NBER WORKING PAPER SERIES

CLASS SIZE AND CLASS HETEROGENEITY

Giacomo De Giorgi
Michele Pellizzari
William Gui Woolston

Working Paper 16405
http://www.nber.org/papers/w16405

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2010

We thank Joe Altonji, Pascaline Dupas, Caroline Hoxby, Seema Jayachandran, Ed Lazear, Aprajit Mahajan, John Pencavel, Kathryn Shaw, Chris Taber and seminar participants at the NBER-Summer Institute 2009. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2010 by Giacomo De Giorgi, Michele Pellizzari, and William Gui Woolston. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Class Size and Class Heterogeneity
Giacomo DeGiorgi, Michele Pellizzari, and William Gui Woolston
NBER Working Paper No. 16405
September 2010
JEL No. A22,I23,J30

ABSTRACT

We study how class size and class composition affect the academic and labor market performance of college students, two crucial policy questions given the secular increase in college enrollment. Our identification strategy relies on the random assignment of students to teaching classes. We find that a one standard deviation increase in class-size results in a 0.1 standard deviation deterioration of the average grade. Further, the effect is heterogeneous as it is stronger for males and lower income students. Also, the effects of class composition in terms of gender and ability appear to be inverse U-shaped. Finally, a reduction of 20 students (one standard deviation) in one's class size has a positive effect on monthly wages of about 80 Euros (115 USD) or 6% over the average.

Giacomo DeGiorgi
Department of Economics
Stanford University
579 Serra Mall
Stanford, CA 94305
and NBER
degiorgi@stanford.edu

Michele Pellizzari
Bocconi University
Department of Economics and IGIER
via Roentgen 1
20136 - Milan - Italy
and IZA
michele.pellizzari@unibocconi.it

William Gui Woolston
Stanford University
william.woolston@stanford.edu

1 Introduction

This paper estimates the effect on the grades and earnings of college students of two controversial educational policies: reducing class size and changing the degree of student heterogeneity within a class. The literature on the education production function finds inconsistent results of the effect of class size on student achievement. For example, Angrist and Lavy (1999) and Krueger (1999) find a substantial positive effect of class size reduction, while Hanushek (1996) and Hoxby (2000) find no impact, a result that is also confirmed in the review of the literature by Hanushek (2006) and by the experimental study of Duflo et al. (2009) in Kenya.

The effect of class heterogeneity on performance is even less clear. Econometric complications have largely prevented well-identified work on this issue (Manning and Pischke, 2006). Only by using a purposely designed experiment, Duflo et al. (2008) show that tracking according to ability has positive effects on all students.

Estimating the causal impact of class size on student achievement is important from a policy perspective because reducing class size for fixed student population requires hiring more teacher-hours, an expensive proposition. On the other hand, manipulating class composition may have substantial effects on student achievement at much lower costs.

While most of the literature has focused on primary and secondary schools, we concentrate on university students. We believe that focusing on post-secondary education is important for at least two reasons. First, because of the difference in students, professors, and pedagogy between secondary and post-secondary education, results from pre-college academic settings may not be informative for understanding educational interventions at universities. Moreover, there is little research on the effect of class size on test scores in the post-secondary setting. Evidence of significant negative effects of class size on test scores has been presented only by Bandiera et al. (2008), Pinto Machado, and Vera-Hernandez (2009), although in different settings and with different identification strategies from ours. In addition, as the fraction of individuals attending college rises around the world, estimates that refer directly to the production of higher education are likely to become more and

more interesting to policy makers.[1] In particular we are also able to explore the link between those policies and labor market outcomes directly.

In this paper we exploit experimental variation in class size and class heterogeneity that arises from random allocation of students to teaching classes at Bocconi University. Such allocation mechanism was not adopted for research purposes but rather with the aim of encouraging wide interactions among students. Nevertheless, as we discuss later on, the allocation is performed according to a computerized random algorithm, as in a purposely designed experiment.

Besides the focus on higher education and the use of experimental variation, our work differs from the bulk of the existing literature in a third dimension: our data includes information on the labor market outcomes of the students in our sample. Thus, we are able to measure the direct wage effect of class size and heterogeneity, both conditional and unconditional on academic performance. To our knowledge this is the first study to present such evidence, although Moffitt (1996) points out that a separate strand of the school quality literature has indeed looked at earnings. For example, Johnson and Stafford (1973) and Card and Krueger (1992) find substantial positive effects on earnings of increasing expenditure per pupil. Dearden et al. (2002) find that the pupil-teacher ratio has no impact on educational qualifications or on men's wages, but they do find an effect on women's wages at the age of 33, particularly those of low ability. Other papers in this area as Betts (1995), Heckman et al. (1996) find no significant effects.

The policy relevance of the questions we ask is widely recognized. Since the *Coleman Report* (1966), the discussion of improving students' performance has focused on reductions in class sizes (Angrist and Lavy, 1999; Hoxby, 2000) and, to a somewhat smaller extent, on changing the composition of students in a classroom.[2] While the first policy is costly, as it entails the hiring of extra staff-hours, changing the composition of classes according to some underlying observable characteristics of the students is an intervention that could be implemented at zero cost and still guarantee

[1] According to the US census, in 1940 4.6% of adults over 25 had a BA. By 2000, 24.4% held a BA. See www.census.gov/population/socdemo/education/phct41/US.pdf for the full figures. On average in the OECD Countries 56% of school-leavers enrolled in tertiary education in 2006 versus 35% in 1995. The same secular trends appear in non-OECD countries (OECD, 2008). Further, the number of students enrolled in tertiary education has increased on average in the OECD countries by almost 20% between 1998 and 2006, with the US having experienced a higher than average increase from 13 to 17 millions.

[2] The NBER working paper version of Hoxby (2000), (Hoxby, 1998), did actually analyze the effect of class composition and performance. See also Betts and Shkolnik (2000a, 2000b) and Duflo et al. (2008) and the literature cited therein.

possibly large positive effects.

Our main results are that class size is an important determinant of student academic and labor market performances. In our main specification, an increase of class size by one standard deviation, or 20 students from a mean of 131, is associated with a reduction of the mean grade by about 1/3 of a grade point or about 0.14 of a standard deviation. Moreover, we find that this effect does not disappear when the size of the class becomes large, as we cannot reject the linear specification of the class size effect. We find that the effect is largest for males and for students from lower income families. On the other hand, our results suggest no heterogeneity of the effect of class size across students of different academic abilities.

When we explore the role of class heterogeneity on academic performance, we find an inverse U-shaped relation between the share of women in the classroom and academic performance. We find a similar, although less robust, relation in terms of heterogeneity in ability. The effects of the gender and the ability composition of the class are non-linear and open up the possibility of increasing academic performance by reshuffling students into an optimal class allocation without the need to invest in additional resources. We explore this issue in detail in Section 6.

Finally, although the effects of class size on labor market outcomes are less precisely estimated, we find that having experienced larger classes on average is associated with lower wages. Namely, our point estimates suggest that increasing the average class size by 20 students reduces entry monthly wages by 80 euros (approximately 115 USD net of taxes) or 6 percent. This is a very important result, given the substantial impact of initial conditions in the labor market (Oyer, 2006). Our baseline estimates imply that reducing class size is likely to be a very cost effective intervention. Our rich data also allows us to explore the mechanism through which class size may influence labor market outcomes. Conditioning on academic performance reduces the magnitude of the link between class size and earnings by a mere 10 percent, suggesting that class size affects labor market outcomes in ways that are not fully captured by grades.

There are many different mechanisms that could link class size to learning, achievement, and labor market performance. For example, smaller classes allow closer student-teacher interactions,

and are subject to lower disruption levels (Lazear, 2001). In particular, if the effect is generated by disruption, one would expect it not to fade away too rapidly as class size increases. It is also plausible that teachers are able to target the educational content to the interests and abilities of all students in a smaller class. However, when faced with a smaller class, teachers may provide less effort, partly offsetting the benefits of a smaller class size (Duflo et al., 2009). In addition, if students learn from their peers, smaller classes may result in lower student achievement. Similar contrasting arguments might be made regarding the students' composition of the classroom: while it is plausible that a diverse student body has positive effects because of possible complementarities in abilities and types, a very heterogeneous class also makes teaching as well as peer interactions harder (Dobblesteen et al. (2002); Figlio and Page (2002); Duflo et al. (2008)).

Our empirical results may shed new light on this issue and suggest which mechanisms are more likely to be at work. For example, the linearity of the effect of class size on academic achievement seems more consistent with a disruption mechanism than with teachers not being able to adjust their teaching methods to the heterogeneity and size of the class. One important difference between college and school (either primary or high school) classes is their relative sizes. While in primary and secondary schools, class size rarely goes above 50 in developed countries (although it might be larger in the developing world (Duflo et al., 2009)), our classes contain on average around 130 students with a standard deviation of 20. Significant effects for such large classes are more likely to be generated by disruption than by any other mechanisms. In fact, the ability of teachers to adjust their teaching methods to student heterogeneity probably declines quickly with the size of the class, and it seems implausible to expect large differences in this dimension across classes above 70-80 students. We interpret our results as consistent with Lazear (2001).

The paper is structured as follows: Section 2 describes the data and the institutional details of Bocconi University and it also provides evidence on the random allocation procedures. Section 3 discusses the empirical strategy; Section 4 presents the results on academic performance and Section 5 the analysis of labor market outcomes. In Section 6 we present a simple model of optimal class formation. Finally, Section 7 concludes.

2 Data and institutional details

We use data from the administrative archives of Bocconi University, an institution of higher education located in Milan, Italy, that offers degree programs in Economics and Management. There are three features of the data and the institutional setting that are crucial for our analysis.

First and most importantly for our identification strategy, the roughly 1,500 students in each of the two cohorts that we consider were repeatedly randomly assigned to compulsory classes during their first, second, and part of their third academic years. Because classrooms are of different physical sizes, the number of students in each class varies within both cohort and program. Moreover, the random assignment of students generates variation in the amount of heterogeneity within a student's group of classmates. Given the importance of the random variation in class size for our identification strategy, we return to this issue in Section 2.2, where we provide evidence that teachers were also (effectively) allocated randomly.

Second, the administrative data contains a wide array of student characteristics and outcomes that are precisely measured. For each student, we have a wealth of information on her academic curriculum, demographic and socio-economic characteristics. In addition, we have several pre-enrollment variables such as high school leaving grade, type of high school, family income and a good indicator of ability - a cognitive test score that all students take as part of the admission procedure. These variables are important because they allow us to test for random allocation of students into classes and to decompose the effect of the interventions by the predetermined characteristics of the students. From the academic register, we have information on the grades obtained by each student in each exam, which we use as our main outcome variable.

Finally, in addition to these administrative data, we also have access to a series of graduates' surveys that cover all students after 1 to 1.5 years since graduation. These data allow us to understand how educational policies influence labor market outcomes. These surveys collect detailed information on the labor market trajectories of the former students. In Section 2.3 we describe the graduates' surveys in more detail.

In our analysis, we focus on two cohorts of students who matriculated in the academic years

1999-2000 and 2000-2001.[3] At that time, Bocconi offered 7 degree programs. However, only three degree programs were large enough to require the splitting of lectures into more than one class: *Economics*, *Management*, and *Economics and Finance*.[4] The official duration of all programs was 4 years, and during the first two years and most of the third, all students were required to take a fixed sequence of compulsory courses specific to their program. Students could then choose elective courses within program-specific guidelines.

We exclude elective courses from our analysis for three reasons. First, elective courses typically had only one class each year. Differences in class size would therefore originate from differential enrollment across years, a source of variation that is plausibly correlated with student ability and professor quality. If, for example, enrollment in a particular elective is high only when the professor is highly effective, differences in class size would be confounded with teacher quality. Second, because students choose to take elective classes, the interpretation of estimates from these courses would be complicated by issues of differential selection into each class. Finally, while compulsory courses were, in general, graded centrally by a group of graders rather than the instructor of a specific class, the grading of elective courses was more decentralized and was conducted by the instructor herself, sometimes with the aid of a grader. Centralized grading is important because when we compare grades across classes, we can be sure that differences in performance do not originate from differential grading practices on the part of an individual instructor.

