Economics of Education Review 000 (2016) 1–10

Contents lists available at ScienceDirect

Economics of Education Review

journal homepage: www.elsevier.com/locate/econedurev

Class size and teacher effects in higher education

Claudio Sapelli[a,*], Gastón Illanes[b]

[a] Economics Department, Pontificia Universidad Catolica de Chile, Chile
[b] Economics Department, MIT, USA

ARTICLE INFO

Article history:
Received 23 October 2013
Revised 3 November 2015
Accepted 4 January 2016
Available online xxx

JEL Classification:
I21
I23
I28

Keywords:
Class size
Teacher effects
Student evaluations

ABSTRACT

Using student evaluations of their instructor as an outcome measure, we estimate and compare class size and teacher effects for higher education, with an emphasis on determining whether a comprehensive class size reduction policy that draws on the hiring of new teachers is likely to improve educational outcomes. We find that first time teachers perform significantly worse than their peers, and we find substantial class size effects. Hence higher education institutions face a tradeoff if they wish to increase admission. This tradeoff implies that as class size increases, at first the negative class size effect is smaller than that of introducing a first time teacher. However, beyond a certain level, the class size effect dominates and it is better to create a new class with a first time teacher.[1]

1. Introduction

Several studies have estimated the effect of class size on learning outcomes, highlighting that smaller classes foster learning. However, when recommending smaller classes as a policy, it is often forgotten that the teachers hired to work in those classes may not be of the same quality as those currently teaching. Thus, the effect of reducing class size on outcomes will depend crucially on the balance between the positive effect of a smaller class and the potentially negative effect of the quality gap between inframarginal and marginal teachers. This work gives insights for higher education on the decision whether to increase class size with existing teachers or hire new (first time) teachers. We provide evidence on class size and first time teacher effects, using teacher evaluation surveys from the Economics Department and the Business School at Pontificia Universidad Católica of Chile (FACEAPUC). First time teacher effects are relevant for this discussion because the most likely avenue for an increase in the number of teachers in Higher Education is hiring first time teachers.

We use student evaluation data as an outcome measure. It can be thought of as an indicator of student learning or an indicator of student satisfaction. Although interpreting student evaluations as an indicator of learning has its problems (see Braga, Paccagnella, & Pellizzari, 2014; Carrell & West, 2010), this method also has distinct advantages over other output measures for evaluating teachers, such as test scores. Hanushek (2003) and Krueger (2003) argue that estimating the effect of class size on learning using test scores raises major concerns, since results are sensitive to the econometric specification used and to the outcome variable in question. Also, there is research linking student satisfaction to effective learning (Theall & Franklin, 2001),

* Corresponding author at: Vicuña Mackenna 4860, Santiago, Chile. Tel.: +562 23544003.
E-mail addresses: csapelli@uc.cl (C. Sapelli), gastonillanes@gmail.com (G. Illanes).

[1] We would like to thank Matías Covarrubias and Fernanda Rojas for excellent research assistance. We would also like to thank comments received in the internal workshop of Pontificia Universidad Católica de Chile's Economics Department and in the 2011 Yearly Congress of the Economics Society of Chile. We also thank two anonymous referees whose advice greatly improved the paper. The usual disclaimer applies.

http://dx.doi.org/10.1016/j.econedurev.2016.01.001
0272-7757/

and research on student evaluations that provides evidence that student ratings are reliable, valid, unbiased, and useful (Murray, 1994). Finally, Bedard and Kuhn (2008) build on this, arguing that student evaluations are better indicators of student learning. We build on this research by using a FRDD methodology to identify causal links, taking advantage of a discontinuity in class size we observe in our data.

We find that there is a negative effect of increasing class size by one standard deviation of roughly 0.187 SDs of our outcome measure. This is similar in size to the lower bound of those found in the literature (Hanushek & Rivkin, 2010). We also find that the average impact of a first time teacher is −0.41 standard deviations. That is, a first time teacher is substantially worse than infra-marginal teachers. As we will show, there is also substantial risk in hiring a new teacher. In higher education, where teaching loads for full time professors are not flexible, administrators often face the decision of increasing class size or hiring a first time teacher. We find that both choices entail a drop in student satisfaction, and hence that the decision rule would imply increasing the class size up to a certain level and then splitting the class and hiring a first time teacher. We give evidence on the magnitude of these effects and discuss how we infer decisions are taken in this context, particularly considering that administrators face uncertainty.

The paper proceeds as follows: Section 2 summarizes the relevant literature on the education production function and on student evaluations, Section 3 presents our data, Section 4 explains the econometric methodology used, Section 5 presents our results, and Section 6 concludes.

2. Literature review

Studies that estimate teachers' effects on achievement using longitudinal data, such as Rockoff (2004), have become a first step in solving many puzzles in the production function of achievement. Estimates suggest that the best teacher may raise achievement by as much as half a standard deviation. Though this literature also finds that credentials do not explain teacher effects for the most part, the exception is that very inexperienced teachers have worse effects, and that the effects of increased experience plateau after four to five years.

This finding has led to the need to measure teacher effects and class size effects and trade off one against the other. If we are to go by the median estimate in the literature then teacher effects are between two times and six times larger than class size effects. Though results in the literature vary with methodology and data set (see Meghir and Rivkin, 2011 for a thorough treatment), there is an emerging consensus regarding the great heterogeneity of teacher quality and its importance. It is in this area of the literature that we wish to contribute.

