Evaluating Students' Evaluations *

Michela Braga
Università Statale di Milano

Marco Paccagnella
Bank of Italy and Bocconi University

Michele Pellizzari
Bocconi University, IGIER, IZA and C.F. Dondena Centre

First Version: September 2010

Current version: November, 2010

VERY PRELIMINARY: PLEASE DO NOT QUOTE.

*We would like to thank Bocconi University for granting access to its administrative archives for this project. In particular, the following persons provided invaluable and generous help: Giacomo Carrai, Mariele Chirulli, Mariapia Chisari, Alessandro Ciarlo, Alessandra Gadioli, Roberto Grassi, Enrica Greggio, Gabriella Maggioni, Erika Palazzo, Giovanni Pavese, Cherubino Profeta, Alessandra Startari and Mariangela Vago. We are also indebted to Tito Boeri, Giacomo De Giorgi, Marco Leonardi, Tommaso Monacelli and Tommaso Nannicini for their precious comments. We would also like to thank seminar participants at Bocconi University. The views expressed in this paper are solely those of the authors and do not involve the responsibility of the Bank of Italy. The usual disclaimer applies. *Corresponding author*: Michele Pellizzari, Department of Economics, Bocconi University, via Roentgen 1, 20136 Milan - Italy; phone: +39 02 5836 3413; fax: +39 02 5836 3309; email: michele.pellizzari@unibocconi.it.

Abstract

This paper contrasts measures of teacher effectiveness with the students' evaluations for the same teachers using administrative data from Bocconi University (Italy). The effectiveness measures are estimated by comparing the subsequent performance - both in follow-on coursework and in the labor market - of students who are randomly assigned to teachers in each of their compulsory courses. We find that, even in a setting where the syllabuses are fixed and all teachers in the same course present exactly the same material, teachers still matter substantially. The average difference in subsequent performance between students who were assigned to the best and worst teacher (on the effectiveness scale) is approximately 23% of a standard deviation in the distribution of exam grades, corresponding to over 3% of the average grade. Moreover, teacher effectiveness appears to be negatively correlated with the students' evaluations; in other words, teachers who are associated with better subsequent performance receive worst evaluations from their students. On the other hand, teachers who are associated with high grades in their own exam are associated with good evaluations. We rationalize these results with a simple model where teachers can either engage in real teaching or teach to the test, the former requiring higher students' effort than the latter. Teaching to the test guarantees high grades in the current course but does not improve future outcomes. Hence, if students provide higher evaluations to courses in which they get higher grades, the model is capable of predicting our empirical finding that good teachers get bad evaluations, especially when teaching to the test is very effective (for example, with multiple choice tests). Consistently with the predictions of the model, we also find that classes in which high-skill students are over-represented produce evaluations that are more in line with estimated teacher effectiveness.

JEL Codes: I20
Keywords: Teacher quality, Postsecondary Education.

1 Introduction

The use of anonymous students' evaluations to measure teachers' performance has become extremely popular in universities around the world (Becker and Watts, 1999). These normally include questions about the clarity of lectures, the logistics of the course, and many others. They are either administered to the students during a teaching session toward the end of the term or, more recently, filled on-line.

From the point of view of the university administration, such evaluations are used to solve the agency problems related to the selection and motivation of teachers, in a context in which neither the types of teachers, nor their level of effort, can be observed precisely. In fact, students' evaluations are often used to make hiring and promotion decisions (Becker and Watts, 1999) and, in institutions that put a strong emphasis on research, to avoid strategic behavior in the allocation of time or effort between teaching and research activities (Brown and Saks, 1987).[1]

The validity of anonymous students' evaluations rests on the assumption that students are in a better position to observe the performance of their teachers. While this might be true for the simple fact that students attend lectures, there are also many reasons to question the appropriateness of such a measure of performance. For example, the students' objectives might be different from those of the principal, i.e. the university administration. Students may simply care about their grades whereas the university (or parents or society as a whole) cares about their learning and the two (grades and learning) might not be perfectly correlated, especially when the same professor is engaged both in teaching and in grading the exams. Consistently with this interpretation, Krautmann and Sander (1999) show that teachers who give higher grades also receive better evaluations, a finding that is confirmed by several other studies and that is thought to be responsible for grade inflation (Carrell and West, 2010; Weinberg, Fleisher, and Hashimoto, 2009).

The estimation of teaching quality is complicated also becasuse it appears to be uncorrelated with observable teachers' characteristics (Hanushek and Rivkin, 2006; Krueger, 1999;

[1] Although there is some evidence that a more research oriented faculty also improve academic and labor market outcomes of graduate students (Hogan, 1981).

Rivkin, Hanushek, and Kain, 2005). Despite such difficulties, there is also ample evidence that teachers quality matters substantially in determining students achievement (Carrell and West, 2010; Rivkin, Hanushek, and Kain, 2005) and that teachers respond to incentives (Duflo, Hanna, and Kremer, 2010; Figlio and Kenny, 2007). Hence, understanding how professors should (or should not) be monitored and incentivized is of primary importance, the more so for economics departments, which appear to receive worse evaluations from the students compared to other disciplines (Cashin, 1990).

In this paper we evaluate the content of the students evaluations by contrasting them with 'hard' measures of teacher effectiveness. We construct such measures by comparing the performance in subsequent coursework and in the labor market of university students who are randomly allocated to different teachers in their compulsory courses. For this exercise we use data on a cohort of students at Bocconi University - the 1998/1999 freshmen - who were required to take a fixed sequence of compulsory courses and who where randomly allocated to a set of teachers for each compulsory course. Additionally, the data are exceptionally rich in terms of observable characteristics, in particular they include measures of cognitive ability and family income.[2]

We find that, even in a setting where the syllabuses are fixed and all teachers in the same course present exactly the same material, professors still matter substantially. The average difference in subsequent performance between students who were assigned to the best and worst teacher (on the effectiveness scale) is approximately 23% of a standard deviation in the distribution of exam grades, corresponding to over 3% of the average grade. Moreover, the effectiveness measures appear to be negatively correlated with the students' evaluations; in other words, teachers who are associated with better subsequent performance receive worst evaluations from their students. On the other hand, teachers who are associated with high grades in their own exams receive better evaluations.

We rationalize these results with a simple model where teaching is defined as the combination of two types of activities: real teaching and 'teaching-to-the-test', the former requiring higher students' effort than the latter. Practically, we think of real teaching as competent pre-

[2]The same data are used in De Giorgi, Pellizzari, and Redaelli (2010).

sentations of the course material with the aim of making students understand and master it, and teaching to the test as mere repetition of exam tests and exercises with the aim of making students learn how to solve them, even without fully understanding their meaning.