The academic curricula of the three degree programs considered are described in Table A.1 in the Appendix. The table reports the list of the compulsory courses for each of the three programs, split by academic year and broad subject areas. The table also reports the number of teaching hours for each course. There are usually 7-8 courses in each academic year, and each of them involves on average approximately 60 hours of teaching/lecturing, although some courses are as long as 80 hours or as short as 32 hours.[5]

[3] We have access to data for many cohorts of students (starting with the enrolment year 1989) but, due to a series of changes in the academic structure and to the unavailability of some crucial information, the cohorts considered here are the only ones that could be used in this particular analysis.

[4] The other programs were *Economics and Management of the Public Administration*, *Economics and Law*, *Law*, *Economics and Management in Arts,* and *Culture and Communication.* For students in these four programs, there was only one class per cohort per program; variation in class size for these students originates only from differences in program or cohort size. Therefore, we exclude them from our analysis.

[5] The terms *class* and *lecture* often have different meanings in different countries and sometimes also in different schools within the same country. In most British universities, for example, lecture indicates a teaching session where

To summarize, the institutional setting and the data available for our exercise are ideally suited to analyze the role of class size and composition on academic and labor market performance. First, variation in both the size and the composition of the classes is randomly generated, as in a purposely designed experiment. Second, rather than relying on a standardized test score that may only partly proxy for the skills that school administrators value, we have an individual's performance in each exam. Third, our data contains information on wages. Fourth, because we have administrative data, we are able to observe the entire student population, not just a sample, and can therefore precisely measure the amount of heterogeneity within a class. Fifth, our data contains a wealth of individual level variables such as gender, family income, and the results of a cognitive admission test that are all very precisely measured and used in the analysis to provide evidence on the random allocation students and, more importantly, to analyze the role of class heterogeneity.

Table 1 reports some descriptive statistics on selected variables for the students in our sample. 42 percent of the students are females and 22 percent have family incomes in the highest paying fee bracket, above 90 thousands euros of gross yearly income (approximately 140,000 USD).[6] On average the GPA at this University is about 26/30, which would be about a $B+$ in the US grading system.[7] The figures in Table 1 also show some interesting differences across the two cohorts that we consider. The fraction of female students rises significantly from 39 percent to 45 percent. At the same time, the average entry test score (which is normalized within the 0-100 range) declines from around 73 down to 56, the size of the cohort increases by almost 10 percent, and the average grade increases by 2/3 of a grade point.

[TABLE 1]

an instructor - typically a full faculty member - presents the main material of the course. Classes are instead practical sessions where a teacher assistant solves problem sets and applied exercises with the students. At Bocconi there was no such distinction, meaning that the same randomly allocated groups were kept for both regular lectures and applied classes. Hence, in the remainder of the paper we use the two terms interchangeably.

[6] Family income is recorded by the university for determining student fees. There are 6 income brackets but students whose parental income falls into the highest income bracket are not required to submit any financial statement and their income is top coded.

[7] Grades at Bocconi, like in all other Italian universities, are given on a scale 0 to 30 with pass equal to 18.

2.1 Class allocation and measurement of class size

At the beginning of each academic year, students were randomly assigned a class identifier: a single digit number which identified the classes in which a student would sit. For the remainder of the academic year, students were instructed to take lectures for all courses in the classroom(s) associated with their identifier. At the beginning of the next academic year, the allocation was repeated. This procedure ensures that a student's peers and class sizes are randomly assigned and vary across each academic year (De Giorgi et al. (2009) and De Giorgi and Pellizzari (2009)). Elective courses were usually much smaller in size and could easily be taught in a single class. To avoid issues related to the endogenous choice of such courses, we exclude them from our analysis.

Although Bocconi's allocation mechanism is crucial for our analysis, the administration adopted the randomization technique for reasons unrelated to our research. Courses were split into several classes for the explicit purpose of keeping class sizes relatively small and to avoid clustering of students in some classes. The yearly repetition of the random allocation was justified by the desire to encourage interactions among all students. Moreover, for organizational reasons, students allocated to a specific class were also taking most of their courses in exactly the same classroom. This is an important feature of Bocconi's organization because it implies that variation in class size comes mostly from variation in the physical size of the classrooms.

Bocconi is scattered around several buildings that have been built or refurbished at different times so that not all classrooms have the same physical capacity. However, despite the differences in physical size, classrooms are very homogeneous in terms of both equipment and furniture, i.e. all classrooms have PCs and overhead projectors and are furnished with essentially the same chairs, benches and desks. Figure A1 in the Appendix shows pictures of a representative small, medium and large classroom to confirm that, other than these differences in size, all other physical features of the rooms are very comparable.[8]

Both the 1999/2000 and the 2000/2001 cohorts of Management students (around 1,100) are divided into 8 classes that range in size from 113 to 147, while both the 300 students in Economics

[8]The pictures were taken at the time of writing but similar furniture was available also during the time covered by our data. The providers of boards, desks and benches, projectors and computers have not changed since then.

and Finance and the roughly 150 students in the Economics degree program are split in two groups each, with sizes ranging 138 to 158 and from 54 to 95, respectively.

Our main measure of class size comes from the student academic records, where the class identifier is reported next to each student's exam result. Thus, we can count the number of students in any given cohort and year who have the same class identifier. We call this variable the *student count* and it corresponds to the number of students who effectively attended the lectures in the same classroom.[9]

However, we know that this measure of class size differs somewhat from the number of students who were originally given the same class identifier. From the teaching planning office we obtained the exact number of students who were given the same class identifier at the beginning of each academic year. This is the number of students who were allocated to the same class by the university administration at the beginning of each academic year. We call this variable the number of *enrolled students*. In Table 2 we report the basic descriptive statistics of these measures of class size. Overall, the student count varies between 54 and 158, with a mean size of about 130 and a standard deviation of roughly 20 students. The number of enrolled students is generally slightly larger: it ranges from a minimum of 64 to a maximum of 172, with an overall mean of 135 and a standard deviation of 28.

The within academic degree program variation in both measures of class size is more limited. The mean (standard deviation) value of student count is 75 (16) in Economics. This compares to an average class of 133 (6.5) in Management, and 149 (7) in Economics and Finance. When we measure class size with the number of officially enrolled students we find slightly larger averages and, consistently larger standard deviations as shown in Table 2. The degree of variation in the size of the classes in our data is sometimes limited, especially if one looks within programs and academic years, as we do in the regression analysis of section 4.

[TABLE 2 and FIGURE 1]

[9] Small variation may come from students taking the exam without attending the lectures or by informally switching across classes. Both these instances, however, are very limited. Attendance is always strongly encouraged and (nominally) tightly enforced at Bocconi, especially for compulsory courses. Moreover, attendance levels are monitored both during the academic year, by random visits of administrative attendants, and at the end of the course, with the teaching evaluation questionnaires, that are regularly administered to the students. The data show very high and stable attendance levels. Also, class switching is formally forbidden. Informally switching classes is theoretically possible however, since students are given personalized calendars based on their class allocation, those who want to do so would also have to reorganize their entire schedule.

In Figure 1 we provide a more detailed comparison of our two measures of class size. The dark and gray bars show the distributions of the student count and enrolled students variables, whereas the dashed lines indicate the respective averages. We also plot the percentage difference between enrolled students and students counts (the little x's) in relation to the number of students originally assigned to that class identifier (on the horizontal axis). Such differences are close to zero (on average about 6 percent),[10] and they appear to be unrelated to the original official size of the class. We also check this relationship by running a simple regression of the percentage difference between our two measures on the number of officially enrolled students. The estimated coefficient is 0.0006 (with a standard error of 0.0004). In a few cases, however, the differences are larger than 15 percent (namely in 15 classes out of 72).

Differences between the student count and the number of enrolled students come from students requesting changes to their original class allocation later on in the year, either for the entire year or for some specific courses. Such requests were (and still are) usually rare and needed to be well motivated. One common reason for such changes are health problems that might prevent a student from accessing some parts of the building (e.g. because of a broken leg) where the class is located.

Overall, in our data less than 6 percent of the students ever switch a class, either for a single course or for an entire academic year. Switches are approximately five times more likely to occur in the first academic year than later years, also girls are 30 percent less likely to switch. Moreover, high ability students are more likely to switch. Although these changes are rare and could not be explicitly requested, by the student, for academic reasons, we cannot rule out a priori that some of them were implicitly driven by factors, like teacher quality or class size, that are endogenous to our process of interest (academic achievement or labor market performance). It should however be noticed that in the data the probability of switching is not significantly affected by the size of the class, in a probit for the probability of switching the coefficient attached to the official class size (our enrolled variable) is very small and insignificant (.0002 with standard error .0002). However, as mentioned above, students with different characteristics may be more or less prone to advance

[10] Note that the enrolled students variable tends to be larger than the student count variable because a small fraction (5.7%) of students were enrolled in the class but do not have valid entries in the student database. These individuals include students from different cohorts who are retaking a class or who delayed taking a course by 1 or more years and students from different degree programs who switched programs.

such requests, raising the possibility that student count is endogenous.

For these reasons, in the empirical application we present results using both OLS and an IV procedure, where we instrument effective class size measured by the student count with the number of officially enrolled students, which, being the outcome of the random allocation algorithm, is purely exogenous. Moreover, the reduced form estimates of our empirical model also have an interesting interpretation. These estimates are the effect of changing the policy variable that the university administration can more easily manipulate: the number of officially enrolled students. In Section 3 we further discuss our empirical strategy.

Table 3 summarizes the extent of heterogeneity in the classroom and across the 72 different classes.[11] There is a non-negligible variation in one's peer group composition although, as we will show, the amount of heterogeneity is consistent with random assignment of students into classes. For example, the share of females is on average equal to 0.4 with a between-class standard deviation of 0.08. At the extremes, class 11 in the third year of the Finance degree program, for the 2001 cohort, has a share of 0.23; while class 4 of the first year Business, 2000 cohort has a share of 0.6. Similarly, the share of high income students is on average of 0.23, with a range of variation 0.12-0.35. We also detect considerable variation within a degree program for each cohort.

[TABLE 3]

In the next subsection we provide evidence of the effectiveness of the random allocation mechanism of students as well as some evidence of the essentially random allocation of teachers to classes.

2.2 Evidence of random allocation

In this section, we provide evidence that both students and teachers are randomly assigned to classes, as in De Giorgi and Pellizzari (2009) and De Giorgi et al. (2009), and as in Guryan et al. (2009). To establish that, conditional on a cohort, degree program, and academic year, students are not differentially selecting certain classes, we marshal three pieces of evidence.

[11] Students in the degree program in Management (around 1,100) are divided into 8 classes, while the students in both Economics and Economics and Finance are split in two groups each. Additionally, classes change in each of the 3 academic years and 2 cohorts, hence the total number of classes that we observe is $(8\times3\times2)+(2\times3\times2)+(2\times3\times2) = 72$.

First, we demonstrate that the distribution of students' entry test scores is consistent with random assignment. The upper panel of Figure 2 compares these distributions of entry test scores of students in the 8 classes of the Management program in each academic year. The middle and lower panels of Figure 2 plot the same distributions for the 2 classes of Economics and Economics and Finance, respectively.[12]

[FIGURE 2]

As it is evident from these graphs, the distributions of test scores are very similar. Table A2 in the Appendix confirms that this visual evidence is consistent with random assignment. In this table, we report the p-values of a complete battery of Kolmogorov-Smirnov tests for the equality of the distribution of ability in all possible pairs of classes within the same degree program, cohort and academic year. Only in 7 out of the 180 admissible pairs of classes (i.e. 4 percent of the cases) are the distributions statistically distinguishable at the 95 percent level.

Our second piece of evidence of the random assignment of students is presented in Table 4 (Panel A), where we check for random assignment on other observable characteristics. Here, we report tests for the equality of the mean percentage of female, the mean percentage of students from top income families and the mean entry test score across classes within each cohort-degree program-academic year cell.[13] In none of the cases is it possible to detect differences that are significant at conventional statistical significance levels.