The most influential studies of class size reduction are those based on the Student Teacher Achievement Ratio, or STAR, a study conducted in Tennessee in the late 1980s. Among them possibly Krueger's (2003) analysis is the most cited one. He finds that elementary school students randomly assigned to small classes outperformed their classmates assigned to regular classes by about 0.22 standard deviations after four years. Other credible studies that also find positive effects of class size reduction find smaller effects. For example, Rivkin, Hanushek, and Kain (2005) examine the effects of natural variation in class size in Texas in the mid-1990s. The estimated effects were about half the size of the effects found in Krueger (2003). International studies also provide positive evidence for the effects of class-size reduction. Angrist and Lavy (1999) take advantage of a class-size limit in Israel of 40 students. They find positive effects of smaller classes, with effect sizes that are on the lower end of those found in the STAR study. Jepsen and Rivkin (2009) examine the class size reduction enacted in 1996 in California. The program was designed to reduce class size by ten students per class, from 30 to 20. They also find positive effects for class-size reduction that are about half as large as those found in Tennessee. Interestingly Jepsen & Rivkin (2009) study also the changes in the teachers required by this change. They find that increases in the numbers of new and not-fully-certified teachers offset much of these gains. In other words, students who ended up in the classrooms of teachers new to their classrooms and grades suffered academically from the teacher's inexperience by almost the same amount as they benefited from being in a smaller class. Summarizing, it appears that large class-size reductions, on the order of magnitude of 7–10 fewer students per class, can have important long-term effects on student achievement. The largest estimates of the magnitude of class-size effects are those produced by Krueger (1999), who found that the students in classes that were 7 to 8 students smaller on average than regular-sized classes performed about 0.22 standard deviations better on a standardized test. This means that students performed about 3 percent of a standard deviation better for every 1 student less in the class.This leads to think that if there is a reduction of 10 students, the effect will be of 0.30 standard deviations. Since most other studies find results that are about half of these (or somewhat lower than that) this has led (Hanushek & Rivkin, 2010) to argue that the literature shows that the effect of a ten student reduction in class size is between 0.10 and 0.30 standard deviations of the dependent variable. At the postsecondary level, Bedard and Kuhn (2008) argue that student evaluations may be a useful indicator of a teacher' s performance. Relative to this work, we tackle the problem with an identification strategy that better deals with endogeneity in class size.

There is value in using student ratings for teacher evaluation. Cashin (1999) performs a meta-analysis of the research and concludes that "student ratings tend to be statistically reliable, valid and relatively free from bias or need for control; probably more so than other data used for evaluation" . There is, however, no consensus regarding the adequacy of student ratings as a measure of instructor or course effectiveness. Be that as it may, they are indicators of student satisfaction (Theall & Franklin, 2001). Moreover, there are positive and significant correlations between student ratings and student learning; and between student ratings and observer, peer and alumni ratings (Greenwald, 1997; McKeachie, 1997). However, there are several drawbacks to using student evaluations as an outcome measures. There is controversy regarding the correlation

Please cite this article as: C. Sapelli, G. Illanes, Class size and teacher effects in higher education, Economics of Education Review (2016), http://dx.doi.org/10.1016/j.econedurev.2016.01.001

Table 1
Percentage of valid responses.

Year	Average (%)	Minimum (%)	Maximum (%)	Year	Average (%)	Minimum (%)	Maximum (%)
1996	85.7	14.3	100	2003	72.3	22.7	95.0
1997	79.6	11.3	100	2004	73.6	20.0	100
1998	79.3	19.4	100	2005	77.2	30.8	100
1999	75.0	18.2	100	2006	71.8	32.4	98.3
2000	70.3	19.4	100	2007	59.7	23.3	85
2001	61.3	18.4	98.1	2008	71.4	14.2	100
2002	68.7	17.9	100				

Table 2
Correlations between the five evaluation indexes.

	Course Aspects	Evaluation Aspects	Recommendation	Satisfaction	Teacher's Work
Course aspects	1.00	0.65	0.66	0.67	0.80
Evaluation aspects	0.65	1.00	0.71	0.72	0.77
Recommendation	0.66	0.71	1.00	0.90	0.85
Satisfaction	0.67	0.72	0.90	1.00	0.81
Teacher's work	0.80	0.77	0.85	0.81	1.00

of student evaluations with factors such as class size and severity of grading. Researchers and critics of student evaluation have suggested several factors which may bias student ratings of teacher effectiveness, such as class size, grade leniency, instructor personality, gender, course workload, time that the class meets, and type of course (academic discipline, etc.). Braga et al. (2014) find that teacher effectiveness and student evaluations are negatively correlated, although this is less so for classes where high-skill students are over-represented. Overall, we wish to sidestep the discussion of the controversial link between student evaluations and learning, believing instead that student satisfaction has value in its own right for Higher Education administrators.

3. Data

Course evaluation data comes from courses taught at FACEAPUC between the second semester of 1996 and the second semester of 2008. Overall, the dataset consists of 25 semesters, 276 courses, and 539 teachers, for a total of 3421 observations. FACEAPUC consists of two entities, the Economics Institute and the School of Administration, and our course data comes from the Commercial Engineering, Master of Economics, Master of Administrative Sciences, and PhD in Economics programs. Commercial Engineering is a professional degree that is a mixture of economics and administration[2], and most of our data comes from courses that are either core or elective courses for this program. However, some courses are electives for this program and required for the more advanced programs. Furthermore, some courses in our dataset are supervised by FACEAPUC but taught to students from other faculties. Overall, this suggests that there could be differences in the students that attend different courses, but there have been no significant policy changes that would create differences across time in the students that attend the same course. Therefore, course fixed effects should solve any problems that arise from this issue. Furthermore, students at FACEAPUC are relatively homogenous, as they are drawn from the right tail of the distribution of scores in the Chilean university admission test, and classrooms are physically very similar.

