1$.

(at least at 10%) and smaller than β_1^E. β_1^E is significantly positive for student participation (column 3) and insignificant elsewhere. The equality $\beta_1^E = \beta_1^M$ is rejected in three out of four specifications ($p < 0.05$). Overall, the results are fairly similar to those reported in section 5 (Tables 4 and 6).

Second, I observe that the frequency distribution of parallel sections across electives is dominated by three courses: Introduction to Psychology, Introduction to Socio-Cultural Anthropology, and Law, Society and Justice (see Supporting Information Table A.1). To check whether my main results are driven by one of those courses I rerun the regressions for student grade and for overall course quality excluding observations corresponding to each of those three courses. Table 8 reports the results. Columns 1 and 2 drop Introduction to Psychology; columns 3 and 4 drop Introduction to Socio-Cultural Anthropology; columns 5 and 6 drop Law, Society and Justice. Across columns β_1^M is negative, significant (at least at 5%) and smaller than β_1^E. β_1^E is positive, twice significant at 10% and once at 5%. The equality $\beta_1^E = \beta_1^M$ is rejected at least at 10%. Overall, the results are again similar to those reported in Section 5 (Table 4).

Third, I rerun the main regressions including an additional control for student body characteristics—the average ability of classmates. To find a suitable control, I tap into the college's admissions data. The admissions process is summarized in Supporting Information Appendix B. As part of this process every student is interviewed by a faculty member. After the interview the

faculty member fills in a standard form, in which she uses a 1-to-5 scale to rate the candidate on several criteria. These criteria include intelligence, social maturity, intellectual curiosity, communication skills, self-confidence, ability to work hard, depth and breadth of intellectual interest, being concerned for others, willingness to contribute to community, international orientation, and chance to succeed at college. To control for ability I use the last of these ratings—chance to succeed at college. I choose this rating for two reasons. First, occupying the last position on the list and being fairly general, it appears to aim at a holistic evaluation of whether the student is able to do well at college. Second, in internal communication the college has shown that this rating correlates with student subsequent performance. For every student in every course section, I calculate the average college success chance rating of her classmates. I then add this average rating as a control in the regressions of Table 4. Table 9 reports the results. The sample size is down because admission data are not available, in digital form, before spring 2009. In other respects, the results are similar to those reported in Table 4. Controls, including college success chance, are insignificant.

To sum up, the main results survive the three robustness checks performed in this section. Additional results from less demanding specifications (i.e., without semester-course-instructor fixed effects) can be found in the working paper version of this manuscript (Karas, 2019).

7 | CONCLUSION

In this paper, I use administrative data from a small Dutch liberal arts college to study the effect of class size on student grades and course evaluations. My identification strategy exploits variation in class size across parallel sections of the same course taught by the same instructor. I show that class size has a negative effect on student grades in mandatory courses but not in electives. I show similar results for various components of student course evaluations: perceived overall course quality, perceived amount learned, student participation and engagement.

I interpret these findings to support the engagement hypothesis (Finn et al., 2003). Because class size works as a substitute for engagement, its effects should be particularly pronounced in courses with low engagement. That is exactly what I find: in mandatory courses, which students often take because they have to, class size has a negative effect on educational outcomes; in electives it does not. The takeaway for policy is clear: holding all else constant, class size reductions should prioritize mandatory courses (or other courses with low student engagement). And vice versa, class size increases would likely do less harm in electives.

In addition, my results help us to better understand why the estimated size (and sign) of the class size effect varies in the literature. While existing studies emphasize student characteristics as a potential explanation, this paper suggests an alternative–course characteristics.

ACKNOWLEDGMENTS

I thank Remy Rikers for useful comments.

ORCID

Alexei Karas https://orcid.org/0000-0002-6149-1208

REFERENCES

Babcock, P., & Betts, J. R. (2009). Reduced-class distinctions: Effort, ability, and the education production function. *Journal of Urban Economics*, *65*(3), 314–322.

Bandiera, O., Larcinese, V., & Rasul, I. (2010). Heterogeneous class size effects: New evidence from a panel of university students. *Economic Journal*, *120*(549), 1365–1398.

Bedard, K., & Kuhn, P. (2008). Where class size really matters: Class size and student ratings of instructor effectiveness. *Economics of Education Review*, *27*(3), 253–265.

Blatchford, P., Bassett, P., & Brown, P. (2005). Teachers' and pupils' behavior in large and small classes: A systematic observation study of pupils aged 10 and 11 years. *Journal of Educational Psychology*, *97*(3), 454–467.

Blatchford, P., Bassett, P., & Brown, P. (2011). Examining the effect of class size on classroom engagement and teacher–pupil interaction: Differences in relation to pupil prior attainment and primary vs. secondary schools. *Learning and Instruction*, *21*(6), 715–730.