Finally, we demonstrate that class size is uncorrelated with the observable characteristics of students in the class. Ruling out this relationship is important because if higher ability students were assigned to smaller classes, results showing a systematic relationship between class size and subsequent performance could reflect underlying differences between the students. In Panel B of Table 4, we run a series of regressions where the unit of observation is a single class (i.e. with 72 observations in total). These regressions help determine whether either of our two measures of class size (the student count and the number of officially enrolled students) is correlated with a

[12] For expositional brevity, all the distributions refer to only one cohort (2000), although the results are similar if we use the other cohort.

[13] The reported F-tests are derived from regressions of the mean characteristics of the class on dummies for the class identifiers, controlling for cohort and academic year fixed effects. The regressions are run using class-level observation, i.e. 48 observation for Management and 12 each for Economics and Economics and Finance.

classes' share of females, share of students from high income families, or average test scores. All results condition on the full three-way interaction of cohort, degree program, and academic year fixed effects. Results show that class size is never significantly correlated with any of the observable characteristics of the student body that we consider. In addition, the reported coefficients are very small in magnitude.

[TABLE 4]

We now turn to studying the assignment of teachers to classes. One potential concern is that teachers select the size of the class they want to teach. If, for example the best teachers are allocated to teach smaller classes, our estimates would reflect both the direct effect of class size and the indirect effect of teacher quality.

We have several reasons to believe that this concern does not apply to our data. First, in personal conversations with university administrators, they have indicated that the assignment of teachers was completely unrelated to the process of allocating students to classes. In fact, the two processes were carried out by distinct bodies: secretaries in each department would assign teachers to class identifiers and officers in a centralized teaching planning office allocated students to class identifiers.

Second, the available empirical evidence is consistent with the hypothesis that a teacher's identity is uncorrelated with the size of her class. Although for privacy reasons we lack the information to identify individual teachers, we demonstrate that teachers who are assigned to teach small classes during one year are not more likely to teach small classes in subsequent years. In particular, we were able to reconstruct the identifiers of the teachers of 4 courses in the Management program. Figure 3 shows the size of the classes allocated to these teachers over the academic years 1999-2000 and 2000-2001. On the horizontal axis we report the (anonymized) teacher identifier and the vertical bar indicating the size of each of the classes taught by that teacher in those academic years. The graphical evidence strongly is consistent with random assignment; for example, instructors of relatively small classes in the 1999-2000 school year appear to have average sized classes during the 2000-2001 school year. In Panel B of Figure 3, we show the same data for the following 4 academic years, 2001-2002 to 2004-2005, thus increasing the number of observations. While the structure of the degree programs

changed for cohorts entering after 2000, we believe that the assignment of teachers to classes was similar in the 2001-2002 to 2004-2005 cohorts. Evidence from these later years is therefore helpful for understanding the assignment of teachers. In those later years, the within-teacher standard deviation in the size of the assigned class is larger than the between-teachers variation and, indeed, quite close to the overall variation. To reiterate, a teacher could be assigned 121 students (about the average) in 2001-02 and then 160 students the following year (the second largest class).[14]

Finally, we supplement this graphical evidence with a simple statistical test for random assignment of teachers. A regression, omitted for brevity, of enrollment on teacher fixed effects shows that certain teachers are not systematically assigned to small or large classes. Out of the 53 teacher fixed effects that we can identify with our data (we use the same data of Panel A in Figure 3 for the regression) only 8 of them were significant at the 90 percent level and 4 of them were significant at the 95 percent level. The F-test for the joint significance of the full set of teacher fixed effects never rejects the hypothesis that they are jointly equal to zero. These results are very robust to the inclusion of additional dummies for academic year and degree program.

[FIGURE 3]

2.3 Survey of graduates

In addition to administrative records, Bocconi regularly surveys its graduates through a questionnaire administered to every student about one and a half years after graduation (De Giorgi et al. 2009). These surveys focus on the labor market experience of the graduates and contain information on the employment profiles, wages, and job satisfaction.

While we view the ability to link detailed information about students while in school with labor market outcomes as an important contribution of our paper, we recognize there are two potential problems with these surveys. First, the response rates are not particularly high; overall, we are able to match slightly more than 50 percent of the students in our cohorts (not unusual for survey data). However, there is no relation between non-response and class size, i.e. we did not find any significant

[14] In the main analysis we do not pool data for the academic years 1999-2001 and 2001-2004 because, starting with 2001-2002, the entire structure of the degree programs was changed.

correlation between non-responses and our class size measures (student count or enrolled).

These response rates are mostly due to the compulsory military service for men, which males typically completed after graduation. On average only about 34 percent of them answer the survey, as opposed to almost 73 percent of females. While we are concerned that selection into the survey may bias our results, we can partially alleviate these concerns by comparing results between our two cohorts. Military service was 10 months long and was abolished in 2001 for all citizens born after 1985. Although the students in our cohorts were born before 1985, in the years prior to the abolition of the service, the set of reasons that allowed exemption were gradually expanded.[15] Hence, the number of people required to serve declines substantially between our two cohorts. While the response rates for females was similar across the two cohorts, the response rate for males increased from 24 percent to 47 percent. If differential response rates were driving our results, we would expect that the results would look different across the two survey years. Instead, we find that the results are generally consistent.

A second issue relates to the measure of wages, which are recorded in 11 intervals. The large majority of respondents (over 90 percent) do report wage information, which is asked to anyone who has had a job between the day of her graduation and the day of the interview (96 percent of the respondents). The intervals range from below 750 to over 5,000 euros per month (net of taxes) and are spaced by either 250 or 500 euros. The descriptive statistics (means and standard deviations) reported in Table 1 refer to an imputed measure of wages computed at the mid-point of the interval indicated by the respondent.[16] All monetary values are in euros at current prices. The mean entry wage is around 1,300 euros net per month, approximately 1,700-1,800 USD. The mean wage increases over time at a rate (around 9 percent) higher than inflation.

3 Empirical Strategy

With a few important exceptions (Krueger (1999), Krueger and Whitmore (2001), Duflo et al. (2008 and 2009)), the existing literature on class size and class heterogeneity has mostly exploited

[15] For example, around the year 2000 a set of new rules allowed permanent exemption from the service to students who enrolled in a PhD programme (one of the author benefitted from it).

[16] For the lowest and the highest intervals we take the upper and the lower limit, respectively.

natural experiments as source of identification. In this work, we exploit the experimental variation arising from the random allocation of students to classes followed at Bocconi University. This random allocation produces exogenous variation in the size and composition of classes and allows us to cleanly identify the effect of class size and heterogeneity on academic performance and labor market outcomes.

Variation in the size of the class is generated by differences in the physical capacities of the classrooms across the university buildings and can be considered exogenous to other inputs in the education production function. In particular, all classrooms have exactly the same equipment and the same furniture except that larger classrooms have larger blackboards and screens.

In the next section, we explore the effect of class size and class heterogeneity on both academic performance and labor market outcomes. Here we briefly discuss our empirical strategies for the identification of these two effects.

Let us start with academic performance. To avoid complications due to the endogenous choice of elective courses, we concentrate exclusively on compulsory courses that comprise the vast majority of a student's courses during the first three academic years. For the two cohorts used in this analysis, students are randomly allocated to different classes three times, one at the beginning of each academic year. Hence, in our empirical specification we use the average grade in the courses of each academic year as a measure of student performance, and we regress it on the class size in each year. We have three observations for each student (one per year), which allow us to control for individual effects as well as for year and program effects. Notice that because the average grade per academic year is computed over a slightly different number of courses across degree programs and academic years, we weight observations accordingly.[17]

We derive our empirical specification from the following model:

$$y_{ijtcd} = \alpha \; size_{jtcd} + \eta_i + \gamma_{tcd} + u_{ijtcd} \tag{1}$$

[17] Each student-year observation is weighted by the number of exams taken by the student in that specific academic year. Table A1 in the Appendix shows that there is some small variation in such number across degree programs and years.

where y_{ijtcd} is the average grade of student i in class j, year t, cohort c and degree program d, $size_{jtcd}$ is the size of the class j in the same tcd cell, η_i is an individual student fixed effect, γ_{tcd} is a fixed effect that varies by year-cohort-program cells and u_{ijtcd} is a residual random term.

The parameters of equation 1 cannot be identified through simple OLS if students defy the random assignment and change classes in a way that is correlated with teacher quality or class size. To describe the nature of the potential endogeneity of $size_{jtct}$, assume that the random term u_{ijtcd} is the sum of an unobservable class component ζ_{jtcd} that is common to all students who are allocated to class j in the tcd cell and a purely random idiosyncratic term v_{ijtcd}:

$$u_{ijtcd} = \zeta_{jtcd} + v_{ijtcd} \tag{2}$$

The most obvious interpretation of ζ_{jtcd} is teacher quality, but it could represent any class specific unobservable shock.

Then, the student count $size_{jtcd}$ results from the aggregation of the individual re-allocation decisions of all the students in the same cohort and degree program. Students who were originally allocated to class j may decide to switch class, while others who were originally allocated elsewhere may request to be moved to class j:

$$size_{jtcd} = enroll_{jtcd} + \sum_{i \in j} in_{ij} - \sum_{i \in j} out_{ij} \tag{3}$$

where $enroll_{jtcd}$ is the number of students originally allocated by the administration to class j in the tcd cell, in_{ij} is an indicator function that is equal to 1 if student i (who was originally allocated to a class different from j) moves to class j and out_{ij} is an indicator function that takes value 1 if student i (originally allocated to class j) manages to be moved elsewhere.

The key endogeneity concern arises because the functions in_{ij} and out_{ij} might be influenced by ζ_{jtcd}, i.e. teacher quality in class j (or any other class-specific shock). More formally, we can define

the two functions as follows:

$$in_{ij} = f\left(X_{ijtcd}, \zeta_{ijtcd}\right) \tag{4}$$

$$out_{ij} = g\left(X_{ijtcd}, \zeta_{ijtcd}\right) \tag{5}$$

where X_{ijtcd} is a set of observable characteristics of the ij pair in the tcd cell and ζ_{ijtcd} can be interpreted either as a single unobservable shock or as a vector of unobservable characteristics of the ij in the same tcd cell.

In this setting, applying simple OLS to equation 1 does not produce consistent estimates of the parameters, particularly of α. The OLS orthogonality assumption fails because $\mathbb{E}\left(size_{jtcd} \cdot u_{ijtcd}\right) \neq 0$. In fact, as equations 3, 4 and 5 demonstrate, the unobservable class shock ζ_{jtcd} is both a determinant of $size_{jtcd}$ and a component of the error term u_{ijtcd} of equation 1.

This discussion also clarifies that $enroll_{jtcd}$ is a perfect instrument for $size_{jtcd}$ in equation 1. In fact, while the observed class size may be correlated with the error term, $\mathbb{E}\left(size_{jtcd} \cdot u_{ijtcd}\right) \neq 0$, $enroll_{jtcd}$ is merely the outcome of the random allocation algorithm. Hence it is exogenous by construction and $\mathbb{E}\left(enroll_{jtcd} \cdot u_{ijtcd}\right) = 0$. At the same time, equation 3 clarifies that $enroll_{jtcd}$ and $size_{jtcd}$ are correlated. Theoretically, if the process of reallocation of students across classes was substantial, $enroll_{jtcd}$ could be a weak instrument. Given what we know from discussions with the university administrators and from our analysis of the raw data in Section 2.1, we do not expect this to be a serious concern. In fact, the results of the first stage regressions for all the specifications that we present in Section 4 confirm this expectation (the F-tests of the excluded instruments range from 51 to 7,000.) Our solution to this identification problem resembles closely the approach of Krueger (1999).

Like all IV empirical strategies in which there are no "defiers", our estimates recover the average treatment effect that is local to the population who "complies" with the instrument (Imbens and Angrist, 1994). In this application, we think that the local average treatment effect (LATE) is likely to be interesting for at least two reasons. First, because switching is rare, a large fraction of the students are likely to comply with their random assignment, and the LATE may be similar to the

average treatment effect. Second, the LATE is policy relevant because it recovers the effect of class size for students who comply with the university's assignment.