Student evaluations are performed twice a year, at the end of each semester, and consist of an online questionnaire containing a series of ordered response questions about different aspects of a course. Table 1 presents yearly averages for the percentage of students in a class who complete the evaluation. Although these percentages fluctuate between years, for every year in our sample more than half of all students have completed the evaluation. However, since evaluations are voluntary, classes with low response percentages may suffer from selection bias. In order to determine whether this affects the estimation of class size effects, we will compare the results from our preferred specification to results obtained under different response rate sample restrictions. We find that restricting the sample to classes with high or low response rates does not affect our results.

The questionnaires' answers are processed and converted into five indexes, that correspond to each student's perception on the following topics: Course Aspects, Evaluation Aspects, Recommendation, Satisfaction, and Teacher's Work[3]. Table 2 presents the correlations between the different indexes, showing that all five are positively correlated but that these correlations are not always high or stable. In fact, they range between 0.65, the Course Aspects

[2] Commercial Engineering students follow a curriculum that is equivalent to that of an Economics student in some countries and to that of a Business student in others.

[3] Each index summarizes the answers to different questions about the course experience. Course Aspects pertains to logistical aspects of the course, Evaluation Aspects to the course's evaluations, Recommendation to whether the student would recommend the teacher, Satisfaction to whether the student is satisfied with the course, and Teacher's Work to the degree of work put in by the teacher during the course.

Please cite this article as: C. Sapelli, G. Illanes, Class size and teacher effects in higher education, Economics of Education Review (2016), http://dx.doi.org/10.1016/j.econedurev.2016.01.001

Table 3
Mean student evaluation, by class size.

Mean student evaluation, by class size		
Class size	Mean	Frequency
$<$ 20 Students	76.5	216
$\leq$ Students $<$ 40	75.7	780
40 $\leq$ Students $<$ 50	73	510
50 $\leq$ Students $<$ 60	74.3	651
60 $\leq$ Students $<$ 70	74.9	766
70 $\leq$ Students $<$ 80	75.4	355
80 $\leq$ Students $<$ 90	71.1	115
Students $\geq$ 90	72.1	28

Index and Evaluation Aspects Index correlation, and 0.9, the correlation between the Recommendation Index and the Satisfaction Index. In the interest of clarity, we will restrict our attention to one index, Satisfaction. We believe that focusing on whether a student is satisfied with the course is more relevant for our research question than students' opinions about evaluations, how hard the teacher worked, or logistical aspects of the course[4]. Furthermore, the correlation between the Recommendation and Satisfaction indeces is high, and results do not vary significantly between them. Thus, in what follows we will focus on student satisfaction, and all of our results will be presented in standard deviation units of this measure.

Table 3 shows class means for the satisfaction index, by class size. Interestingly, class means decrease slightly as class size rises, and the naive interpretation would be that class size does not have a strong negative impact on evaluations. However, we know these means are also affected by the fact that better teachers are more likely to be assigned to larger classes. In fact, at FACEAPUC class size is determined by the students' demand up to a cap, since the administration only sets limits on the maximum number of students that are allowed to take a class. More specifically, class size is determined after a two stage bidding process. In the first stage, the department opens classes and sets a maximum class size. Students have an endowment of points, which are spend bidding for different classes. The students with the highest bids are assigned to each class, until the class is full or demand is satisfied. If classes are full, the department can increase their size to accommodate demand, or open new course offerings. After this process, a second round of bidding is opened for unused slots, and once again classes may be expanded. In both of these stages, classes almost never go beyond 85 students due to classroom constraints. Hence, we build an identification strategy for class size effects based on the 85 student cap induced by classroom constraints. This strategy will be discussed further in the next section.

Finally, data on teaching experience was built by looking at the first time a teacher appears in the sample. If he or she appears for the first time after 1999, we assume that it is their first time teaching, while if they appear before that date, we look at FACEAPUC's records for 1995 and 1996 to check whether they had taught before[5]. Table 4 summarizes the percentage of first time teachers for every year in the sample. Wary of the fact that 1996 and 1997 show abnormally high numbers of first time teachers, we drop the observations for these two years and repeat the analysis, and find no significant differences.

Table 4
Percentage of first time teachers.

Year	Percentage of first time teachers (%)	Year	Percentage of first time teachers (%)
1996	27.3	2003	8.9
1997	20.8	2004	10.4
1998	15.2	2005	7.9
1999	14.0	2006	10.2
2000	13.8	2007	7.9
2001	9.1	2008	9.8
2002	16.1		

4. Methodology

A naive approach to estimating the causal effect of class size on student satisfaction would be to estimate the following OLS regression:

$$s_{cpt} = f(size_{cpt}) + \epsilon_{cpt} \quad (1)$$

where s_{cpt} is the student satisfaction index for course c, taught by professor p, in period t; and $f(size_{cpt})$ is some function of class size for that course, professor and time combination. This regression is problematic, as the effect of any omitted variable that is correlated with class size will be loaded onto the class size coefficient estimates. For example, if better teachers are assigned to bigger (smaller) classes, this regression will underestimate (overestimate) the effect of class size. Other possible sources of omitted variable bias include differential teacher-student match quality at different class sizes, and selection of different ability students into different class sizes.

We use a quirk in FACEAPUC's course scheduling methodology to obtain quasi-experimental variation in class size: administrators would like class size to be capped at 85 students. As a result, whenever enrollment in a class crosses 85 students, the probability that a new class is opened increases, and students are funneled into that new class. This is similar in spirit to the Maimonides' Rule discussed in Angrist and Lavy (1999). This implies that we can use a fuzzy regression-discontinuity design to estimate the causal effect of class size on student satisfaction, by looking at the variation in class size induced by classes going over the 85 student threshold. To do so, we calculate a classes' predicted size if each instance of a course being taught in a semester was firmly capped at 85 students:

$$Predicted\ class\ size = \frac{Total\ enrollment}{\left\lfloor \frac{Total\ enrollment - 1}{85} \right\rfloor + 1} \quad (2)$$

[4] The question used to construct the index is "How satisfied are you with the course?", with options ranging from 1 to 5.