Professors are heterogeneous in their teaching methodology, i.e. in the combination of real teaching and teach-to-the-test. Grades are the outcome of teaching and are less dispersed the more the professor teaches to the test. The type of the exam defines the effectiveness of teach-to-the-test. To the one extreme, one can think of an exam as a selection of multiple-choice questions randomly drawn from a given pool. In such a situation, teaching to the test merely consists in going over all the possible questions and memorizing the correct answer. This is a setting in which teaching to the test can be very effective and lead to all students performing very well, regardless of their ability. The other extreme are essays, where there is no obvious correct answers and one needs to personally and originally elaborate on one's own understanding of the course's material; in this type of exam teaching to the test is unlikely to be particularly effective.

Students evaluate teachers on the basis of their utility levels, at least when they are asked about their general satisfaction with the course. We assume that students' utility depends positively on grades and negatively on effort. Further, we also introduce heterogeneity by assuming that good students face a lower marginal disutility of effort.

This simple model is able to predict our empirical findings, namely that good teachers get bad evaluations. This is more likely to occur with exam types that are more prone to teaching-to-the-test and when low ability students are over represented. Consistently with these predictions, we also find that the evaluations of classes in which high-skill students (identified by their score in the cognitive admission test) are over-represented are more in line with the estimated real teacher quality. Furthermore, the distributions of grades in the classes of the most effective teachers are more dispersed, which supports our specification of the learning function.

There is a large literature that investigates the role of teacher quality and teacher incentives in improving educational outcomes, although most of the existing studies focus on primary and secondary schooling (Figlio and Kenny, 2007; Jacob and Lefgren, 2008; Kane and Staiger,

2008; Rivkin, Hanushek, and Kain, 2005; Rockoff, 2004; Rockoff and Speroni, 2010; Tyler, Taylor, Kane, and Wooten, 2010). The availability of standardized test scores facilitates the evaluation of teachers in primary and secondary schools and such tests are currently available in many countries and also across countries. The large degree of heterogeneity in subjects and syllabuses in universities makes it impossible to design common tests that would allow to compare the performance of students who were exposed to different teachers, especially across subjects. At the same time, the large increase in college enrollment experienced in almost all countries around the world in the past decades (OECD, 2008) calls for a specific focus on higher education, as in this study.[3]

Only a couple of other papers investigate the role of students' evaluations in universities, namely Carrell and West (2010) and Weinberg, Fleisher, and Hashimoto (2009). Compared to these papers we improve in various directions. First of all, the random allocation of students to teachers in our setting differentiates our approach from that of Weinberg, Fleisher, and Hashimoto (2009), who cannot purge the estimates of teacher effectiveness from the potential bias due to best students selecting the courses of the best professors. Rothstein (2009) and Rothstein (2010) show that correcting such a selection bias is pivotal to producing reliable measures of teaching quality. The study of Carrell and West (2010) uses data from a U.S. Air Force Academy, while our empirical application is based on a more standard institution of higher education and is therefore more likely to be generalizable to other settings.[4] Additionally, our data allow us to cross-check the validity of our estimates of teacher effectiveness by looking at various students' outcomes, including graduation marks and entry wages. Finally, we also provide a theoretical framework for the interpretation of our results, which is absent in Carrell and West (2010).

[3]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. 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.

[4]Bocconi is a selective college that offers majors in the wide area of economics, management, public policy and law, hence it is likely comparable to US colleges in the upper part of the quality distribution. For example, faculty in the economics department hold PhDs from Harvard, MIT, NYU, Stanford, UCLA, LSE, Pompeu Fabra, Stockholm University. Recent top Bocconi PhD graduates landed jobs (either tenure trak or post-docs) at the World Bank and the University College of London. Also, the Bocconi Business school is normally ranked in the same range as the Georgetown University McDonough School of Business or the Johnson School at Cornell University in the US and to the Manchester Business School or the Warwick Business School in the UK (Financial Times Business Schools Rankings).

More generally, this paper is also related and contributes to the wider literature on performance measurement and performance pay. For example, one concern with students evaluations of teachers is that they might divert professors from activities that have a higher learning content for the students (but that are more demanding in terms of students' effort) and concentrate more on classroom entertainment (popularity contests) or change their grading policies. This interpretation is consistent with the view that teaching is a multi-tasking job, which makes the agency problem more difficult to solve (Holmstrom and Milgrom, 1994). Subjective evaluations, which have become more and more popular in modern human resource practices, can be seen as a mean to address such a problem and, given the very limited extant empirical evidence (Baker, Gibbons, and Murphy, 1994; Prendergast and Topel, 1996), our results can certainly inform also this area of the literature.

The paper is organized as follows. Section 2 describes our data and the institutional details of Bocconi University. Section 3 describes our strategy to estimate teacher effectiveness and presents the results. In section 4 we correlate such class effects with the students' evaluations. Section 5 presents some robustness cheks. In Section 6 we present a simple theoretical framework that rationalizes our results, while section 7 presents some additional evidence that corroborates our model. Finally, section 8 concludes.

2 Data and institutional details

The empirical analysis in this paper is based on data for one cohort of undergraduate students at Bocconi university, an Italian private institution of tertiary education offering degree programs in economics, management, public policy and law. We select the cohort of the 1998/1999 freshmen for technical reasons, being the only one available in our data where students were randomly allocated to teaching classes for *each* of their compulsory courses.[5] In later cohorts, the random allocation was repeated at the beginning of each academic year so that students

[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 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.

would take *all* the compulsory courses of each academic year with the same group of classmates (so that we would only be able to identify the joint effectiveness of the entire *set* of teachers).[6] For earlier cohorts the class identifiers, which are the crucial piece of information for our study, were not recorded in the university archives.

The students entering Bocconi in the 1998/1999 academic year were offered 7 different degree programs, although only three of them attracted a sufficient number of students to require the splitting of lectures into more than one class: Management, Economics and Law&Management[7]. Students in these programs were required to take a fixed sequence of compulsory courses for the entire duration of their first two years, for a good part of their third year and, in a few cases, also in their last year. Figure 1 lists the exact sequence for each of the three programs that we consider, breaking down courses by the term (or semester) in which they were taught and by subject areas (classified with different colors: red for management, black for economics, green for quantitative subjects, blue for law).[8] In Section 3 we construct measures of teacher effectiveness for the professors of these compulsory courses. We do not consider elective subjects, as the endogenous self-selection of students would complicate the identification of teaching quality.

[INSERT TABLE 1 ABOUT HERE]

Most (but not all) of the courses listed in Table 1 are taught in multiple classes or sections. The number of such classes varies across both degree programs and specific courses. For example, Management is the degree program that attracts the most students (over 85% in our cohort) who are normally divided into 8 to 10 classes for their compulsory courses. Economics and Law&Management students are much fewer and are rarely allocated to more than just two classes. Moreover, the number of classes also varies within degree programs depending on the number of available teachers. For instance, at the time of our cohort Bocconi did not

[6] De Giorgi, Pellizzari, and Woolston (2010) use data for these later cohorts for a study of class size.