Although we use *enrolled students* primarily as an instrument for *student count*, the reduced form estimates are interesting in their own right, as they may be interpreted as the relevant policy effect from the perspective of a university administrator. In fact, while changing the number of officially enrolled students (*enrolled students*) is a relatively easy task, the enforcement and manipulation of the actual (size) *student count* would depend on the university's enforcement capabilities, which might vary across colleges. At a minimum, the reduced form estimates are of interest to Bocconi's administrators.

Regardless of how we measure class size (student count or enrolled students) or the estimation procedure used (OLS, IV, or reduced form), the computation of the standard errors of the estimated coefficients from equation 1 for correct inference poses some additional problems. First, the individual fixed effect η_i induces correlation across the observations that refer to the same student. Second, we also need to cluster the standard errors to take account of the fact that students in the same class-year-cohort-program cell share the same class size $size_{ijtcd}$.

We address the first problem by transforming the model in orthogonal deviations, a transformation that eliminates the individual effect η_i from the equation and, in a standard setting, also preserves homoskedasticity. Specifically, orthogonal deviations are computed as the difference between the individual observation and the mean of all future observations for the same individual, adjusted by a constant factor to ensure that all cross-sectional observations share the same variance within-group, see Arellano (2003). Like first differencing or the within-group transformation (the transformation used by most econometric software packages, like Stata, to produce fixed-effect estimators), also orthogonal deviations eliminate the individual unobservable fixed term from the model (η_i in our equation 1). Moreover, in a standard setting, transforming the model in first differences, within-group deviations or orthogonal deviations is equivalent in terms of asympthotic results. Nevertheless, in order to obtain the most efficient estimates, it is necessary to apply GLS to the first difference or within-group models while simple OLS is sufficient with orthogonal deviations.

As a consequence, orthogonal deviations allow to deal more flexibly with other complications in the covariance structure of the error terms, like clustering.

In fact, in the specific case of equation 1, homoskedasticity is not guaranteed in the transformed model because class size does not vary at the same level of the dependent variable. While academic performance varies at the level of the single student and across academic years, class size is constant for all students who are allocated the same identifier within the same year-cohort-program group. Hence, we cluster the standard errors of the transformed model at the correct level of the class-year-cohort-program cell (there are 72 such cells in total). Applying such a clustering with an alternative fixed-effect transformation, like first differences of within-group deviations, would be more complicated as one would have to simultaneously account for correlation in the error terms induced by the panel structure of the data as well as for the clusters.

In Section 4, we investigate the effect of class size on academic performance using a series of variants of equation 1: we look at heterogeneity of the effect of class size across different types of students, and we consider the effect of class composition on academic performance, in this latter case measures of class heterogeneity are added to 1. In all cases, the empirical strategy for the estimation of equation 1 and its variants remains the same.

When we investigate the effect of class heterogeneity, we construct the instruments for the actual class composition using information from the university administration about the original official class allocation of each single student and we construct the corresponding measures of heterogeneity (e.g. share of females) among students who were officially allocated to the same class.

In Section 5 we look at the effect of class size and class composition on labor market performance. In this case the choice of the empirical model is less obvious. While we observe only one outcome for each student (her wage after entering the labor market), we observe at least three different class sizes for each student in compulsory courses over her academic career. We choose the most obvious specification, where we include the average class size ($\overline{size}$) a student has been exposed to according to the following equation:

$$w_{icd} = \beta \ \overline{size}_{icd} + \delta X_{icd} + \epsilon_{icd} \tag{6}$$

where w_{icd} is the wage reported by student i in cohort c and degree program d, $\overline{size}$ is the average of the 3 class sizes a student has experienced in her first three academic years and X_{icd} is a large set of controls determined prior to a student's matriculation. These controls include gender, the score obtained in the cognitive entry test, household income, geographical residence, type of high school, plus survey wave, cohort and degree program fixed effects.

Similarly to equation 1, identification rests on the random allocation mechanism and we address the potential endogeneity with the same approach discussed above. The standard errors are clustered at the same level of variation of $\overline{size}_{icd}$, i.e. the intersection of cohort, degree program and the three class identifiers of each academic year.[18]

4 The effect of class size and class composition on academic performance

We estimate equation 1 both by OLS and IV, to account for the possible endogeneity in the class size measure as given by the student count. Later (Section 4.1) we investigate the interaction of class size with the individual characteristics of the student, to test whether some type of individuals benefits or suffers more from smaller classes. Finally, in Section 4.2 we estimate the direct effect of class composition on academic performance.

Table 5 reports our main results for academic performance. Using a simple linear specification (columns 1 to 4) we find a significant effect of class size on academic performance.[19] The OLS estimate in column 2 indicates that one additional student in the class reduces the individual mean grade in the corresponding academic year by 0.01 grade points over an average of 26 ($B+$ for US universities) and a standard deviation of 2.33. Since our estimates are often significant at the margin of the 95 percent level, we report both the standard errors and the p-values (in square parentheses). For example, the OLS estimate in column 2 is significant at the 93.1 percent level.

[18] Because averaging over three years of data reduces the amount of variation in course heterogeneity, we have not enough variation to directly identify the effects of heterogeneity on wages.

[19] The specification in column 1 of Table 5 does not include student's fixed effects, while column 2 does include these fixed effects. If class size were randomly assigned to students, adding a student fixed effect should result in an estimate of similar magnitude but smaller standard errors. Consistent with this hypothesis, the estimates in columns 1 and 2 are similar in magnitude, but that the standard error is substantially smaller in column 2 than in column 1.

The IV estimate is a bit larger in magnitude and equal to -0.017, although it is not statistically different from the OLS. As we expected, the F-test of the first stage is very strong (F-stat of 242) and it allows to rule out the usual concerns due to weak instruments. Finally, the reduced form estimate is between the OLS and IV. Both the IV and the RF estimates are significant at the 94 percent level.

To put the magnitude of the estimated effects into a better perspective, take the IV coefficient and consider the effect of increasing class size by one standard deviation (computed over the entire sample). This corresponds to approximately 20 students or about 15 percent over an average class size of around 131. Such a change would reduce the mean grade by about 0.34 points or about 0.15 of a standard deviation, an effect that is consistent with the existing literature that finds significant effects (see Angrist and Lavy (1999), Krueger (1999), Bandiera et al. (2008), Pinto Machado and Vera-Hernandez (2009)).

Looking at the heterogeneity of the effects by major, we get a substantial effect for the Economics major, point estimate -.027 significant at the 99 percent level which translates into a reduction of .46 points for standard deviation increase in the class size (17 students) or a 20 percent of a standard deviation decrease in the average grade. For the Management major we get a point estimate of -.014 with standard error of .009, in terms of the effect we have that a standard deviation increase in the class size (7 students) reduces the average grade by .1 or 4 percent of a standard deviation. The results for the Economics and Finance major are larger in terms of point estimate, .1 points reduction in the GPA (significant at the 99 percent level), which translates into a fall of 30 percent of a standard deviation for a standard deviation increase in the class size (7 students).[20]

[TABLE 5]

In addition, we also attempted to determine if the relationship between class size and student performance is non-linear. In a setting where class size is one of the inputs of a standard human capital production function with decreasing return to scale, we would find that the impact of class

[20] To keep the discussion concise, we focus on the results from the pooled regressions (across degree programs). Although the heterogeneity of the effects is quite interesting, we believe it to be beyond the scope of the current paper. The estimation results by degree program are available from the authors.

size flattens out at larger sizes. To determine if this model is consistent with our data, we ran two types of specifications. First, we estimated spline regressions where we allow the effect of class size to vary at each quartile of the distribution. Next, we included class size and its square as regressors. For each model, we estimated the OLS, IV, and reduced form specifications. In none of these specifications is it possible to reject the null that the relationship between class size and student performance is linear. The lack of evidence of non-linearities suggests that a possible mechanism for explaining the class size negative effects even in large classes is that suggested by Lazear (2001), where students are subject to disruption shocks that affect one single student and then propagate by disturbing the entire class (or students in a neighborhood of who is first affected). In that setting, the class size effect is indeed negative even for large classes if the probability of no-disruption is large, as one would expect among college students.[21] Essentially in the Lazear's model classroom teaching is a typical public good subject to the negative externalities produced by the students asking (meaningless) questions or simply chatting.

To conclude this section, we present some simple evidence to show that students themselves have the perception that larger classes are detrimental to their learning. As is now customary in most universities, Bocconi regularly administers evaluation questionnaires to its students for gathering their opinions about various aspects of the teaching environment. In particular, Bocconi students are asked to indicate on a scale from 1 (disagree) to 5 (agree) if they agree with the following statement: *"the number of students in the classroom allows all the teaching activities to be regularly and efficiently carried out."* In Figure 4 we plot the average answer to this question in each of the 72 class-year-program-cohort cell against the corresponding class size measured either by the students count (upper panel) or by the number of officially enrolled students (lower panel).

As the figures clearly show, the two variables are negatively (and significantly at the 95 and 89 percent levels, respectively) correlated, for example a standard deviation increase in the Students count variable results into a .1 points fall in that measure over a mean of 3.8. Notice that the R-squared of these simple regressions are however relatively small (9.3 percent when the students

[21] Borrowing from Lazear (2001), if the probability that a student is not disrupting her or others' learning is p then the probability that disruption takes place in a classroom of n students is $1 - p^n$, which behaves essentially linearly when $p \to 1$ even when the class-size is between 1 and 200 students.

count is used as a measure of class size and 4.3 percent when we use the number of officially enrolled students).

[FIGURE 4]

4.1 Heterogeneous effects

After having established that class size reduces student achievement, we now explore the heterogeneity of the effect across students of different ability (as measured by the pre-enrollment admission test), gender, and family income. These results are interesting for at least two reasons. First, studying the heterogeneity of the effect is informative of the distributional consequences of lowering class size. If, for example, students from poorer families benefited more than others, reducing class size could be an efficient means of redistribution. Second, if school administrators face a budget constraint and cannot provide small classes for all students, they may want to allocate spots in small classes to students who are likely to benefit the most.

Table 6 reports the results obtained by augmenting our basic specification (columns 1 to 3 of Table 5) with interactions of class size and three crucial characteristics of the students: ability, gender, and income. The OLS estimates are never significant, while in both the IV and the reduced form estimation we find that the negative effect of larger classes essentially disappears for female and students from wealthier families.

[TABLE 6]

One possible explanation for the above results is that students from wealthier families can access remedial tools for less effective lectures (such as better textbooks, better study environment, remedial private teachers). Given that females have more pro-social behavior in general (they drink less (Sloan et al., 1995), they smoke less (Gruber, 2001), they commit less crime (Ludwig, 2001)), they may also be less disruptive in the class. If there is some degree of clustering of study mates by gender, women may suffer less from disruption because they sit close and interact more with other girls than with boys, who disrupt more.[22] This interpretation would also be consistent with the results that we

[22] The literature also documents that women are more risk averse (Schubert et al., 1999) and "shy-away" from competition (Gneezy et al. (2003)).

obtain in the next section on class composition. A puzzling result is that the OLS estimates appear to be all insignificant while the IV results on class size are similar to the earlier ones, although with larger standard errors.

4.2 Class Heterogeneity

In this section we explore how the heterogeneity of a student's peers influences her academic performance. This exercise is interesting for at least two reasons. First, growing evidence suggests that the composition of one's peer group is an important determinant of individual behavior and of students' achievement.[23] Further, recent evidence in Duflo et al. (2008) shows that tracking has positive effects on students' performance; in the same spirit Pinto-Machado and Vera-Hernandez (2009) find positive and heterogeneous effects of peers' ability on student performance. Cooley (2009) finds evidence of peer effects on performance within race-based reference groups.

Here, we investigate whether the composition of classmates in terms ability, income, and gender affects student performance. We compute a measure of dispersion for ability, income, and gender for each class. Because of the process of repeated random allocation that we described earlier (Section 2.1), each student is exposed to a different set of randomly selected peers in each academic year. For each of these groups, we compute the fraction of female in the class, the fraction of students from high income families and the mean and the standard deviation of the (log) entry test score for those effectively in the classroom (similarly to the student count definition).[24] Further, having obtained from the administration the original class identifier assigned to each student by the random allocation mechanism, we can produce the same measures of class heterogeneity based on this purely exogenous and theoretical class composition (similarly to what we do for the enrolled students measure). Hence, in Table 7 we report both the OLS, IV, and reduced form estimates.