[5] One could object that we are building a "First Time Teaching in FACEAPUC" variable, since we do not know whether they have taught elsewhere. However, we do not believe that this is a significant issue in our sample, as most first time teachers are very young.

Please cite this article as: C. Sapelli, G. Illanes, Class size and teacher effects in higher education, Economics of Education Review (2016), http://dx.doi.org/10.1016/j.econedurev.2016.01.001

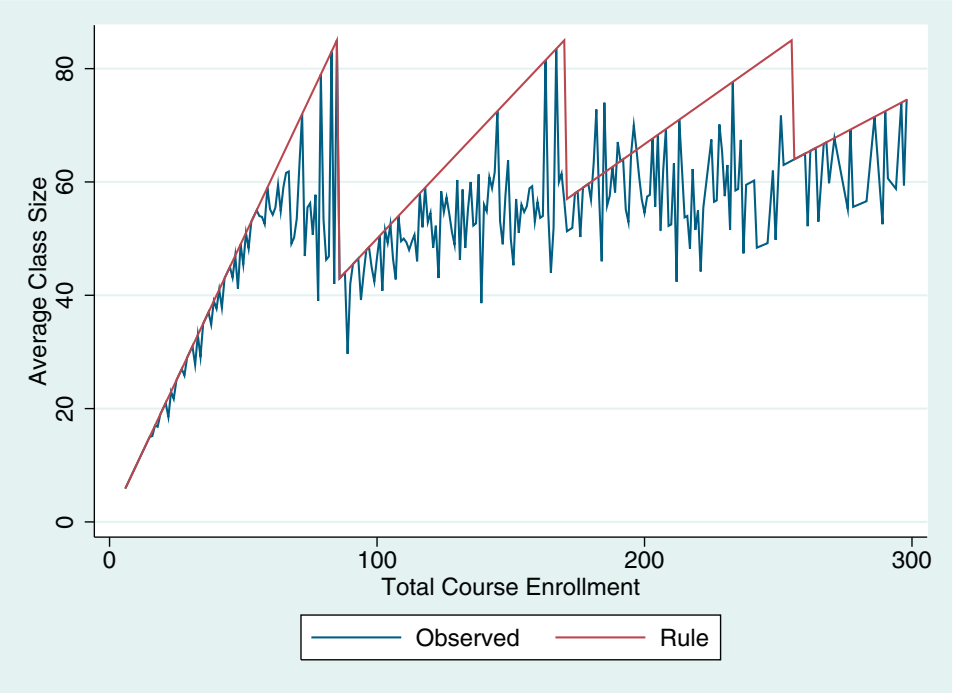


Fig. 1. Maimonides' Rule at PUC.

Table 5
Teacher observables around cutoff.

Comparison of teacher observables around discontinuities							
	Has PhD	Has masters	Has MBA	Full time	Company	REPEC citations	Age
Above a cutoff	0.046*	0.025	−0.063***	−0.013	−0.042	−0.130	0.298
	(0.026)	(0.029)	(0.020)	(0.029)	(0.029)	(1.645)	(0.646)
Observations	431	431	431	440	436	439	445

Notes: This table presents results of the OLS regression of different teacher observables on whether a class is above or below the cutoff, conditional on being 10 students above or below a cutoff. "Full time" refers to whether a teacher is employed full time by FACEAPUC. Company refers to whether the teacher works in a company. Heteroskedasticity-robust standard errors in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.1$

and use this variable as an instrument for class size. That is, for every course being taught each semester, we sum enrollment over all instances of the course being taught (classes), predict enrollment in each class, and use this variable as an instrument. Fig. 1 plots observed class size and actual class size. Note that there is in fact a sharp discontinuity at 85 students, and a smaller but still distinct discontinuity at 170 students.

What does this variation buy us? If students cannot differentially select into classes on different sides of the thresholds, then this instrument will deal with any omitted student-level heterogeneity that is correlated with class size. Furthermore, if the administration does not differentially determine whether to add a new class or not depending on teacher quality or course characteristics, then this instrument also deals with any omitted teacher and course variables that are correlated with class size. The former seems reasonable, as students cannot accurately predict whether a class will be full or not and more classes will be opened when selecting courses, but the latter does not. One would expect the administration to be more lenient in allowing class size to go above 85 students for high quality teachers or for certain courses where a suitable teacher for a new class is harder to find. As a result, we view this instrument as controlling for omitted student characteristics, and will use teacher, course and time fixed effects to deal with unobserved heterogeneity in course and teacher characteristics that is correlated with class size.

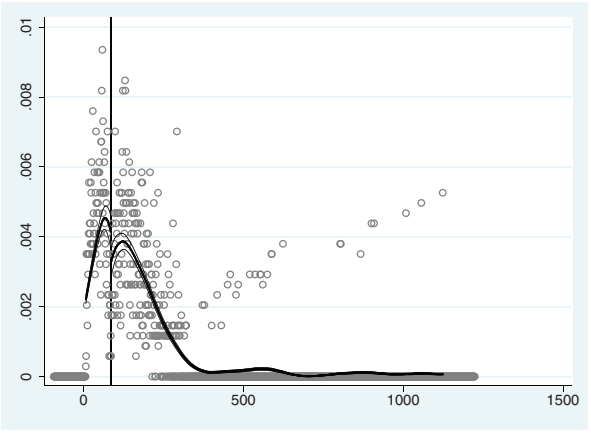
We do not have individual-level information that would allow us to test whether students differentially sort into classes above or below the cutoff. However, we can test whether teacher characteristics significantly vary around the cutoff. If teachers are systematically assigned to classes below or above the cutoff, part of the effect this methodology captures will be due to teacher quality, and not to class size effects. Table 5 presents results of the OLS regression of different teacher observables on a dummy variable for whether a class is above the cutoff, conditional on enrollment being within 10 students of a