[7] The other degree programs were Economics and Social Disciplines, Economics and Finance, Economics and Public Administration.

[8] Notice that Economics and Management share exactly the same sequence of compulsory courses in the first three terms. Indeed, students in these two programs did attend these courses together and made a final decision about their major at the end of the third term. De Giorgi, Pellizzari, and Redaelli (2010) study precisely this choice. In the rest of the paper we abstract from this issue and we treat the two degree programs as entirely separated. In the Appendix we present some robustness checks to justify this approach.

have a law department and all law professors were contracted from other nearby universities. Hence, the number of classes in law courses are normally fewer than in other subjects (e.g. 4 in Management). Similarly, since the management department was (and still is) much larger than the economics or the quantitative department, courses in the management areas were normally split in more classes than courses in other subjects.

Regardless of the specific class to which students were allocated, they were all taught the same material. In other words, all professors of the same course were required to follow exactly the same syllabus, although some variations across degree programs was allowed (i.e. mathematics was taught slightly more formally to economics students than Law&Management ones). Additionally, the exam questions were also the same for all students, regardless of their class. Specifically, one of the teachers in each course (normally a senior person) acted as coordinator for all the others, supervising that all classes progressed similarly during the term, defining changes in the syllabus and addressing specific problems that might have arisen. The coordinator also prepared the exam paper, which was administered to all classes. Grading was usually delegated to the individual teachers, each of them marking the papers of the students in his/her own class, typically with the help of one or more teaching assistants. Before communicating the marks to the students, the coordinator would check that there were no large discrepancies in the distributions across teachers.

[INSERT TABLE 2 ABOUT HERE]

Table 2 reports some descriptive statistics that summarize the distributions of (compulsory) courses and their classes across terms and degree programs. For example, in the first term Management students take 3 courses, divided into a total of 24 different classes: management I, which is split into 10 classes; private law, 6 classes; mathematics, 8 classes. The table also reports basic statistics (means and standard deviations) for the size of these classes.

Our data cover in details the entire academic history of the students in these programs, including their basic demographics (gender, place of residence and place of birth), high school leaving grades as well as the type of school, the grades in each single exam they sat at Bocconi together with the date when the exams were sat. Graduation marks are observed for all non-

dropout students.[9] Additionally, all students take a cognitive admission test as part of their application to the university and such test scores are available in our data for all the students. Moreover, since tuition fees depend on family income, this variable is also recorded in our dataset. Importantly, we also have access to the random class identifiers that allow us to identify in which class each students attended each of their courses.

[INSERT TABLE 3 ABOUT HERE]

Table 3 reports some descriptive statistics for the students in our data by degree program. The vast majority of them are enrolled in the Management program (88%), while Economics and Law&Management attract 11% and 14%. Female students are generally under-represented in the student body (43% overall), apart from the degree program in Law&Management. About two thirds of the students come from outside the province of Milan, which is where Bocconi is located, and such a share increases to 75% in the Economics program. Family income is recorded in brackets and one quarter of the students are in the top bracket, whose lower threshold is in the order of approximately 100,000 euros at current prices. Students from such a rich background are under-represented in the Economics program and over-represented in Law&Management. High school grades and entry test scores (both normalized on the scale 0-100) provide a measure of ability and suggest that Economics attracts the best students, a fact that is confirmed by looking at university grades, graduation marks and entry wages in the labor market.

Data on wages come from graduate surveys that we were able to match with the administrative records. Bocconi runs regular surveys of all alumni approximately one to one and a half years since graduation. These surveys contain a detailed set of questions on labor market experience, including employment status, occupation, and (for the employed) entry wages. As it is common with survey data, not all students respond to the survey but we are still able to match almost 60% of the students in our cohort, a relatively good response rate for surveys.[10]

[9]The dropout rate, defined as the number of students who, according to our data, do not appear to have completed their programs at Bocconi over the total size of the cohort, is just above 10%. Notice that some of these students might have transfered to another university or still be working towards the completion of their program, whose formal duration was 4 years. In Section 5 we perform some robustness check to show that excluding the dropouts from our calculations is not crucial for our results.

[10]The response rates are highly correlated with gender, because of compulsory military service, and with the

Finally, we complement our dataset with students' evaluations of teachers. Towards the end of each term (typically in the last week), students in all classes are asked to fill an evaluation questionnaire during one lecture. The questions gather students' opinions about and satisfaction with various aspects of the teaching experience, including the clarity of the lectures, the logistics of the course, the handiness of the professor and so on. For each item in the questionnaire, students answer on a scale from 1 (very negative) to 5 (very positive).

In order to allow students to evaluate their experience without fear of retaliation from the teachers at the exam, such questionnaires are anonymous and it is impossible to match the individual student with a specific evaluation of the teacher.[11] However, each questionnaire reports the name of the course and the class identifier, so that we can attach average evaluations to each class in each course. Figure 1 shows, as an example, the first page of the evaluation questionnaire used in the academic year 1998-1999.[12]

[INSERT FIGURE 1 ABOUT HERE]

2.1 The random allocation

In this section we present evidence that the random allocation of students into classes was successful. De Giorgi, Pellizzari, and Redaelli (2010) use data for the same cohort (although for a smaller set of courses and programs) and provide similar evidence. The randomization was (and still is) performed via a simple random algorithm that assigned a class identifier to each student, who were then instructed to attend the lectures for the specific course in the class labeled with the same identifier. The university administration also adopted the policy of repeating the randomization for each course with the explicit purpose of encouraging wide interactions among the students.

[INSERT TABLE 4 ABOUT HERE]

graduation year, given that Bocconi has improved substantially over time in its ability to track its graduates. Until the 1985 birth cohort, all Italian males were required to serve in the army for 10-12 months but were allowed to postpone the service if enrolled in full time education. For college students, it was customary to enroll right after graduation.

[11] We are not aware of any university in the world where the students evaluations of their teachers are not anonymized.

[12] The questionnaires were changed slightly over time as new items were added and questions were slightly rephrased. We focus on a subset of questions that are consistent over the period under consideration.

Table 4 reports results from a battery of OLS regressions for each course in each program (i.e. each of the cells in Table 1) with selected observable characteristics (by columns in the table) on the left hand side and a full set of class fixed effects on the right. For each of these regressions we run an F-test of the joint significance of such dummies, which corresponds to testing the hypothesis that the average of the observable variable under consideration is the same in all classes. In the table we report the average value of such F-tests for each observable characteristic in each degree program, together with the percentage of tests that reject the null. Results show that the average F-test is always smaller than 1 but in one case. Only in a few cases the fraction of tests that reject the null is above 10%.