[TABLE 7]

In columns 1 to 3, we concentrate on a simple linear specification, and we find that a larger share

[23] See the large literature on social interactions and peer effects summarized by Jackson (2008).

[24] Notice that the mean proportion of females and high income students are sufficient statistics for the distribution of these dichotomous variables within each class. The entry test score, instead, is a continuous variable therefore we compute both the mean and the standard deviation.

of female students in the class is beneficial for academic achievement: increasing the percentage of females in an average class (which is approximately 40 percent female) by 10 percentage points increases GPA of the average student by 0.14-0.15 of a grade points or 0.05-0.06 of a standard deviation. Increasing the fraction of high-income students has the opposite effect: adding 10 students of this type to an average class (which has approximately 28 out of 130 students) reduces the GPA of the average classmate by 0.16 grade points or 0.07 of a standard deviation, although this effect disappears in both the IV and the reduced form specifications.

In the following columns 4 to 6 of Table 7 we experiment with a simple quadratic specification to determine if the effect of class composition is linear. In the OLS results, both the linear and the quadratic effect of the dispersion in ability (measured by the standard deviation of the log entry test score) are significant, suggesting that more diverse classes perform better but that such effect is decreasing. The results for gender composition are qualitatively similar; for classes with the average share of female students, an increase in the fraction of female students increases performance, but the marginal effect declines (and eventually becomes negative) as the share of female students increases. Finally, the incidence of high income students still has no impact on performance. In columns 5 and 6, we replicate the estimates using our IV and reduced form specifications. Here, only the effects of gender composition remain significantly different form zero, although the sign and magnitude of the test score results are broadly similar across specifications.

Our results for gender composition highlight the importance of estimating non-linear effects. The results from the linear specification, columns (1)-(3), suggest that for our sample, increasing the share of female students increases performance. Because the students in our sample are predominately (58 percent) male, these results are consistent with at least two different hypotheses: (a) students always learn better when the share of female students increases (b) students tend to learn best when the ratio of males and females is approximately even. Our quadratic results cast doubt on (a). Taken at face value, these coefficients suggest that the optimal gender composition is 49.4 percent female.[25]

[25] The baseline results suggest that the effect of share female are given by: $5.836 \cdot (\text{share female}) - 5.895 \cdot (\text{share female})^2$. This function is maximized when $5.836 = 2 \cdot 5.895 \cdot (\text{share female})$ or when share female $= .494$.

To give a sense of the magnitude of these non-linear estimates, Figure 5 plots the marginal effects of our three measures of class composition (dispersion in ability, gender composition and income composition) derived from both the linear (corresponding to the estimates in columns 1 to 3 of Table 7) and the quadratic (corresponding to the estimates in columns 4 to 6 of Table 7) specifications. In the left panels of Figure 6 we show the OLS results while the IVs are plotted in the right panels.

[FIGURE 5]

The OLS quadratic effect of the dispersion in test scores (top left panel) shows that increasing the diversity in ability among classmates by one standard deviation (approximately 0.015) from the mean increases performance by 0.56 of a grade point or 1/4 of a standard deviation. Performing the same exercise, i.e. increasing test dispersion by one standard deviation, starting from an already diversified class, say one with dispersion in ability that is 2 standard deviations above the mean (corresponding approximately to the top 5 percent of the distribution), increases performance by 0.48 of a grade point, that is about 15 percent less than before. The entire effect of test score dispersion, however, disappears under the IV specification due to large standard errors.

Our results clearly show that gender composition has a robust and large effect on performance. The point estimates on the share of females remain significant in all specifications: OLS and IV, linear and quadratic. The marginal effects are plotted in the middle panels of Figure 6. Let us focus on the IV specification (although the OLS estimates are not substantially different, and using a Hausman test, we cannot reject the null of equality between the OLS and IV) and notice that increasing the percentage of female classmates by one standard deviation (approximately 0.04) from the mean (which is equal to about 40 percent) increases performance by 0.23 of a grade point or 10 percent of a standard deviation. Performing the same exercise, i.e. increasing the incidence of females by one standard deviation, starting from a class that is already female dominated, say one with female incidence that is 2 standard deviations above the mean (corresponding approximately to the top 5 percent of the distribution), increases performance by 0.19 of a grade point, that is about 17 percent less than before.

The results for income dispersion are significant only in the linear OLS specification and indicate

that a larger share of high income classmates reduces performance. However, such effect disappears in all other specifications.

The positive effect of the incidence of female students in the class seems consistent with the idea that girls have a more pro-social behavior and, are less disruptive. Along the same lines, the negative effect of wealthier students may be rationalized by arguing that, given their ability to make up for less productive lectures with private resources, they may be more prone to disruption to the detriment of the entire class. The positive effect of dispersion in ability indicates that students' skills are complements in the classroom production function.

The non-linearities in some of these effects open up the possibility reshuffling students in classrooms and increasing average performance without necessarily requiring additional resources. We return to the issue of optimal class formation in Section 6.

5 The effect of class size on labor market performance

In this section we test whether the negative effect of class size on academic performance affects labor market outcomes around 18 months after graduation. The literature on school resources and labor market performance (Moffit, 1996; Hanushek, 2006) finds a substantial positive effect of school resources, measured as class size or teacher per pupil ratios. As explained in Section 2.3, we observe our students once they graduate, typically around one and a half years after graduation and, although we have no longer term outcomes, we believe it is important to study the short-run impacts (Oyer, 2006).

Given that our students are assigned to a different class in each of the 3 years of required courses, the most natural way to specify our empirical model is to consider the average of those 3 class sizes as our measure of treatment. The results are reported in Table 8, where we produce estimates of equation 6 using a variety of specifications. In columns 1 and 2 we adapt the estimation procedure to the original wage information (recorded in intervals), and we apply interval regression. To avoid the technicalities involved in adopting an IV procedure in this model, we report only results computed using either the students count or the enrolled students as a measure of class size. In the following

columns (3 to 5), we use as a dependent variable a continuous version of the wage information computed at the mid points of the intervals indicated by each respondent. Then, we can apply the standard techniques and we report OLS, IV and reduced form estimates.

The estimates reported in the first 5 columns of Table 8 indicate that the effect of the average class size in college on entry wages is negative and of non-trivial magnitude across all specifications, although the effect is not always precisely estimated.

[TABLE 8]

The magnitude of the coefficients suggests that an increase of 20 students in the size of the average class would reduce monthly wages by approximately 80 to 85 euros on average or around 115 USD or 6 percent over the average monthly wage. This is a substantial effect, particularly if such a penalty is never recovered over the course of one's working life, as suggested by Oyer (2006). In the last 5 columns of Table 8 we repeat all the estimates by conditioning on academic performance. Interestingly, across specifications, the magnitude of the effects decreases by approximately 10 percent, suggesting that class size affects labor market outcomes both through its impact on academic performance and also independently through some other mechanism, possibly the development of non academic skills.[26]

6 The *optimal* class allocation

A crucial policy question is whether an optimal class can be designed by the administrators in terms of size and composition.[27] We address this question given the estimated parameters from Section 4.

First, we must take a stand on the planner's objective function and on the constraints she faces. For simplicity, we assume that the planner seeks to maximize the sum of individual expected

[26] This result is robust to controlling flexibly for the graduation mark, e.g. quintiles of the graduation mark.

[27] One has to bear in mind the caveats highlighted by Carrell et al. (2009b) who performed an experiment on class composition by constructing classes that were "optimal" according to the experimental estimates of exogenous peer effects from Carrell et al. (2009a). Carrell et al. (2009b) found that instead of improving students performance there actually was a fall in it. The authors explain the surprising result with changes in the interaction process induced by the extreme sorting applied. Our analysis instead predicts "optimal" composition that are well within the sample variation and clearly do not support extreme sorting of students.

performances, although we recognize that there are many other reasonable objective functions, i.e. maximum performance, dispersion of the distribution or other objectives.

We further assume that the total number of classrooms and teachers, the size of each classroom, and the student population are fixed. This assumption corresponds to solving the short run problem where resources are supplied inelastically. If we solved the problem under several different scenarios, these results would also help the social planner optimally allocate resources to higher education. Here, the planner will change the type of students in each classroom, keeping the number of classes and professors fixed.

We write student $i's$ expected performance in class j as follows:

$$P_{ij} = \alpha size_j + \beta_1 female_j + \beta_2 female_j^2 + \gamma_1 \sigma\left(ability\right)_j + \gamma_2 \left(\sigma\left(ability\right)_j\right)^2.$$

So that the planner's objective function is:

$$\begin{aligned}\sum_j w_j P_j &= \alpha \sum_j w_j size_j + \beta_1 \sum_j w_j fem_j + \beta_2 \sum_j w_j fem_j^2 \\ &\quad + \gamma_1 \sum_j w_j \sigma(abil)_j + \gamma_2 \sum_j w_j \left(\sigma(abil)_j\right)^2 = M\end{aligned}$$

where w_j is the size of class j.

Hence the full program is:

$$\begin{aligned}\max_{fem,\sigma(abil)} \quad & M = \alpha \sum_j w_j size_j + \beta_1 \sum_j w_j female_j + \beta_2 \sum_j w_j fem_j^2 + \gamma_1 \sum_j w_j \sigma(abil)_j \\ & + \gamma_2 \sum_j w_j \left(\sigma(abil)_j\right)^2 \\ s.t. \quad & N = \overline{N} \\ & size_j = \overline{N_j} \\ & E[fem_j] = \frac{\overline{F}}{\overline{N}} = f \\ & abil_i \sim \Phi(abil).\end{aligned}$$

where Φ is the appropriate cdf.

A simple example will help clarify how our estimates can inform the allocation of students into classes based on gender. Assume that the administrators have three classrooms of fixed size. Let these sizes be given by N_1, N_2, and $(N - N_1 - N_2)$, respectively. To simplify the analysis so that it focuses specifically on the gender margin, assume further that ability does not influence performance $(\gamma_1 = \gamma_2 = 0)$, or that the ability of males and females is drawn from an identical distribution and that the planner does not know the ability of the individual student. Because the size of the classroom j is fixed at $\overline{N_j}$, we can re-write the problem as solving for the share of females in each classroom $(f_j = \frac{F_j}{\overline{N_j}})$. This transformation ensures that our results are directly comparable to our empirical section. We need to define only $J - 1$ classrooms as the Jth one would be just the complement to the others. Let's also define $f = \frac{F}{N}$ as the fraction of females in the population of interest.

The solutions to this problem are:

$$
\begin{aligned}
f_1^* &= \frac{2\beta_2 n_1 F - \beta_1 N\left(1 - n_1(3 - 2n_1 - 2n_2) - 2n_2(1 - n_2)\right)}{2\beta_2 N\left(1 + 2n_1(n_1 + n_2 - 1) + 2n_2(n_2 - 1)\right)}, \\
f_2^* &= \frac{2\beta_2 n_2 F - \beta_1 N\left(1 - n_1(2 - 2n_1 - 3n_2) - 2n_2(1 - n_2)\right)}{2\beta_2 N\left(1 + 2n_1(n_1 + n_2 - 1) + 2n_2(n_2 - 1)\right)}, \\
f_3^* &= \frac{f - f_1^* - f_2^*}{1 - n_1 - n_2}.
\end{aligned}
$$

For example, where $N = 1000, \quad F = 400, \quad n_1 = .5, \quad n_2 = 1/3, n_3 = 1/6$, with the estimated $\widehat{\beta_1} = 5.836, \widehat{\beta_2} = -5.895$ would give $f_1^* = .37, f_2^* = .41, f_3^* = .46$ and $F_1^* = 186, F_2^* = 138, F_2^* = 76$.