Please cite this article as: C. Sapelli, G. Illanes, Class size and teacher effects in higher education, Economics of Education Review (2016), http://dx.doi.org/10.1016/j.econedurev.2016.01.001

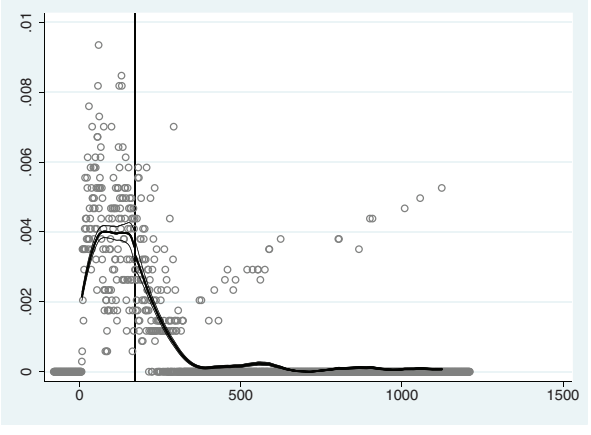
predicted class size discontinuity. Classes above cutoffs are marginally more likely to be taught by instructors with a PhD, and significantly less likely to be taught by instructors with an MBA, although none of these effects is large. There is no difference around the cutoffs in the probability that an instructor has a Masters degree or works in a company, as well as no effect on instructor age and citations. Overall, we think this evidence suggests that teacher observables do not significantly vary around the cutoffs. Nonetheless, our preferred specifications include teacher fixed effects, in order to deal with the possibility that the administration has information on unobserved teacher quality and non-randomly selects teachers into classrooms.

Another test of the validity of our methodology would be to calculate whether there is selective bunching at the discontinuities, following McCrary (2008). Fig. 2 presents results of the McCrary (2008) test of manipulation of the running variable. The test for the discontinuities at total enrollments of 170 and 255 students cannot reject the hypothesis that the distribution of the running variable is smooth at the cutoff. Meanwhile, the test at 85 enrollees rejects the hypothesis that the distribution of enrollees is smooth at the cutoff, but in the opposite direction that selective bunching due to class size effects would predict. That is, the test finds that the mass above 85 enrollees is smaller than the mass below 85 enrollees. However, if there was selective bunching at the cutoff due to class size effects, we'd expect greater mass above the cutoff than below, as classes above the cutoff are more likely to be broken up. Instead, we interpret this finding as being driven by administrative constraints: sometimes demand for a class is greater than 85 students, but an additional teacher or classroom cannot be found. In those cases, the class is simply taught for 85 students. This failure to open a second class of a course leads to greater mass below 85 students than above 85 students. Since we have already shown that teacher observables do not vary around the cutoffs, we interpret the probability of being administratively constrained as being uncorrelated with the unobservable.

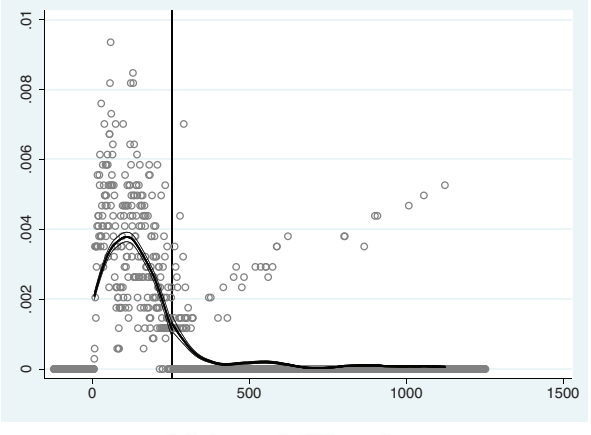
After obtaining class size estimates, we partial out the effect of class size on student satisfaction, and regress residual satisfaction on a first time teacher dummy. The goal is to obtain estimates of the effect of being a first time teacher without dropping individuals who only teach once, as would happen if we simply included the first time teacher dummy as an additional control in the previous estimation strategy. We do not have an instrument for being a first time teacher, however, so any omitted variable that is correlated with said variable and residual satisfaction will bias our results. We do not believe this to be a first order concern, and that this methodology is suitable to gain an understanding of the effect of being a first time teacher on course satisfaction, as will be discussed in the following section.



(a) Around 85 students



(b) Around 170 students



(c) Around 255 students

Fig. 2. Selective bunching at discontinuities.

5. Results

Panel A of Table 6 reports coefficient estimates for the class size effect, obtained under different sets of controls. Column 1 implies that a 1 standard deviation increase in class size leads to a 0.08 standard deviation drop in student satisfaction, without including any controls. Column 2 incorporates time effects, and the effect increases to 0.097 standard deviations. Column 3 adds teacher and course fixed effects, and the effect now increases to 0.187 standard deviations. Finally, the last column presents the OLS

Please cite this article as: C. Sapelli, G. Illanes, Class size and teacher effects in higher education, Economics of Education Review (2016), http://dx.doi.org/10.1016/j.econedurev.2016.01.001

Table 6
Fuzzy RD estimates of class size on student satisfaction.