Blatchford, P., Bassett, P., Goldstein, H., & Martin, C. (2003). Are class size differences related to pupils' educational progress and classroom processes? Findings from the institute of education class size study of children aged 5–7 years. *British Educational Research Journal*, *29*(5), 709–730.

Blatchford, P., Moriarty, V., Edmonds, S., & Martin, C. (2002). Relationships between class size and teaching: A multimethod analysis of English infant schools. *American Educational Research Journal*, *39*(1), 101–132.

Bound, J., Lovenheim, M. F., & Turner, S. (2010). Why have college completion rates declined? An analysis of changing student preparation and collegiate resources. *American Economic Journal: Applied Economics*, *2*(3), 129–157.

Bound, J., & Turner, S. (2007). Cohort crowding: How resources affect collegiate attainment. *Journal of Public Economics*, *91*(5–6), 877–899.

Cheng, D. A. (2011). Effects of class size on alternative educational outcomes across disciplines. *Economics of Education Review*, *30*(5), 980–990.

Correia, S. (2016). REGHDFE: Stata module to perform linear or instrumental-variable regression absorbing any number of high-dimensional fixed effects. *Statistical software components*. Chestnut Hill, MA: Boston College.

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

De Paola, M., Ponzo, M., & Scoppa, V. (2013). Class size effects on student achievement: Heterogeneity across abilities and fields. *Education Economics*, *21*(2), 135–153.

Deci, E. L., & Ryan, R. M. (2000). The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry*, *11*, 227–268.

Denny, K., & Oppedisano, V. (2013). The surprising effect of larger class sizes: Evidence using two identification strategies. *Labour Economics*, *23*, 57–65.

Dobbelsteen, S., Levin, J., & Oosterbeek, H. (2002). The causal effect of class size on scholastic achievement: Distinguishing the pure class size effect from the effect of changes in class composition. *Oxford Bulletin of Economics and Statistics*, *64*(1), 17–38.

Finn, J. D., & Achilles, C. M. (1999). Tennessee's class size study: Findings, implications, misconceptions. *Educational Evaluation and Policy Analysis*, *21*(2), 97–109.

Finn, J. D., Pannozzo, G. M., & Achilles, C. M. (2003). The why's of class size: Student behavior in small classes. *Review of Educational Research*, *73*(3), 321–368.

Harris, D. N., & Sass, T. R. (2011). Teacher training, teacher quality and student achievement. *Journal of Public Economics*, *95*(7-8), 798–812.

Hattie, J. (2005). The paradox of reducing class size and improving learning outcomes. *International Journal of Educational Research*, *43*(6), 387–425.

Karas, A. (2019). The effect of class size on grades and course evaluations: Evidence from multi-section courses. *USE Working Paper series*, *19*(3).

Kokkelenberg, E. C., Dillon, M., & Christy, S. M. (2008). The effects of class size on student grades at a public university. *Economics of Education Review*, *27*(2), 221–233.

Krueger, A. B., & Whitmore, D. M. (2001). The effect of attending a small class in the early grades on college-test taking and middle school test results: Evidence from Project STAR. *Economic Journal*, *111*(468), 1–28.

Lazear, E. P. (2001). Educational production. *Quarterly Journal of Economics*, *116*, 777–803.

Mandel, P., & Sussmuth, B. (2011). Size matters. The relevance and Hicksian surplus of preferred college class size. *Economics of Education Review*, *30*(5), 1073–1084.

Mishel, L. R., & Rothstein, R. (2002). *The class size debate*. Krueger, A. B., Hanushek, E. A. E. A., & Rice, J. K. (Eds.). Washington, DC: Economic Policy Institute.

Monks, J., & Schmidt, R. M. (2011). The impact of class size on outcomes in higher education. *B.E. Journal of Economic Analysis & Policy*, *11*(1). Retrieved from https://bepress.com/

Papay, J. P., & Kraft, M. A. (2015). Productivity returns to experience in the teacher labor market: Methodological challenges and new evidence on long-term career improvement. *Journal of Public Economics*, *130*, 105–119.

Sapelli, C., & Illanes, G. (2016). Class size and teacher effects in higher education. *Economics of Education Review*, *52*, 19–28.

Stroet, K., Opdenakker, M.-C., & Minnaert, A. (2013). Effects of need supportive teaching on early adolescents' motivation and engagement: A review of the literature. *Educational Research Review*, *9*, 65–87.

Westerlund, J. (2008). Class size and student evaluations in Sweden. *Education Economics*, *16*(1), 19–28.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Karas A. The effect of class size on grades and course evaluations: Evidence from multisection courses IthankRemy Rikers for useful comments. *Bull Econ Res*. 2021;*73*:624–642. https://doi.org/10.1111/boer.12274