The same type of optimization can be performed taking into account the heterogeneity in ability, as well as changing the number and size of the classrooms. One would need to take into account the actual joint distributions of the choice variables. Further, although we fixed the student body in this exercise, nothing prevents an institution from adjusting along several dimensions subject to some budget constraints. For example, from our analysis it is obvious that a larger share of women would benefit the overall performance. Given the estimated parameters, the share of females which maximizes performance is given by the ratio $-\frac{\widehat{\beta_1}}{2\widehat{\beta_2}} \approx .495$ while the current share of women is .42.

7 Conclusions

In this paper we investigate the effects of two controversial policies, class size and class composition, on student performance in school and in the labor market. We contribute to a large literature on policy interventions designed to improve student outcomes by adopting a novel approach that differs from most of the existing research in four important ways. First, we focus on university education rather than primary or secondary schooling. Because the pedagogy, average class size, and student population differ in important ways between university and pre-university education, we believe that these results provide evidence more directly applicable to higher education. In addition, Heckman (2007) argues that interventions early in life are likely to have larger impacts on the development of human capital than are interventions later in life, raising the possibility that effectiveness of educational interventions may differ between pre-tertiary and a university settings. Second, we rely on random variation in the size and composition of the classes. Such randomization was not the intended purpose of the administrators, hence our design helps avoiding concerns that teachers and students alter their behavior because of the experiment itself: the Hawthorne effect. Third, our paper studies the impact of class size and student heterogeneity on labor market outcomes rather than just on test scores. Finally, we provide a useful example on the construction of the *optimal* class composition.

Our results suggest four findings. First, class size has a small but substantial impact on student academic performance. A reduction in class size by 20 students increases the average grade by 0.1 standard deviations. Second, we show that the effect of class size on student performance is larger for men and for lower-income students. Third, we show that a larger share of females has, up to a certain threshold, a positive impact on average grades, i.e. performance is inverse U-shaped in the share of females. The same can be said in terms of ability heterogeneity: some heterogeneity improves the average performance, but a very heterogeneous class is detrimental. In contrast, we find no evidence that heterogeneity in family income has an effect on performance.

Finally, we turn to labor market outcomes. Our baseline results suggest that increasing class size by 20 students reduces a student's wage by approximately 6 percent. Given this estimate, it would

be hard to dismiss class size reduction as an ineffective and inefficient policy. Suppose that the 1,500 students at Bocconi were divided in 14 rather than the actual 12 classes, so that average class size would be reduced by 20 students. Such an intervention would generate a gain of 80 euros *per month* $\times$ 1,500 students, or 120,000 euros in total *each month*, which are likely to be more than enough to pay the costs of acquiring the additional resources necessary to activate the two extra classes. Further, we provide evidence that a *zero-cost* intervention, e.g. reshuffling the class composition in terms of share of females would increase overall average performance.

References

[1] Angrist, Joshua and Lavy, Victor, (1999), "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement," The Quarterly Journal of Economics, 114: 533-575.

[2] Arellano, Manuel, (2003), Panel Data Econometrics. Oxford University Press.

[3] Bandiera, Oriana, Larcinese, Valentino, and Rasul, Imran, (2008), "Heterogeneous Class Size Effects: New Evidence from a Panel of University Students," forthcoming, the Economic Journal.

[4] Betts, Julian, (1995), "Does School Quality Matter? Evidence from the National Longitudinal Survey of Youth," Review of Economics and Statistics, 77: 231-247.

[5] Campbell, Ernest, Coleman, James, Hobson, Carol, McPartland, James, Mood, Alexander, Weinfeld, Frederic and York, Robert, (1966), "Equality of Educational Opportunity," Washington DC: U.S. Government Printing Office.

[6] Card, David. and Krueger, Alan., (1992), "Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States," Journal of Political Economy, 100: 1-40.

[7] Card, David. and Krueger, Alan., (1996), "School Resources and Student Outcomes: An Overview of the Literature and New Evidence from North and South Carolina," The Journal of Economic Perspectives, 10: 31-50.

[8] Carrell, Scott, Fullerton, Richard. and West, James, (2009a): "Does Your Cohort Matter? Estimating Peer Effects in College Achievement," Journal of Labor Economics, 27, 439-464.

[9] Carrell, Scott, Sacerdote, Bruce and West, James, (2009b), "Beware of Economists Bearing the Reduced Forms? An Experiment in How Not To Improve Student Outcomes," mimeo UC-Davis.

[10] Cooley, Jane, (2009), "Desegregation and the Achievement Gap: Do Diverse Peers Help?" mimeo University of Wisconsin-Madison.

[11] Dearden, Lorraine, Ferri, Javier and Meghir, Costas, (2002), "The Effect of School Quality on Educational Attainment and Wages," Review of Economics and Statistics, 84, 1-20.

[12] De Giorgi, Giacomo, Redaelli, Silvia and Pellizzari, Michele, (2009), "Be as Careful of the Books You Read as of the Company You Keep. Evidence on Peer Effects in Educational Choices," NBER DP: 14948.

[13] De Giorgi, Giacomo and Pellizzari, Michele, (2009), "Understanding Peer Effects," mimeo Stanford University.

[14] Dobblesteen, Simone, Levin, Jesse and Oosterbeek, Hessel, (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, 17-38.

[15] Duflo, Ester, Dupas, Pascaline and Kremer, Michael, (2008), "Peer Effects and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya," forthcoming, American Economic Review.

[16] Duflo, Ester, Dupas, Pascaline and Kremer, Michael, (2009): "Inputs versus Accountability: Experimental Evidence from Kenya," mimeo UCLA.

[17] Figlio, David and Page, Marianne, (2002), "School Choice and the Distributional Effects of Ability Tracking: Does Separation Increase Inequality?" Journal of Urban Economics, 51: 497-514.

[18] Gneezy, Uri, Niederle, Muriel and Rustichini, Aldo, (2003), "Performance in Competitive Environments: Gender Differences," The Quarterly Journal of Economics, 118: 1049 - 1074.

[19] Gruber, Jonathan (2001), "Tobacco at the Crossroads: The Past and Future of Smoking Regulation in the United States," The Journal of Economic Perspectives, 15: 193-212.

[20] Guryan, Jonathan, Kroft, Korry and Notowidigo, Matthew, (2009), "Peer Effects in the Workplace: Evidence from Random Groupings in Professional Golf Tournaments," forthcoming AEJ: Applied.

[21] Hanushek, Eric, (1996), "School Resources" in Eric A. Hanushek and Finis Welch (ed.), Handbook of the Economics of Education, Amsterdam: Elsevier.

[22] Hanushek, Eric, (2006) "School Resources" in E. Hanushek and F. Welch (ed.) Handbook of the Economics of Education. Amsterdam: Elsevier.

[23] Hanushek, Eric, Kain, John and Rivkin, Steven, (1995), "Teachers, Schools, and Academic Achievement," Econometrica, 73: 417-458.

[24] Heckman, James, (2007). "The Economics, Technology, and Neuroscience of Human Capability Formation," Proceedings of the National Academy of Sciences, 104: 13250–13255.

[25] Heckman, James, Layne-Farrar, Anne, and Todd, Petra, (1996), "Human Capital Pricing Equations with an Application to Estimating the Effect of Schooling Quality on Earnings," Review of Economics and Statistics, 78: 562-610.

[26] Hoxby, Caroline, (1998), "The Effects of Class Size and Composition on Student Achievement: New Evidence from Natural Population Variation," NBER Working Paper 6869.

[27] Hoxby, Caroline, (2000), "The Effects of Class Size on Student Achievement: New Evidence from Population Variation," Quarterly Journal of Economics, 115(4): 1239-1285.

[28] Imbens, Guido, and Angrist, Joshua, (1994), "Identification and Estimation of Local Average Treatment Effects," Econometrica, 62: 467-475.

[29] Jackson, Mathew, (2008) "Social and Economic Networks," Princeton: Princeton University Press.

[30] Johnson, George and Stafford, Frank, (1973), "Social Returns to Quantity and Quality of Schooling," Journal of Human Resources 8: 139-155.

[31] Krueger, Alan, (1999), "Experimental Estimates of Education Production Functions Experimental Estimates of Education Production Functions," The Quarterly Journal of Economics, 114: 497-532.

[32] Krueger, Alan, Whitmore, Diane, (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," The Economic Journal, 111(468): 1-28.

[33] Lazear, Edward, (2001): "Educational Production," The Quarterly Journal of Economics, Vol. 116: 777-803

[34] Ludwig, Jens, Duncan, Greg and Hirschfield, Paul, (2001), "Urban Poverty and Juvenile Crime: Evidence from a Randomized Housing-Mobility," The Quarterly Journal of Economics, 116: 655-679.

[35] Manning, Alan and Pischke, Jörn-Steffen, (2006), "Comprehensive versus Selective Schooling in England and Wales: What do we Know?" NBER WP: 12176.

[36] Moffitt, Robert, (1996), "Symposium on School Quality and Educational Outcomes: Introduction," The Review of Economics and Statistics, 78: 559-561.

[37] OECD, (2009), Education at a Glance 2008, Paris.

[38] Oyer, Paul, (2006), "Initial Labor Market Conditions and Long-Term Outcomes for Economists," Journal of Economic Perspectives, 20: 143-160.

[39] Pinto Machado, Matilde and Vera-Hernandez, Marcos, (2009): "Peer Effects and Class Size in College," mimeo UCL.

[40] Schubert, Renate, Brown, Martin, Gysler, Matthias and Brachinger, Hans (1999), "Financial Decision-Making: Are Women Really More Risk-Averse?" The American Economic Review, 89: 381-385.

[41] Sloan, Frank, Reilly, Bridget, Schenzler, Christoph, (1995), "Effects of Tort Liability and Insurance on Heavy Drinking and Drinking and Driving," Journal of Law and Economics, 38: 49-77.

Table 1. Students' descriptive statistics

	All students		2000 Cohort		2001 Cohort	
	mean	**s.d.**	**mean**	**s.d.**	**mean**	**s.d.**
1=female	0.42	(0.49)	0.39	(0.49)	0.45	(0.50)
1=high income	0.22	(0.42)	0.22	(0.41)	0.23	(0.42)
entry test score	65.18	(15.85)	73.39	(13.85)	56.26	(12.77)
gpa	26.00	(2.33)	25.65	(2.40)	26.39	(2.20)
entry wage	1,303.37	(483.73)	1,242.88	(391.93)	1,358.38	(548.76)

Notes: High income families (above 90 thousands euros of gross yearly income, corresponding to approximately 140,000 USD) pay the maximum fee, hence they are not required to report their actual income to the university administration. Entry wages are originally recorded in intervals. The statistics reported here refer to an imputed measure of wages computed at the mid-point of the interval indicated by the respondent. All monetary values are in euros at current prices.

Table 2. Classes' descriptive statistics

		Students count							Officially enrolled students						
	Academic year:	first		second		third		Overall	first		second		third		Overall
Degree program	**Cohort:**	2000	2001	2000	2001	2000	2001		2000	2001	2000	2001	2000	2001	
Economics	mean	92	61	92	63	85	56	75	91	69	90	72	90	74	81
	std.dev.	*3.09*	*9.37*	*2.12*	*7.68*	*2.12*	*1.94*	*16.28*	*0.80*	*8.22*	*1.62*	*8.49*	*2.47*	*0.71*	*10.73*
	min	90	54	90	58	84	55	54	91	64	89	66	88	73	64
	max	95	67	93	68	87	58	95	92	75	92	78	92	74	93
Management	mean	138	131	133	137	129	132	133	133	145	130	162	131	162	144
	std.dev.	*7.11*	*9.34*	*4.01*	*4.67*	*2.31*	*4.99*	*6.53*	*6.87*	*10.20*	*3.19*	*4.25*	*1.87*	*0.83*	*14.84*
	min	124	113	127	132	126	125	113	119	125	126	157	129	161	119
	max	146	142	140	147	132	138	147	142	155	136	168	133	164	168
Economics and Finance	mean	153	140	157	151	151	141	149	145	154	157	171	156	171	159
	std.dev.	*3.62*	*2.39*	*1.82*	*0.71*	*2.63*	*1.41*	*6.92*	*0.80*	*2.12*	*4.44*	*0.81*	*3.74*	*0.81*	*9.86*
	min	151	138	156	150	150	140	138	145	152	153	170	154	170	145
	max	156	142	158	151	153	142	158	146	155	160	172	159	172	172
Total	mean	133	121	130	127	125	121	126	128	133	128	148	128	149	136
	std.dev.	*20.77*	*29.32*	*20.45*	*30.68*	*20.78*	*30.68*	*25.39*	*18.63*	*31.36*	*20.40*	*35.97*	*20.42*	*35.33*	*28.56*
	min	90	54	90	58	84	55	54	91	64	89	66	88	73	64
	max	156	142	158	151	153	142	158	146	155	160	172	159	172	172

Notes: The students count is the number of students in any given cohort and year who have the same class identifier. The officially enrolled students is the number of students who were allocated to the same class by the university administration at the beginning of each academic year. 2000 corresponds to the 1999/2000 cohort, while 2001 corresponds to the 2000/2001 cohort. One observation per class (72 cells in total).