Fuzzy RD estimates of the effect of class size on student satisfaction				
	(1)	(2)	(3)	(OLS)
Panel A: 2SLS				
Class size	−0.081**	−0.097***	−0.187***	−0.128***
	(0.032)	(0.031)	(0.046)	(0.018)
Time FE		x	x	x
Course FE			x	x
Teacher FE			x	x
Observations	3421	3421	3244	3244
Panel B: first stage				
	(1)	(2)	(3)	
Predicted class size	0.592***	0.596***	0.508***	
	(0.011)	(0.011)	(0.022)	
Time FE		x	x	
Course FE			x	
Teacher FE			x	
Observations	3421	3421	3244	
First stage *F*-test	2849.3	2860.4	426.2	

Notes: This table presents results of the 2SLS regression of the standardized student satisfaction index on standardized class size, using standardized predicted class size as an instrument. Panel A presents the 2SLS results, while Panel B presents the first stage results. The row marked "First Stage *F*-test" presents the *F*-test for instrument significance in the first stage. Heteroskedasticity-robust standard errors in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.1$

results of regressing student satisfaction on class size, including time, course and teacher fixed effects. This estimate implies that a one standard deviation increase in class size leads to a 0.128 standard deviation drop in student satisfaction.

The main point of this table is to show that under our preferred specification, in column 3, class size effects are significant, showing that students value smaller classes. Since the standard deviation of class size is 18.7, this is a per-student effect of 0.01, which is in line with the lower bound of class size effect estimates in Hanushek and Rivkin (2010). Thus, while the effect exists, it is small relative to the literature on K-12 education. Comparing column 3 to the OLS results, we see that omitted student-level heterogeneity is positively correlated with class size, leading the OLS estimates to be an underestimate of the true class size effect. This is consistent with positive matching between class size and students who value class size. A comparison between columns 2 and 3 shows that, as expected, the fuzzy regression discontinuity design by itself is not enough to control for unobserved course and teacher characteristics. This could be the case because the administration is more lenient in allowing class size to cross the enrollment thresholds whenever a teacher is high quality or the course is particularly suited for this. Panel B presents first stage results for the fuzzy regression discontinuity design. Note that all specifications have large first stage *F*-tests, alleviating any concerns about weak instruments.

Table 7 focuses in on each discontinuity, by reporting results of the robust nonparametric confidence interval procedure proposed by Calonico, Cattaneo, and Titiunik (2014). The purpose of this methodology is to improve on local polynomial estimators for regression discontinuity design, by correcting for the bias induced by

Table 7
Class size effects at each discontinuity.

Class size effects at each discontinuity			
	(1)	(2)	(3)
Total enrollment	85	170	255
Class size	−0.151	5.689	−0.626
	(0.827)	(11.951)	(1.895)
Observations	548	940	971

Notes: This table presents results of the robust nonparametric confidence interval procedure for regression-discontinuity designs proposed by Calonico et al. (2014). Column 1 looks at classes around the first enrollment discontinuity (85 students), column 2 at the second enrollment discontinuity (170 students), and column 3 at the third enrollment discontinuity (255 students.) Total enrollment is defined as the total number of students enrolled in a course in the semester, summing across all instances of the class. Heteroskedasticity-robust standard errors in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.1$

bandwidth selection procedures. Column 1 focuses on the discontinuity around 85 enrolled students, column 2 on the discontinuity around 170 enrolled students, and column 3 on the discontinuity around 255 enrolled students. The first discontinuity presents results that are in line with the results from the previous table, albeit with significantly less precision. As discussed in Angrist and Pischke (2009), this is to be expected. Unfortunately, this is not the case for columns 2 and 3, as the instrument is weak around these discontinuities, and does not buy us significant power. Therefore, most of the power of the instrument is coming off the first discontinuity. Overall, these results show that we are gaining power in Table 6 by joining all three discontinuities and by including in the sample courses that are outside the optimal bandwidth. However, we do not think that this is problematic, as the estimates obtained from using only the most powerful discontinuity are in line with the results obtained using the full sample.

When working with student evaluations, response rates and selection into responding are a significant concern, as they could lead to biased estimates. Table 8 deals with this issue by reporting results for our preferred specification by response rate groups. As a reference, column 1 copies the results from column 3 in Table 6, while column 2 looks at classes with response rates that are above the sample average (74.6%). Column 3 looks at classes with response rates above the 75th percentile of response rates (85.3%), and column 4 looks at classes with response rates below the 25th percentile (61.8%). While precision falls, the parameter estimates are stable across subsamples, leading us to conclude that selection into non-response is not an important source of bias in this setting.

Another potential concern with the results obtained in Table 6 is that we are omitting grades as a control variable. If students reward teachers who give better grades with better evaluations, and class size is correlated with grades, then this omission may be problematic, depending on one's interpretation of the relationship between the student satisfaction index and grades. If one interprets student satisfaction as an indicator of learning, and higher grades reflect a better understanding of the material, one would not want to control for grades separately, as doing so would soak up part of the causal effect of class size

Please cite this article as: C. Sapelli, G. Illanes, Class size and teacher effects in higher education, Economics of Education Review (2016), http://dx.doi.org/10.1016/j.econedurev.2016.01.001

Table 8
Fuzzy RD estimates of class size on student satisfaction, by response rates.