[INSERT FIGURE 2 ABOUT HERE]

Testing the equality of means is not a sufficient test of randomization for continuous variables. Hence, in Figure 2 we compare the distributions of our measures of ability (high school grades and entry test scores) for the entire student body and for a randomly selected class in each program. The figure evidently shows that the distributions are extremely similar and formal tests confirm the visual impression.[13]

3 Estimating teacher effectiveness

We use performance data (mainly grades, but also graduation marks and wages) for our students to estimate measures of teacher effectiveness. Namely, for each of the compulsory courses listed in Table 1 we compare the future outcomes of students that attended those courses in different classes, under the assumption that students who were taught by better professors will enjoy better outcomes later on. This approach is similar to the *value-added* methodology that is more commonly used in primary and secondary schooling (Goldhaber and Hansen, 2010; Hanushek, 1979; Hanushek and Rivkin, 2006, 2010; Rivkin, Hanushek, and Kain, 2005; Rothstein, 2009) but it departs from its standard version, that uses contemporaneous outcomes and

[13] We perform Kolmogorov-Smirnov tests for the continuous variable and tests of proportions for dummy indicators.

conditions on past performance, since we use future performance to infer current teaching quality.[14]

One usual concern with the estimation of teacher quality is the non-random assignment of students to professors. For example, if the best students self-select themselves into the classes of the best teachers, then estimates of teacher quality would be biased upward. Rothstein (2009) shows that such a bias can be substantial even in well-specified models and especially when selection is mostly driven by unobservabes.

We avoid these complications by exploiting the random allocation of students in our cohort to different classes for each of their compulsory courses. For this same reason, we focus exclusively on compulsory courses, as self-selection is an obvious concern for elective subjects. Moreover, elective courses were usually taken by fewer students than compulsory courses, and hence they were usually taught in a single class.

To understand our procedure to estimate teacher effectiveness, consider a set of students enrolled in degree program d and indexed by $i = 1, \ldots, N_d$, where N_d is the total number of students in the program. In our application there are three degree programs ($d = \{1, 2, 3\}$): Management, Economics and Law&Management. Each student i attends a fixed sequence of compulsory courses indexed by $c = 1, \ldots, C_d$, where C_d is the total number of such compulsory courses in degree program d. In each course c the student is randomly allocated to a class $s = 1, \ldots, S_c$, where S_c is the total number of classes in course c. Denote by $\zeta \in Z_c$ a generic (compulsory) course, different from c, which students i attends in semester $t \geq t_c$, where t_c denotes the semester in which course c is taught.

Let $y_{ids\zeta}$ denote the grade obtained by student i in course ζ. To control for differences in the distribution of grades across courses, $y_{ids\zeta}$ is standardized at the course level. Then, for each course c in each program d we run the following regression:

$$y_{ids\zeta} = \alpha_{dcs} + \beta X_i + \epsilon_{ids\zeta} \tag{1}$$

where X_i is a vector of student-level characteristics including a gender dummy, a dummy for

[14]For this reason we prefer to use the label *teacher effectiveness* for our estimates. Carrell and West (2010) use our same approach and label it *value-added*.

whether the student is in the top income bracket, the entry test score and the high school leaving grade. The αs are our parameters of interest and measure the effectiveness of the teacher of class s: high values of α indicate that, on average, students attending course c in class s performed better (in subsequent courses) than students taking course c in a different class. The random allocation procedure guarantees that the class fixed effects α_{dcs} in equation 1 are purely exogenous and identification is straightforward.[15]

[INSERT TABLE 5 ABOUT HERE]

Table 5 presents the descriptive statistics of the class fixed effects estimated according to this procedure. Overall, we are able to estimate 226 such fixed effects, the large majority of which are for Management courses.[16]. On average, the standard deviation of the class fixed effects within courses is 0.08. Recall that grades are normalized so that the distributions of the class effects are comparable across courses. This is almost twice as large as what Carrell and West (2010) found for introductory course professors in their sample.

Our results imply that a one-standard deviation change in professor quality translates in a 0.08-standard-deviation change in student achievement, roughly 1 percent of actual grades over an average of approximately 26.[17] This average effect masks a large degree of heterogeneity across courses. The variation in the class effects ranges from essentially zero (the minimum being 0.0003 in the course of accounting for Law&Management students) to almost 0.25 (in macroeconomics for Economics students). Students who took a course in the class of the best and the worst teacher are, on average, 23% of a standard deviation apart in terms of future performance, corresponding to over 3% of the average grade.

[15]Notice that in few cases more than one teacher taught in the same class, so that our class effects capture the overall effectiveness of teaching and cannot be attached to a specific person. Since the students' evaluations are also available at the class level and not for specific teachers, we cannot disaggregate further.

[16]We cannot run equation 1 for courses that have no contemporaneous nor subsequent courses, such as Corporate Strategy for Management, Banking for Economics and Business Law for Law&Management (See Table 1). For such courses, the set Z_c is empty. Additionally, some courses in Economics and in Law&Management are taught in one single class, for example econometrics for Economics students or statistics for Law&Management. For such courses, we have $S_c = 1$.

[17]In Italy, university exams are graded on a scale 0 to 30, with pass equal to 18. Such a peculiar grading scale comes from historical legacy: while in primary, middle and high school students were graded by one teacher per subject on a scale 0 to 10 (pass equal to 6), at university each exam was supposed to be evaluated by a commission of three professors, each grading on the same 0-10 scale, the final mark being the sum of these three. Hence, 18 is pass and 30 is full marks. Apart from the scaling, the actual grading at Bocconi is performed as in the average US or UK university.

In Table 5 we also report measures of the statistical significance of the estimated class effects. First, for each course we run an F-test for the hypothesis that all the effects are equal and the table reports the average value of these F-tests. Moreover, we also report the percentage of the tests that reject the null at the 95% level.

For robustness and comparison, we estimate the class effects in a number of other ways, using alternative outcome measures. First, rather than using performance in subsequent courses, we run equation 1 with the grade in the same course c as a dependent variable. We label these estimates *contemporaneous* effects.

Next, we restrict the set Z_c to courses belonging to the same subject area of course c, under the assumption that good teaching in one course is likely to have a stronger effect on learning in courses of the same subject areas (e.g. a good basic mathematics teacher is more effective in improving students performance in financial mathematics than in business law). The subject areas are defined by the colors in Table 1. We label these estimates *subject* effects. Given the more restrictive definition of Z_c we can only produce these estimates for a smaller set of courses.

Finally, we replace the dependent variable in equation 1 with either graduation mark or the students' wage in their first job. We label these estimates *graduation* and *wage* effects, respectively.

[INSERT TABLE 6 ABOUT HERE]

In Table 6 we investigate the correlation between these alternative estimates of the class effects. Specifically, we report results from a series of simple OLS regressions with our benchmark estimates as dependent variable and, in turn, each of the alternative measures on the right hand side, together with dummies for degree program, term and subject area.

All the alternative versions of class effects are positively correlated with our benchmark, apart from the contemporaneous that are negatively correlated.