Table 3. Descriptive statistics of class composition

Variable		First year 2000	First year 2001	Second year 2000	Second year 2001	Third year 2000	Third year 2001	Overall
Female	mean	0.364	0.42	0.364	0.437	0.371	0.441	0.399
	s.d.	*0.059*	*0.0682*	*0.064*	*0.0764*	*0.0724*	*0.0695*	*0.074*
	min	0.237	0.299	0.209	0.318	0.226	0.323	0.209
	max	0.445	0.527	0.429	0.567	0.511	0.525	0.567
High-income	mean	0.227	0.242	0.231	0.241	0.232	0.243	0.236
	s.d.	*0.0441*	*0.0554*	*0.0466*	*0.0477*	*0.0357*	*0.0535*	*0.046*
	min	0.13	0.148	0.161	0.156	0.201	0.123	0.123
	max	0.298	0.318	0.316	0.322	0.298	0.343	0.343
SD Entry test	mean	0.21	0.244	0.208	0.243	0.208	0.24	0.226
	s.d.	*0.0261*	*0.0172*	*0.0242*	*0.0153*	*0.0222*	*0.0152*	*0.026*
	min	0.169	0.22	0.172	0.215	0.163	0.215	0.163
	max	0.248	0.278	0.249	0.271	0.24	0.263	0.278

Notes: One observation per class (72 cells in total). SD Entry test is the within class standard deviation in the entry test scores.

Table 4. Average class characteristics

	[1] % of female	[2] % of top income	[3] average test score
Panel A: F-test of equality of means across classes:			
Economics	F(1, 7)=1.26	F(1, 7)=0.65	F(1, 7)=1.05
	(0.299)	(0.448)	(0.339)
Management	F(1, 37)=0.96	F(1, 37)=1.17	F(1, 37)=0.70
	(0.471)	(0.343)	(0.674)
Economics & Finance	F(1, 7)=2.87	F(1, 7)=0.31	F(1, 7)=1.59
	(0.134)	(0.593)	(0.247)
Panel B: correlation of class size and average class characteristics			
Student count	0.000	0.001	0.000
	(0.001)	(0.001)	(0.001)
Officially enrolled	0.001	0.001	0.000
	(0.001)	(0.001)	(0.001)
Observations	72	72	72

Notes: The F-tests reported in Panel A are derived from regressions of the mean characteristics of the class on dummies for the class identifiers, controlling for cohort and academic year fixed effects. The coefficients reported in Panel B come from regressions run at the class level (72 obervations) with the average class characteristic on the LHS and the measure of class size on the RHS. All regressions include the full set of three-way interactions of cohort, degree program and academic year fixed effects. Robust standard errors in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 5. Class-size effects on academic performance

		Linear specification		
	No FE	**OLS**	**IV**	**RF**
	[1]	**[2]**	**[3]**	**[4]**
Student count	-0.012	-0.010*	-0.017*	-
	(0.008)	(0.005)	(0.009)	
	[0.181]	[0.069]	[0.060]	
Enrolled students		-	-	-0.014*
				(0.007)
				[0.061]
Observations		4,810	4,810	4,810
F-stats		-	241.87	-

Notes: Robust standard errors in parentheses, clustered by cohort-program-class-year cells. Corresponding p-values in square parentheses. Clustering at the cohort-program-class level produces very similar standard errors to the ones reported in the table. * significant at 10%; ** significant at 5%; *** significant at 1%

Table 6. Heterogeneous effects of class-size on students' academic performance

	OLS [1]	IV [2]	RF [3]
Student count	-0.003	-0.031	-
	(0.012)	(0.025)	
	[0.785]	[0.216]	
Test score x Student count	0.000	0.000	-
	(0.000)	(0.000)	
	[0.439]	[0.858]	
Female x Student count	0.001	0.023**	-
	(0.006)	(0.009)	
	[0.885]	[0.013]	
High income x Student count	0.005	0.026**	-
	(0.006)	(0.010)	
	[0.393]	[0.007]	
Enrolled students	-	-	-0.020
			(0.013)
			[0.135]
Test score x Enrolled students	-	-	0.000
			(0.000)
			[0.813]
Female x Enrolled students	-	-	0.010***
			(0.003)
			[0.003]
High income x Enrolled students	-	-	0.012***
			(0.003)
			[0.001]
Observations	4,810	4,810	4,810
F-stats	-	64.28; 51.62; 62.86; 84.22	-

Notes: Robust standard errors in parentheses, clustered by cohort-program-class-year cells. Corresponding p-values in square parentheses. Clustering at the cohort-program-class level produces very similar standard errors to the ones reported in the table.

* significant at 10%; ** significant at 5%; *** significant at 1%

Table 7. The effects of class heterogeneity on academic performance

	Linear effects			Quadratic effects		
	OLS	**IV**	**RF**	**OLS**	**IV**	**RF**
	[1]	**[2]**	**[3]**	**[4]**	**[5]**	**[6]**
Student count	-0.009	-0.018*	-	-0.012**	-0.020*	-
	(0.006)	(0.010)		(0.006)	(0.011)	
	[0.127]	[0.079]		[0.050]	[0.069]	
Heterogeneity of actual classmates:						
Mean test score in the class	0.034	-0.037	-	0.033	-0.033	-
	(0.022)	(0.035)		(0.023)	(0.035)	
	[0.128]	[0.286]		[0.150]	[0.337]	
S.d. of (log) test scores in the class	-1.729	-2.522	-	37.633**	26.375	-
[SD test]	(1.454)	(1.897)		(15.861)	(26.986)	
	[0.240]	[0.184]		[0.022]	[0.328]	
[SD test] squared	-	-	-	-85.994**	-63.646	-
				(33.711)	(56.855)	
				[0.014]	[0.263]	
Percentage of females in the class	1.402***	1.517**	-	5.836**	8.737**	-
[% female]	(0.483)	(0.648)		(2.883)	(3.737)	
	[0.006]	[0.019]		[0.049]	[0.019]	
[% female] squared	-	-	-	-5.895*	-9.161**	-
				(3.432)	(4.578)	
				[0.092]	[0.045]	
students in the class [% high-	-1.633*	0.087	-	-3.026	0.942	-
income]	(0.877)	(0.734)		(5.161)	(6.964)	
	[0.069]	[0.905]		[0.560]	[0.892]	
[% high-income] squared	-	-	-	3.704	-1.072	-
				(10.282)	(14.219)	
				[0.720]	[0.940]	
Enrolled students	-	-	-0.014*	-	-	-0.014*
			(0.008)			(0.008)
			[0.074]			[0.079]
Heterogeneity of officially enrolled students:						
Mean test score in the class	-	-	-0.034	-	-	-0.026
			(0.024)			(0.023)
			[0.103]			[0.264]
S.d. of (log) test scores in the class	-	-	-2.020	-	-	9.388
[SD test]			(1.578)			(14.260)
			[0.223]			[0.514]
[SD test] squared	-	-	-	-	-	-25.314
						(30.684)
						[0.414]
Share of females in the class	-	-	1.203**	-	-	8.648**
			(0.489)			(3.321)
			[0.003]			[0.012]
Share of females squared	-	-	-	-	-	-9.470**
						(4.246)
						[0.031]
Share of high income students in	-	-	-0.235	-	-	-0.535
the class			(0.577)			(4.492)
			[0.042]			[0.906]
Share of high-income squared	-	-	-	-	-	0.773
						(8.011)
						[0.923]
Observations	4,810	4,810	4,810	4,810	4,810	4,810

Notes: The first stage F-tests of the excluded instruments are the following. For the model in column 2: 41.61; 210.57; 61.82; 223.83; 101.51. For the model in column 3: 36.54; 149.62; 52.53; 50.69; 127.38; 86.00; 66.71; 81.49

Robust standard errors in parentheses, clustered by cohort-program-class-year cells. Corresponding p-values in square parentheses. Clustering at the cohort-program-class level produces very similar standard errors to the ones reported in the table. * significant at 10%; ** significant at 5%; *** significant at 1%

Table 8. Class-size effects on Wages

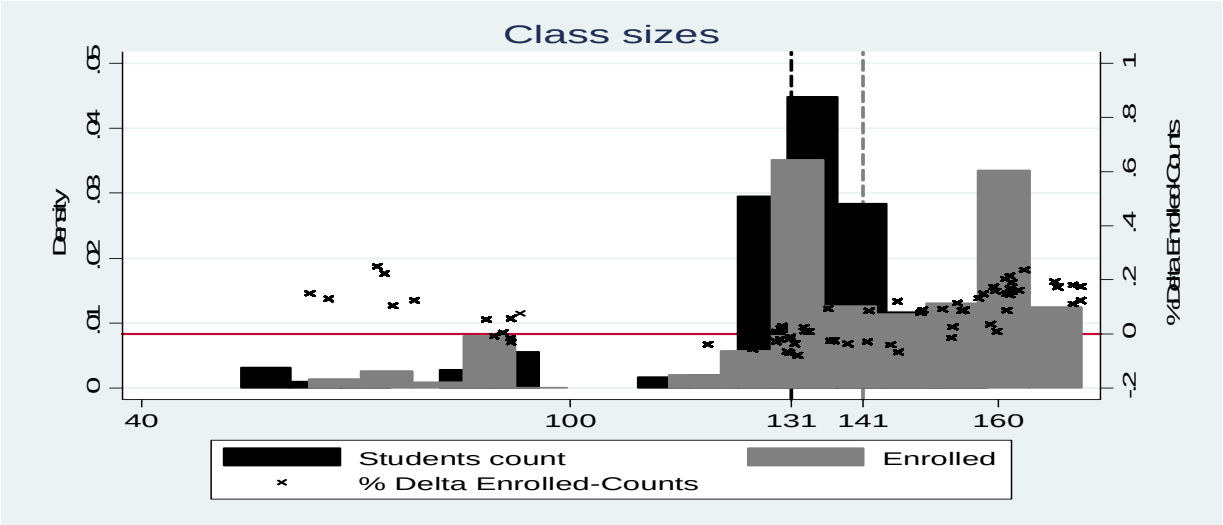
	Interval regression		OLS	IV	RF	Interval regression		OLS	IV	RF
	[1]	**[2]**	**[3]**	**[4]**	**[5]**	**[6]**	**[7]**	**[8]**	**[9]**	**[10]**
Average student count	-4.239*	-	-4.602**	-3.206	-	-3.927*	-	-4.285*	-2.858	-
	(2.334)		(2.343)	(2.551)		(2.295)		(2.303)	(2.512)	
	[0.069]		[0.050]	[0.209]		[0.069]		[0.063]	[0.255]	
Average enrolled students	-	-2.404	-	-	-2.257	-	-2.154	-	-	-2.012
		(1.761)			(1.802)		(1.738)			(1.776)
		[0.172]			[0.211]		[0.215]			[0.258]
Graduation mark	-	-	-	-	-	9.337***	9.357***	9.076***	9.106***	9.110***
						(2.131)	(2.137)	(1.961)	(1.952)	(1.968)
						[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Observations	1,075	1,075	1,075	1,075	1,075	1,075	1,075	1,075	1,075	1,075
F-stats	-	-	-	1,937.86	-	-	-	-	1,933.89	-

Notes: all models include the following set of controls: gender, entry test score, high school final grade, high school type, family income, original residence, cohort, degree program, survey wave.