Fuzzy RD estimates of the effect of class size on student satisfaction, by student evaluation response rates				
	Response rate cutoffs			
	Full sample (1)	> Mean (2)	> 75th percentile (3)	< 25th percentile (4)
Class size	−0.187***	−0.185**	−0.193	−0.209**
	(0.046)	(0.077)	(0.138)	(0.082)
Time, course, teacher FE	x	x	x	x
First stage *F*-test	426.2	96.0	25.1	305.5
Observations	3244	1881	852	854

Notes: This table presents results of the 2SLS regression of the standardized student satisfaction index on standardized class size, using standardized predicted class size as an instrument. Column 1 reproduces the results for the full sample, column 2 restricts the sample to courses whose response rate is above the mean response rate (74.6%), column 3 restricts the sample to courses with response rates above the 75th percentile of response rates (85.3%), and column 4 restricts the sample to courses with response rates below the 25th percentile (61.8%). The row marked "First Stage *F*-test" presents the *F*-test for instrument significance in the first stage. Heteroskedasticity-robust standard errors in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.1$

Table 9
Fuzzy RD estimates of class size on student satisfaction, controlling for grades.

Fuzzy RD estimates of the effect of class size on student satisfaction, controlling for grades		
	(1)	(2)
Class size	−0.187***	−0.157***
	(0.046)	(0.046)
Grades		0.220***
		(0.042)
Time, course, teacher FE	x	x
First stage *F*-test	426.2	421.1
Observations	3244	3244

Notes: This table presents results of the 2SLS regression of the standardized student satisfaction index on standardized class size, using standardized predicted class size as an instrument, with and without including grades as a control. The row marked "First Stage *F*-test" presents the *F*-test for instrument significance in the first stage. Heteroskedasticity-robust standard errors in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.1$

Table 10
First time teacher effect estimates.

Estimates of the effect of first time teachers on residual student satisfaction		
	(1)	(2)
First time teacher	−0.415***	
	(0.058)	
One time teacher		−0.716***
		(0.116)
Continuing first time teacher		−0.295***
		(0.063)
Observations	3421	3421

Notes: Column 1 presents results of the regression of the residual standardized student satisfaction index on a first time teacher dummy, while column 2 breaks down first time teachers into those that only teach once ("One Time Teacher") and those who later go on to teach again ("Continuing First Time Teacher"). Residual standarized student satisfaction is obtained by partialling out the effect of class size on the student satisfaction index. Heteroskedasticity-robust standard errors in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.1$

on learning. However, if the correlation between grades and student satisfaction is not due to learning, and grades correlate with class size, we would want to control for grades. We do not have the ability to identify which of these stories is correct, but fortunately we do not need to, as Table 9 shows that including grades as a control leads to a drop in the impact of a one standard deviation increase in class size from −0.187 to −0.157, which is statistically and economically insignificant. That is, although grades are positively correlated with student satisfaction, the correlation between grades and class size is small enough that including this variable does not significantly change our results. Therefore, we are able to side-step the discussion of the relationship between grades and satisfaction.

Having obtained estimates of the class size effect, we partial them out from student satisfaction and regress the residual on a first time teacher dummy. This allows us to estimate the correlation between being a first time teacher and student satisfaction, controlling for the causal effect of class size. This is important, as first time teachers could be differentially assigned to classes of different size. As mentioned in the previous section, the advantage of this methodology is that it allows us to include teachers who only teach once in the estimation, an important driver of the first time teacher effect. We argue that this correlation is the relevant measure of first time teacher quality, reflecting the average difference in student satisfaction between a first time teacher and an experienced teacher. Table 10 presents the results of this exercise. Column 1 shows that first time teachers are associated with a 0.415 standard deviation drop in student satisfaction. An interesting thought experiment is the comparison between having an 85 student class and breaking up said class into two 42.5 student classes, with one class keeping the same teacher as before and the other being taught by a first time teacher. The class that keeps the same teacher experiences a 0.425 standard deviation increase in satisfaction on average due to lower class size, while the class being taught by a first time teacher on average has a 0.01 standard deviation increase in satisfaction[6]. Thus, on average

[6] $0.187 \times 42.5/18.7 - 0.415$

Please cite this article as: C. Sapelli, G. Illanes, Class size and teacher effects in higher education, Economics of Education Review (2016), http://dx.doi.org/10.1016/j.econedurev.2016.01.001

breaking up an 85 student class is barely Pareto optimal. Column 2 delves into this issue more deeply, by separating first time teachers into those who teach again ("Continuing First Time Teachers") and those who only teach once ("One Time Teacher"). One time teachers are associated with a 0.72 standard deviation drop in satisfaction, while continuing first time teachers are associated with a 0.3 standard deviation drop in satisfaction. Thus, breaking up an 85 student class into two classes, with one being taught by a one time teacher, hurts the students who get the one time teacher by 0.3 standard deviations[7]. This illustrates the fact that while on average breaking up a class and adding a first time teacher is Pareto optimal, getting a bad draw of the first time teacher leads to significant losses for students.

The discussion regarding the Pareto optimality of breaking up a class is interesting, as it could be used to evaluate FACEAPUC's policies regarding the hiring of new teachers. A reasonable stopping rule for hires would be to stop when the marginal teacher on average leaves students indifferent between being on either side of the 85 student enrollment discontinuity. These results suggest that the average new hire satisfies this rule. An important caveat is that this rule ignores dynamics, and that one may be willing to accept worse marginal teachers today in order to build a stock of experienced teachers for the future. Whether this is relevant or not depends on the average expected lifespan at FACEAPUC of the marginal teacher, as if the marginal teacher is a one semester filler this consideration will not matter. We cannot identify marginal teachers, but intuitively they should not be the individuals who are expected to stay at FACEAPUC the longest, such as new tenure track hires. As a result, we feel that FACEAPUC's hiring policies seem to be, to first order, consistent with the class size effect around the 85 student enrollment discontinuity.