The subject and graduation estimates are positively and significantly correlated with our benchmark. Given the larger noise in the wage data, the coefficient on the wage effects is also positive but does not reach conventional levels of statistical significance. Only the contemporaneous effects are negatively and significantly correlated with our benchmark, a result that

is consistent with previous findings (Carrell and West, 2010; Krautmann and Sander, 1999; Weinberg, Fleisher, and Hashimoto, 2009). We will return to this issue in Section 4.

4 Correlating teacher effectiveness and student evaluations

In this section we investigate the relationship between our measures of teaching effectiveness from Section 3 and the evaluations received by the same teachers from their students. We concentrate on a specific set of items from the evaluation questionnaires, namely overall teaching quality, lecturing ability of the teacher, overall lecturing clarity. Additionally, we also look at other items: the teacher's ability in arousing interest for the subject, the course logistics (schedule of classes, combinations of practical sessions and traditional lectures) and the total workload compared to other subjects.

By definition, the class fixed effects capture all those features, both observable and unobservable, that are fixed for all students in the class. These certainly include teaching quality but also other factors that were documented to be important ingredients of the education production function, such as class size and class composition (De Giorgi, Pellizzari, and Woolston, 2010). An important advantage of our data is that most of these other factors are observable. In particular, based on our academic records we can construct measures of both class size as well as class composition. Additionally, we also have access to the identifiers of the teachers in each class and we can recover a large set of observable characteristics, like gender, tenure status and, possibly, research records. We also know which of the several teachers in each course acted as coordinator.

Once we condition on all these observable controls, unobservable teaching quality is likely to be the only remaining factor. At a minimum, it should be uncontroversial that teaching quality is by far the single most important unobservable that generates variation in the estimated class effects.

Formally, we estimate the following equation:

$$q^k_{dtcs} = \lambda_0 + \lambda_1 \widehat{\alpha}_{dtcs} + \lambda_2 C_{tdcs} + \lambda_3 T_{tdcs} + \gamma_d + \delta_t + \upsilon_c + \epsilon_{dtcs} \tag{2}$$

where q^k_{dtcs} is the average answer to question k in class s of course c in the degree program d (which is taught in term t), $\widehat{\alpha}_{dtcs}$ is the estimated class fixed effect, C_{tdcs} is a set of class characteristics (class size, class attendance, number of collected questionnaires), T_{tdcs} is a set of teacher characteristics (gender, course coordinator). γ_d, δ_t and υ_c are fixed effects for degree program, term and subject areas, respectively. ϵ_{dtcs} is a residual error term.

Since the dependent variable in equation 2 is an average, we estimate it using weighted OLS, where each observation is weighted by the square root of the number of collected questionnaires in the class, which corresponds to the size of the sample over which the average answers are taken. Additionally, we also bootstrap the standard errors with 150 replications to take into account the presence of generated regressors (the $\widehat{\alpha}$s).

[INSERT TABLE 7 ABOUT HERE]

Table 7 reports the estimates of equation 2 for a first set of core evaluation items, namely overall teaching quality and lecturing clarity. For each of these items we show results obtained using our benchmark estimates of teacher effectiveness and those obtained using the contemporaneous class effects.

Results show that our benchmark class effects are negatively associated with both core items in the students evaluations. In other words, teachers who are more effective in promoting future performance receive worst evaluations from their students. This relationship is statistically significant for both items and of sizable magnitude. One standard deviation increase in teacher effectiveness reduces the students evaluations on both items by almost 5% of a standard deviation. Such an effect could move a teacher who would otherwise receive a median evaluation down to the 46th-47th percentile of the distribution of either of the items under consideration.

Consistently with the findings of other studies (Carrell and West, 2010; Krautmann and Sander, 1999; Weinberg, Fleisher, and Hashimoto, 2009), when we use the contemporaneous class effects the estimated coefficients turn positive and highly statistically significant for both items. In other words, teachers of classes that are associated with higher grades in their own exam receive better evaluations from the students. The magnitude of the effects is only marginally smaller than those estimated for our benchmark measures (4% of a standard deviation rather than 5%).

The results in Table 7 clearly challenge the validity of students' evaluations of teachers as a measure of teaching quality. Even abstracting from the possibility that professors strategically adjust their grades to please the students (a practice that is made difficult by the timing of the evaluations, that are always collected before the exam takes place), it might still be possible that professors who make the classroom experience more enjoyable do that at the expense of true learning or fail to encourage students to exert effort. Alternatively, students might reward teachers who prepare them for the exam, that is teachers who teach-to-the-test, even if this is done at the expenses of true learning. This interpretation is consistent with the results in Weinberg, Fleisher, and Hashimoto (2009), who provide evidence that students are generally unaware of the value of the material they have learned in a course, and it is the interpretation that we adopt to develop the theoretical framework of Section 6.

Of course, one may also argue that student satisfaction is important per se and, even, that universities should aim at maximizing satisfaction rather than learning, especially private institutions like Bocconi. We doubt that this is the most common understanding of higher education policy.

[INSERT TABLE 8 ABOUT HERE]

In Table 8 we replicate the same regressions of Table 7 with other items in the evaluation questionnaires, namely the teacher ability in arousing interest for the subject, the quality of the course logistics (combination and integration of practical classes and lectures, class schedule, etc.) and total workload.

Our main results are confirmed also for these additional items. In particular, we find that the teacher ability to arouse interest is negatively correlated with benchmark teacher effectiveness and positively correlated with contemporaneous class effects. Good evaluations of course logistics are positively affected by high contemporaneous grades while they are largely unaffected by teacher effectiveness. Effective teachers, who are presumably associated with higher workload, receive lower evaluations on this item, which is largely unaffected by contemporaneous class effects.

5 Robustness checks

In this section we present robustness checks for our main results in Sections 3 and 4.

First, we investigate the role of students' dropout in the estimation of our measures of teacher effectiveness. In our main empirical analysis students who do not have a complete academic record are excluded. These are students who either dropped out of higher education or have transfered to another university or are still working towards the completion of their programs, whose formal duration was 4 years. They total about 10% of all the students who enrolled in their first year in 1998-1999. In order to check that excluding them does not affect our main results, in Figure 3 we compare the benchmark class effects estimated in Section 3 with similar estimates that include such dropout students. As it is evident, the two sets of estimates are very similar and regressing one over the other (controlling for degree program fixed effects) yields an R^2 of over 90%. Importantly, there does not seem to be larger discrepancies between the two versions of the class effects for the best or the worst teachers.

[INSERT FIGURE 3 ABOUT HERE]

Second, one might be worried that students might not comply with the random assignment to the classes. Students might decide to attend one or more courses in a different class from the one to which they were formally allocated for various reasons. For example, they may desire to stay with their friends who are assigned a different class or they may like a specific teacher who is known to present the subject particularly clearly. Unfortunately, such changes would not be recorded in our data, unless the student formally asked to be allocated to a different class, a request that needed to be adequately motivated.[18] Hence, we cannot exclude a priori that some students switch class.