Robust standard errors in parentheses, clustered by cohort-program-class cells. Corresponding p-values in square parentheses.

* significant at 10%; ** significant at 5%; *** significant at 1%

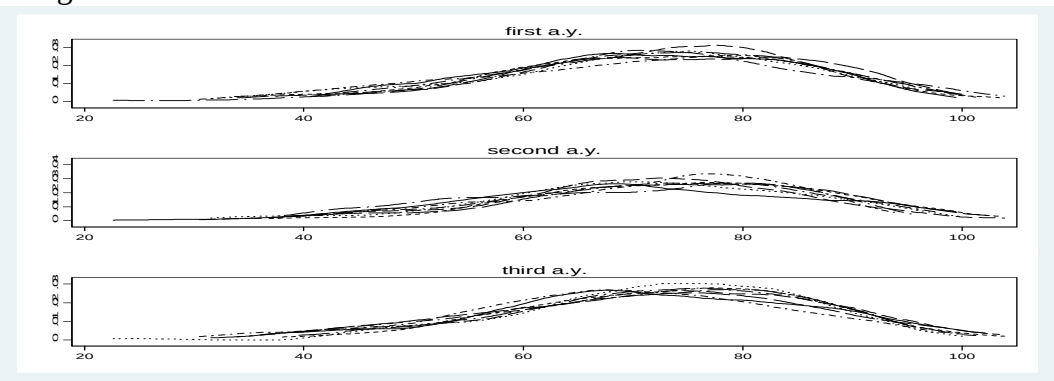
Figure 1. Variation in Class Size



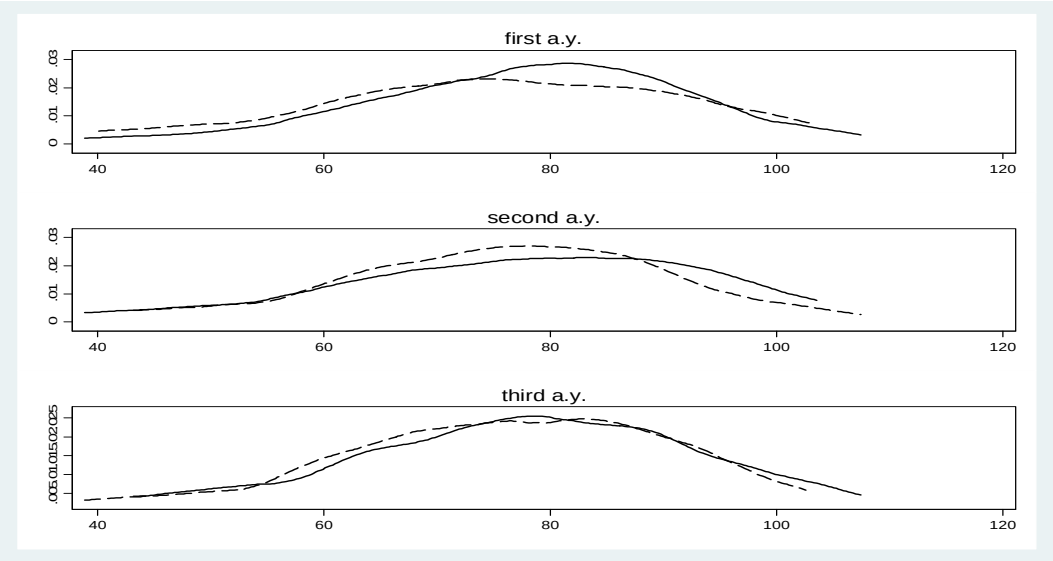
Notes: The darker and lighter bars indicate the density (left axis) of the different class sizes measured as *Students count* and *Enrolled* respectively. Dashed lines of corresponding colors indicate the averages of the two variables. The small x's in the graph show the percentage difference (right axis) between the *Enrolled* and *Students count* variables.

Figure 2. Test score distributions by class

Mangement



Economics



Finance

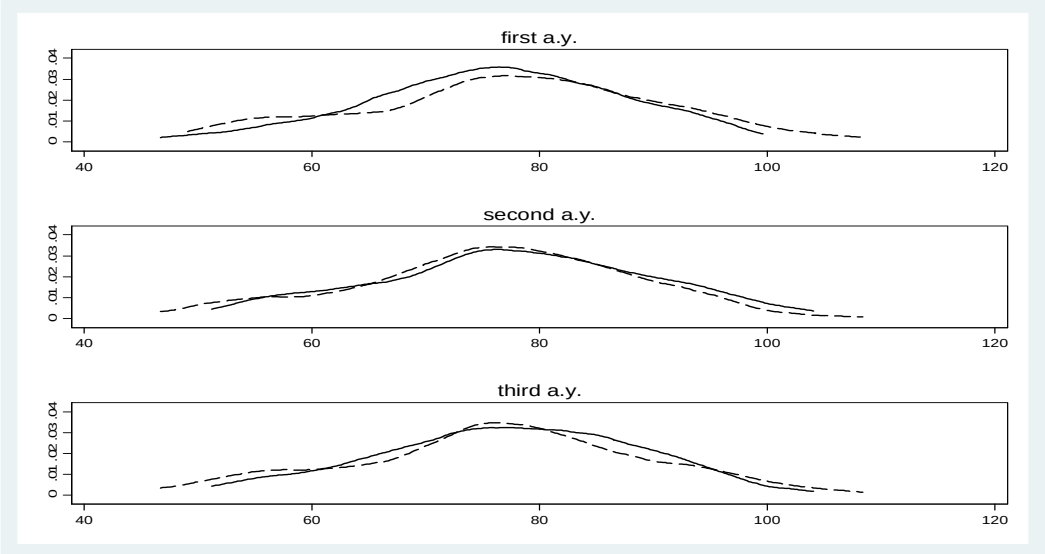
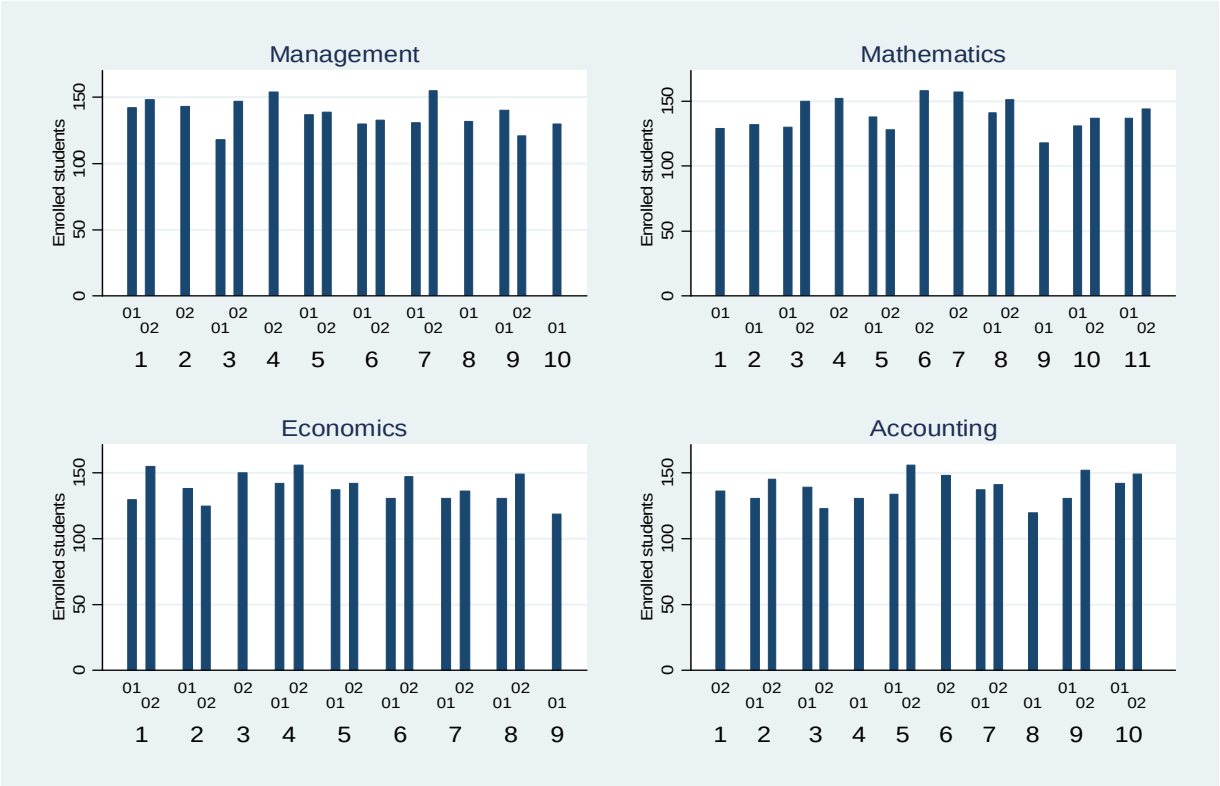
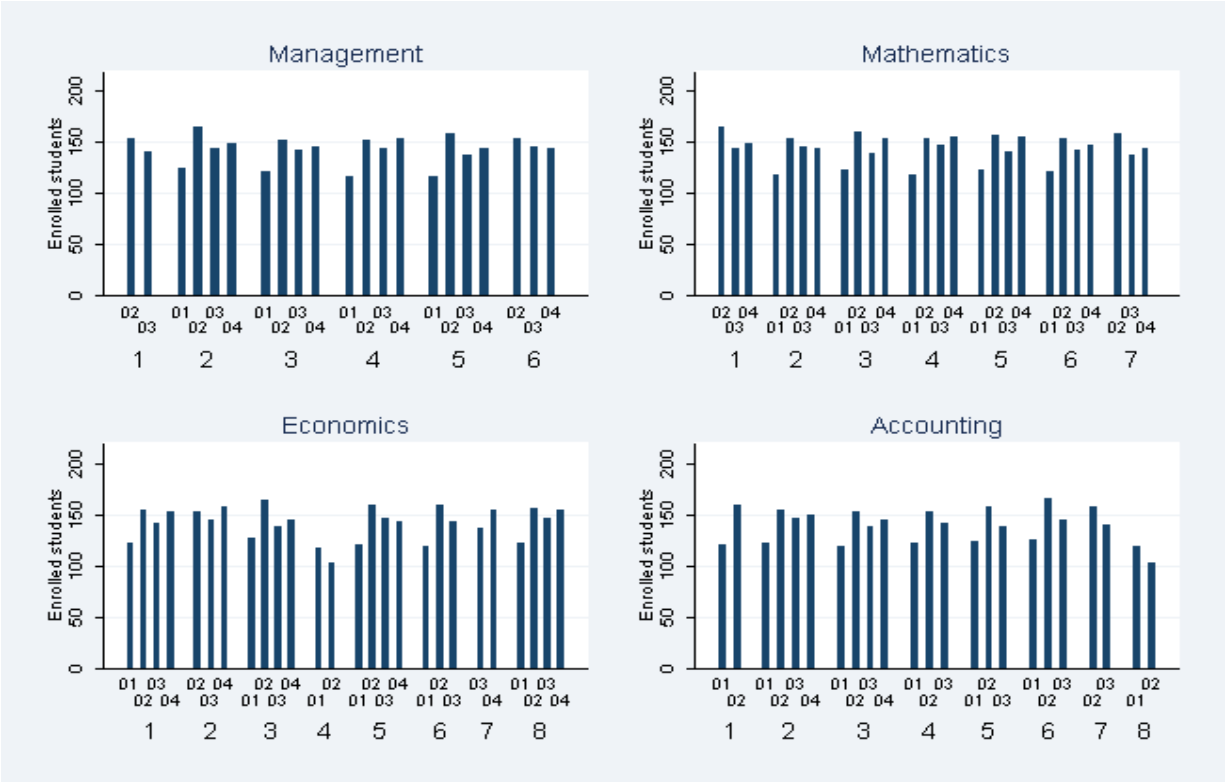


Figure 3. Teachers allocation

Panel A. Academic years 1999-2001



Panel B. Academic years 2001-2004



Notes: number of officially enrolled students in the classes of each teacher in different subject areas.