6. Conclusion

We study the tradeoff between smaller class sizes and teacher effects in the production function for higher education. While reducing class size has a positive effect, we argue that a comprehensive class size reduction policy has to be coupled with an expansion in the number of teachers. If this is the case, then the relative quality of marginal teachers is critical for the success of such a policy. In order to explore whether the quality gap between infra-marginal and marginal teachers dominates the class size effect, we estimate class size and first time teacher effects. Our findings show that a negative class effect does exist, and its' impact when class size is reduced can be offset by the negative impact of a first time teacher. Hanushek and Rivkin (2010) survey various studies and argue that the effect of a ten student reduction in class size is between 0.10 and 0.30 standard deviations of the dependent variable. In comparison, we predict that said impact is roughly 0.10 standard deviations, a relatively small class effect. At the same time, Rockoff (2004) finds that a one standard deviation increase in teacher quality raises learning outcomes in 0.24 standard deviations. We find that a first time teacher lowers outcomes in roughly 0.41 standard deviations, and that first time teachers that are not invited to teach again lower them in 0.7 standard deviations.

These results imply that it is Pareto optimal to break up a class, giving half the students to a first time teacher, for class sizes above 85 students. This is the case because the students who keep the same teacher gain a 0.425 standard deviation increase in satisfaction on average due to lower class size, while the students who are assigned to the new teacher have a 0.01 standard deviation increase in satisfaction. Since this is precisely the rule in place at FACEAPUC, one could judge this institution's hiring policies to be on average getting this decision right. However, the effect of low quality first time teacher, defined as individuals who are not invited to teach again, is so large (−0.71 SDs) that it is Pareto optimal to break up a class only when class size is greater than 140 students. This suggests that the ability to detect low quality first time teachers is important, and while on average FACEAPUC seems to be doing this correctly, we are unable to determine whether the marginal hire is of high or low quality. These results highlight that finding methodologies to identify poor quality first time teachers seems like a relevant area of future research.

Regarding the external validity of our results, to begin we find results that relate well to those in the literature. But many results are dependent on criteria used by administrators, characteristics of the student pool, of the teacher pool one has available, and on the characteristics of full time (or experienced) professors. The stopping rule we described is key in determining the margin at which we are measuring results. However, much of the discussion is relevant to any education institution and in particular to any higher education institution.

[7] $0.187 \times 42.5/18.7 - 0.72$

References

Angrist, J. D., & Lavy, V. (1999). Using maimonides' rule to estimate the effect of class size on scholastic achievement. *The Quarterly Journal of Economics, 114*(2), 533–575. http://ideas.repec.org/a/tpr/qjecon/v114y1999i2p533-575.html

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics*. Princeton, NJ: Princeton University Press.

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. http://ideas.repec.org/a/eee/ecoedu/v27y2008i3p253-265.html

Braga, M., Paccagnella, M., & Pellizzari, M. (2014). Evaluating students' evaluations of professors. *Economics of Education Review, 41*(C), 71–88. http://ideas.repec.org/a/eee/ecoedu/v41y2014icp71-88.html

Calonico, S., Cattaneo, M. D., & Titiunik, R. (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica, 82*(6), 2295–2326.

Carrell, S. E., & West, J. E. (2010). Does professor quality matter? evidence from random assignment of students to professors. *Journal of Political Economy, 118*(3), 409–432. http://ideas.repec.org/a/ucp/jpolec/v118y2010i3p409-432.html

Cashin, W. E. (1999). *Changing practices in evaluating teaching* (pp. 25–44)). Bolton, MA: Anker Publishing Company, Inc.

Greenwald, A. G. (1997). Validity concerns and usefulness of student ratings of instruction. *American Psychologist, 52*, 1182–1186.

Hanushek, E. A. (2003). The failure of input-based schooling policies. *Economic Journal, 113*(485), F64–F98. http://ideas.repec.org/a/ecj/econjl/v113y2003i485pf64-f98.html

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review, 100*(2), 267–271. http://ideas.repec.org/a/aea/aecrev/v100y2010i2p267-71.html

Please cite this article as: C. Sapelli, G. Illanes, Class size and teacher effects in higher education, Economics of Education Review (2016), http://dx.doi.org/10.1016/j.econedurev.2016.01.001

Jepsen, C., & Rivkin, S. (2009). Class size reduction and student achievement: The potential tradeoff between teacher quality and class size. *Journal of Human Resources, 44*(1). http://ideas.repec.org/a/uwp/jhriss/v44y2009i1p223-250.html

Krueger, A. B. (1999). Experimental estimates of education production functions. *The Quarterly Journal of Economics, 114*(2), 497–532. http://ideas.repec.org/a/tpr/qjecon/v114y1999i2p497-532.html

Krueger, A. B. (2003). Economic considerations and class size. *Economic Journal, 113*(485), F34–F63. http://ideas.repec.org/a/ecj/econjl/v113y2003i485pf34-f63.html

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics, 142*(2).

McKeachie, W. (1997). Student rating, the validity of use. *American Psychologist, 52*(11), 1218–1225.

Meghir, C., & Rivkin, S. (2011). *Econometric methods for research in education: (Vol. 3* (pp. 1–87)). Elsevier. http://ideas.repec.org/h/eee/educhp/3-01.html

Murray, H. (1994). *Can teaching be improved?*. Ontario, Canada: Brock University.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*(2), 417–458. http://ideas.repec.org/a/ecm/emetrp/v73y2005i2p417-458.html

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review, 94*(2), 247–252. http://ideas.repec.org/a/aea/aecrev/v94y2004i2p247-252.html

Theall, M., & Franklin, J. (2001). Looking for bias in all the wrong places: A search for truth or a witch hunt in student ratings of instruction? *New Directions for Institutional Research*, 45–56. doi:10.1002/ir.3.

Please cite this article as: C. Sapelli, G. Illanes, Class size and teacher effects in higher education, Economics of Education Review (2016), http://dx.doi.org/10.1016/j.econedurev.2016.01.001