If the process of class switching is unrelated to teaching quality, then it merely affects the precision of our estimated class effects but it is very well possible that students switch in search for the best lectures. We can get some indication of the extent of this problem from the students' answers to an item of the evaluation questionnaires that asks about the congestion

[18]Such requests could be motivated, for example, with health reasons. For example, due to a broken leg a student might not be able to reach classrooms in the upper floors of the university buildings and could ask to be assigned a class that is taught on the ground floor.

in the classroom. Specifically, the question asks whether the number of students in the class is detrimental to one's learning. We can, thus, identify the most congested classes from the average answer to such question in each class.

Courses in which students concentrate in the class of the best professors should be characterized by a very skewed distribution of such a measure of congestion, with one (or a few) classes being very congested and the others being pretty empty. Thus, for each course we compute the difference in the congestion indicator between the most and the least congested classes (over the standard deviation). Courses in which such difference is very large should be the ones that are more affected by switching behaviors.

[INSERT TABLE 9 ABOUT HERE]

In Table 9 we replicate our benchmark class effects (Panel A) by excluding the most switched course (Panel B), i.e. the course with the largest difference between the most and the least congested classes, which is marketing. Results change only marginally, although the significance levels are reduced accordingly with the smaller sample sizes. Next, in Panel C and D we exclude from the estimates also the second most switched course (human resource management) and the five most switched courses, respectively.[19] Overall, this exercise suggests that course switching should not affect our estimates in any major direction.

6 Interpreting the results: a simple theoretical framework

We think of teaching as the combination of two types of activities: *real teaching* and *teaching to the test*. The first consists of presentations and discussions of the course material and leads to actual learning, conditional on the students exerting effort; the latter is aimed at maximizing performance in the exam, it requires lower effort by the students and it is not necessarily related to actual learning.

Practically, we think of real teaching as competent presentations of the course material with the aim of making students understand and master it and teaching to the test as mere repetition

[19]The five most congested courses are marketing, human resource management, mathematics for Economics and Management, financial mathematics and managerial accounting.

of exam tests and exercises with the aim of making students learn how to solve them, even without fully understanding their meaning.

Consider a setting in which teachers are heterogenous in their preference (or ability) to do real teaching. We measure such heterogeneity with a parameter $\mu_j \in [0, 1]$, such that a teacher j with $\mu_j = 0$ exclusively teaches to the test and a teacher with $\mu_j = 1$ exclusively engages in real teaching.

Then, the grade x of a generic student i in the course taught by teacher (or in class) j is defined by the following production function:

$$x_i = \mu_j h(e_i) + (1 - \mu_j)\overline{x} \tag{3}$$

which is a linear combination of a function $h(\cdot)$ of student's effort e_i and a constant $\overline{x}$, weigthed by the teacher's type μ_j. We assume $h(\cdot)$ to be a continuous and twice differentiable concave function. Under full real teaching ($\mu_j = 1$) grades vary with students' effort; on the other hand, if the teacher exclusively teaches to the test ($\mu_j = 0$), everyone gets the same grade $\overline{x}$, regardless of effort.

The parameter $\overline{x}$ measures the extent to which the exam material and the exam format lend themselves to teaching to the test. To the one extreme, one can think of the exam as a selection of multiple-choice questions randomly drawn from a large pool. In such a situation, teaching to the test merely consists in going over all the possible questions and memorizing the correct answer. This is a setting which would feature a large $\overline{x}$. The other extreme are essays, where there is no obvious correct answers and one needs to personally and originally elaborate on one's own understanding of the course's material.

For simplicity, equation 3 assumes that teaching to the test does not require students to exert effort. All our results would be qualitatively unchanged under the weaker assumption that teaching to the test requires less effort by the students. We also assume that μ_j is a fixed characteristic of teacher j, so that the model effectively describes the conditions for identifying teachers of different types, a key piece of information for hiring and promotion decisions. Alternatively, μ_j could be treated as an endogenous variable under the control of the individual teacher, in which case the model would feature a rather standard agency problem where the

university tries to provide incentives to the teachers to choose a μ_j close to 1. Although, such a model would be considerably more complicated that what we present in this section, its qualitative results would be unchanged and the limited information on teachers in our data would makes its additional empirical content redundant in our setting.

In all cases, we assume that μ_j is unobservable by the university administrators (the principal) and, although it might be observable to the students, it cannot be credibly communicated to third parties.

Assume now that students care about their grades but dislike exerting effort, so that the utility function of a generic student i can be written as follows:

$$U_i = x_i - \frac{1}{2}\frac{e_i^2}{\eta_i} \tag{4}$$

where η_i is a measure of student ability. The quasi-linearity of equation 4 simplifies the algebra of the model. Alternatively, we could have introduced some curvature in the utility function and assumed a linear production process without affecting the results. With non-linearities both in the production and the utility function one would have to make explicit a number of additional assumptions to guarantee existence and uniqueness of the equilibrium.

Students choose their optimal level of effort e_i^* according to the following first order conditions:

$$\mu_j \frac{\partial h(e)}{\partial e_i}(e_i^*) = \frac{e_i^*}{\eta_i} \tag{5}$$

Using equation 5 it is easy to derive the following results:

$$\frac{de_i^*}{d\eta_i} > 0 \tag{6}$$

$$\frac{de_i^*}{d\mu_j} > 0 \tag{7}$$

$$\frac{de_i^*}{d\mu_j d\eta_i} > 0 \tag{8}$$

Result 6 shows that more able students exert higher effort. Result 7 shows that more real teaching induces higher effort from the students and result 8 indicates that such an effect is larger for the more able students.

Additionally, using the envelope theorem it is easy to show that:

$$\frac{\partial U_i(e_i^*)}{\partial \mu_j} = h(e_i^*) - \overline{x} \tag{9}$$

Define $\overline{e}$ the level of effort such that $h(\overline{e}) = \overline{x}$. Moreover, since for given μ_j there is a unique correspondence between effort and ability, $\overline{e}$ uniquely identifies a $\overline{\eta}$. Hence:

$$\frac{\partial U_i(e_i^*)}{\partial \mu_j} > 0 \quad \text{if} \quad \eta_i > \overline{\eta} \tag{10}$$

$$\frac{\partial U_i(e_i^*)}{\partial \mu_j} < 0 \quad \text{if} \quad \eta_i < \overline{\eta} \tag{11}$$

Equations 10 and 11 are particularly important under the assumption that, when answering questions about the overall quality of a course, students give a better evaluation to classes or teachers that are associated to a higher level of utility. Equations 10 and 11 suggest that high ability students evaluate better teachers or classes that are more focused on real learning while low ability students prefer teachers or classes that teach to the test. Hence, if the (non-contemporanous) class effects estimated in Section 3 indeed measure the real learning value of a class (μ_j, in the terminology of our model), we expect to see a more positive (or less negative) correlation between such class effects and the students' evaluations in those classes where the concentration of high ability is higher.

7 Further evidence

In this section we present two additional pieces of evidence that are consistent with the implications of model of Section 6.

First, our specification of the production function for exam grades in equation 3 implies a positive relationship between grade dispersion and the professor's propensity to real teaching (μ_j). In our empirical exercise the class fixed effects can be interpreted as measures of teacher effectiveness or, in the terminology of the model, μ_j. Hence, if grades were more dispersed in the classes of the worst teachers one would have to question our specification of equation 3.

[INSERT FIGURE 4 ABOUT HERE]

In Figure 4 we plot the coefficient of variation of grades in each class (on the vertical axis) against our measure of teacher effectiveness for the same class (on the horizontal axis).[20] The two variables are obviously positively correlated and such correlation is statisticall significant at conventional levels: a simple univariate OLS regression of the variable on the vertical axis on the variable on the horizontal axis yields a coefficient of 0.017 with a standard error of 0.006. The evidence from Figure 4 is consistent with the specification of equation 3.

[INSERT TABLE 10 ABOUT HERE]

Next, according to equations 10 and 11, we expect the correlation between our measures of teacher effectiveness and the average student evaluations to be less negative in classes where the share of high ability students is higher. This is the hypothesis that we investigate in Table 10. We define high ability students those who score in the upper quartile of the distribution of the entry test score and for each class in our data we compute the share of such students. Then, we augment the standard specification of equation 2 with a set of dummies for the quartiles of the distribution of the share of high ability students and we interact these dummies with our measure of teacher effectiveness. The results show that, indeed, the negative correlations reported in Table 7 are mostly due to classes with a particularly low incidence of high ability students, i.e. classes in the bottom quartile. Although the interaction terms are not significant, the point estimates suggest that in classes with higher shares of high ability students teacher effectiveness is less negatively correlated with the students' evaluations.

8 Conclusions

Using administrative archives from Bocconi University and exploiting experimental variation in students allocation to teachers within courses we find that, on average, students evaluate positively classes that give high grades and negatively classes that are associated with high grades in subsequent courses, higher graduation marks and higher wages when entering the labor market. The empirical findings can be rationalized with a simple model featuring heterogeneity in

[20] To take proper account of differences across degree programs, the variables on both axes of Figure 4 are the residuals of weighted OLS regressions that condition on degree program fixed effects, as in standard partitioned regressions.

the preferences (or ability) of teachers to engage in real teaching rather than teach to the test, with the former requiring higher effort from students than the latter. Overall, our results casts serious doubts on the validity of students' evaluations of professors as measures of teaching quality or effort.

References

BAKER, G., R. GIBBONS, AND K. J. MURPHY (1994): "Subjective performance measures in optimal incentive contracts," *Quarterly Journal of Economics*, 109(4), 1125–1156.

BECKER, W. E., AND M. WATTS (1999): "How departments of economics should evaluate teaching," *American Economic Review (Papers and Proceedings)*, 89(2), 344–349.

BROWN, B. W., AND D. H. SAKS (1987): "The microeconomics of the allocation of teachers' time and student learning," *Economics of Education Review*, 6(4), 319–332.

CARRELL, S. E., AND J. E. WEST (2010): "Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors," *Journal of Political Economy*, 118(3), 409–32.

CASHIN, W. E. (1990): "Students do rate different academic fields differently," in *Student ratings of instruction; Issues for improving practice*, ed. by M. Theall, and J. Franklin, vol. 43 of *New Directions for Teaching and Learning*, pp. 113–121. Jossey-Bass.

DE GIORGI, G., M. PELLIZZARI, AND S. REDAELLI (2010): "Identification of Social Interactions through Partially Overlapping Peer Groups," *American Economic Journal: Applied Economics*, 2(2), 241–275.

DE GIORGI, G., M. PELLIZZARI, AND W. G. WOOLSTON (2010): "Class Size and Class Heterogeneity," *Journal of the European Economic Association*, forthcoming.

DUFLO, E., R. HANNA, AND M. KREMER (2010): "Incentives Work: Getting Teachers to Come to School," mimeo, MIT.

FIGLIO, D. N., AND L. KENNY (2007): "Individual teacher incentives and student performance," *Journal of Public Economics*, 91, 901–914.

GOLDHABER, D., AND M. HANSEN (2010): "Using performance on the job to inform teacher tenure decisions," *American Economic Review (Papers and Proceedings)*, 100(2), 250–255.

HANUSHEK, E. A. (1979): "Conceptual and empirical issues in the estimation of educational production functions," *Journal of Human Resources*, 14, 351–388.

HANUSHEK, E. A., AND S. G. RIVKIN (2006): "Teacher quality," in *Handbook of the Economics of Education*, ed. by E. A. Hanushek, and F. Welch, vol. 1, pp. 1050–1078. North Holland, Amsterdam.

——— (2010): "Generalizations about using value-added measures of teacher quality," *American Economic Review (Papers and Proceedings)*, 100(2), 267–271.

HOGAN, T. D. (1981): "Faculty research activity and the quality of graduate training," *Journal of Human Resources*, 16(3), 400–415.

HOLMSTROM, B., AND P. MILGROM (1994): "The firm as an incentive system," *American Economic Review*, 84(4), 972–991.

JACOB, B. A., AND L. LEFGREN (2008): "Can principals identify effective teachers? Evidence on subjective performance evaluation in education," *Journal of Labor Economics*, 26, 101–136.

KANE, T. J., AND D. O. STAIGER (2008): "Estimating teacher impacts on student achievement: an experimental evaluation," Discussion Paper 14607, NBER Working Paper Series.

KRAUTMANN, A. C., AND W. SANDER (1999): "Grades and student evaluations of teachers," *Economics of Education Review*, 18, 59–63.

KRUEGER, A. B. (1999): "Experimental estimates of education production functions," *Quarterly Journal of Economics*, 114, 497–532.

OECD (2008): *Education at a Glance*. Organization of Economic Cooperation and Development, Paris.

PRENDERGAST, C., AND R. H. TOPEL (1996): "Favoritism in organizations," *Journal of Political Economy*, 104(5), 958–978.

RIVKIN, S. G., E. A. HANUSHEK, AND J. F. KAIN (2005): "Teachers, Schools and Acedemic Achievement," *Econometrica*, 73(2), 417–458.

ROCKOFF, J. E. (2004): "The impact of individual teachers on student achievement: evidence from panel data," *American Economic Review (Papers and Proceedings)*, 94(2), 247–252.

ROCKOFF, J. E., AND C. SPERONI (2010): "Subjective and Objective Evaluations of Teacher Effectiveness," *American Economic Review (Papers and Proceedings)*, 100(2), 261–266.

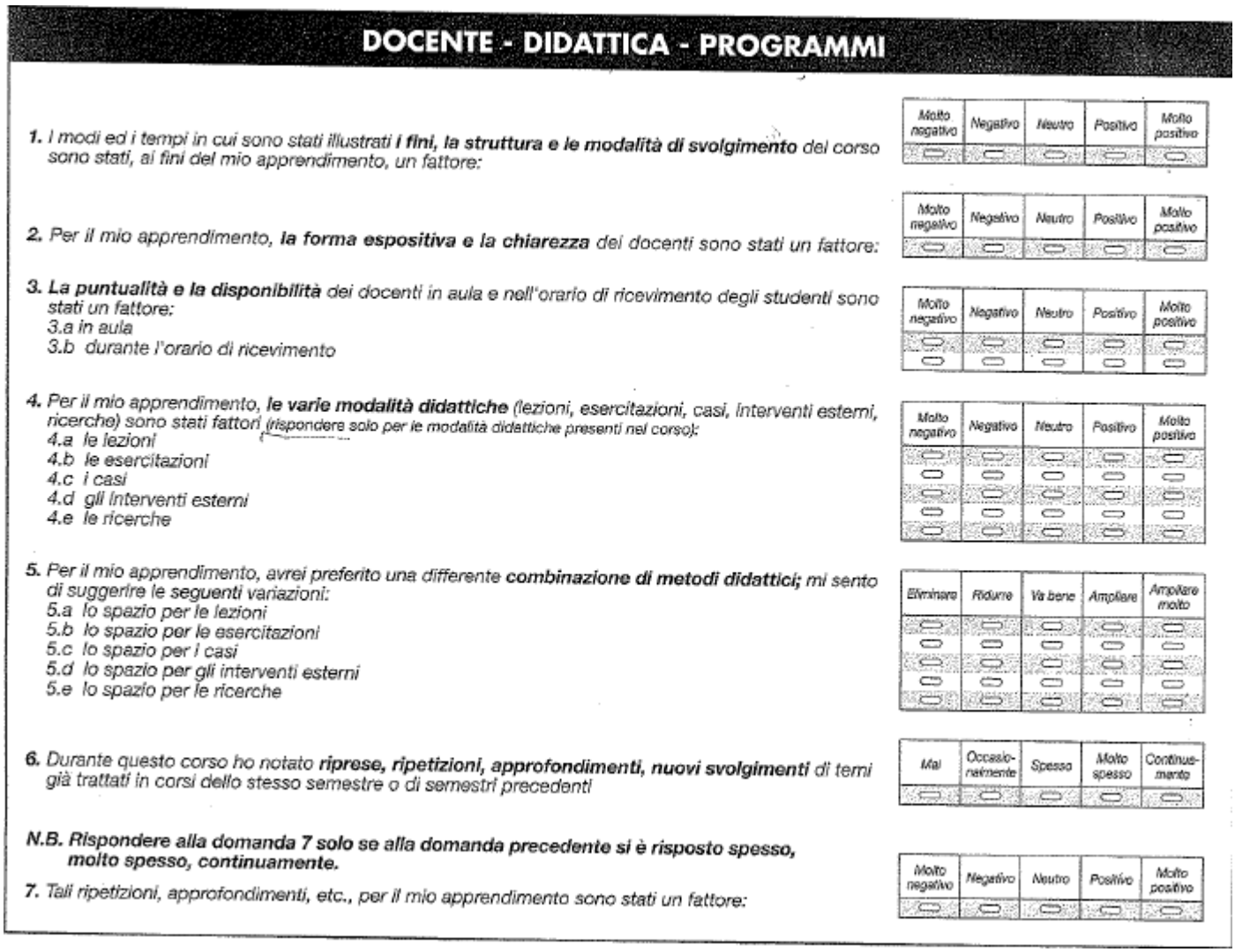
ROTHSTEIN, J. (2009): "Student sorting and bias in value added estimation: selection on observables and unobservables," *Education Finance and Policy*, 4(4), 537–571.

——— (2010): "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement," *Quarterly Journal of Economics*, 125(1), 175–214.

TYLER, J. H., E. S. TAYLOR, T. J. KANE, AND A. L. WOOTEN (2010): "Using student performance data to identify effective classroom practices," *American Economic Review (Papers and Proceedings)*, 100(2), 256–260.

WEINBERG, B. A., B. M. FLEISHER, AND M. HASHIMOTO (2009): "Evaluating Teaching in Higher Education," *Journal of Economic Education*, 40(3), 227–261.

Figures



DOCENTE - DIDATTICA - PROGRAMMI

	Molto negativo	Negativo	Neutro	Positivo	Molto positivo
1. *I modi ed i tempi in cui sono stati illustrati* ***i fini, la struttura e le modalità di svolgimento*** *del corso sono stati, ai fini del mio apprendimento, un fattore:*	○	○	○	○	○
2. *Per il mio apprendimento,* ***la forma espositiva e la chiarezza*** *dei docenti sono stati un fattore:*	○	○	○	○	○
3. ***La puntualità e la disponibilità*** *dei docenti in aula e nell'orario di ricevimento degli studenti sono stati un fattore:*					
3.a in aula	○	○	○	○	○
3.b durante l'orario di ricevimento	○	○	○	○	○
4. *Per il mio apprendimento,* ***le varie modalità didattiche*** *(lezioni, esercitazioni, casi, interventi esterni, ricerche) sono stati fattori (rispondere solo per le modalità didattiche presenti nel corso):*					
4.a le lezioni	○	○	○	○	○
4.b le esercitazioni	○	○	○	○	○
4.c i casi	○	○	○	○	○
4.d gli interventi esterni	○	○	○	○	○
4.e le ricerche	○	○	○	○	○

	Eliminare	Ridurre	Va bene	Ampliare	Ampliare molto
5. *Per il mio apprendimento, avrei preferito una differente* ***combinazione di metodi didattici;*** *mi sento di suggerire le seguenti variazioni:*					
5.a lo spazio per le lezioni	○	○	○	○	○
5.b lo spazio per le esercitazioni	○	○	○	○	○
5.c lo spazio per i casi	○	○	○	○	○
5.d lo spazio per gli interventi esterni	○	○	○	○	○
5.e lo spazio per le ricerche	○	○	○	○	○

	Mai	Occasio-nalmente	Spesso	Molto spesso	Continua-mente
6. *Durante questo corso ho notato* ***riprese, ripetizioni, approfondimenti, nuovi svolgimenti*** *di temi già trattati in corsi dello stesso semestre o di semestri precedenti*	○	○	○	○	○

N.B. Rispondere alla domanda 7 solo se alla domanda precedente si è risposto spesso, molto spesso, continuamente.

	Molto negativo	Negativo	Neutro	Positivo	Molto positivo
7. *Tali ripetizioni, approfondimenti, etc., per il mio apprendimento sono stati un fattore:*	○	○	○	○	○

Figure 1: Excerpt of student questionnaire

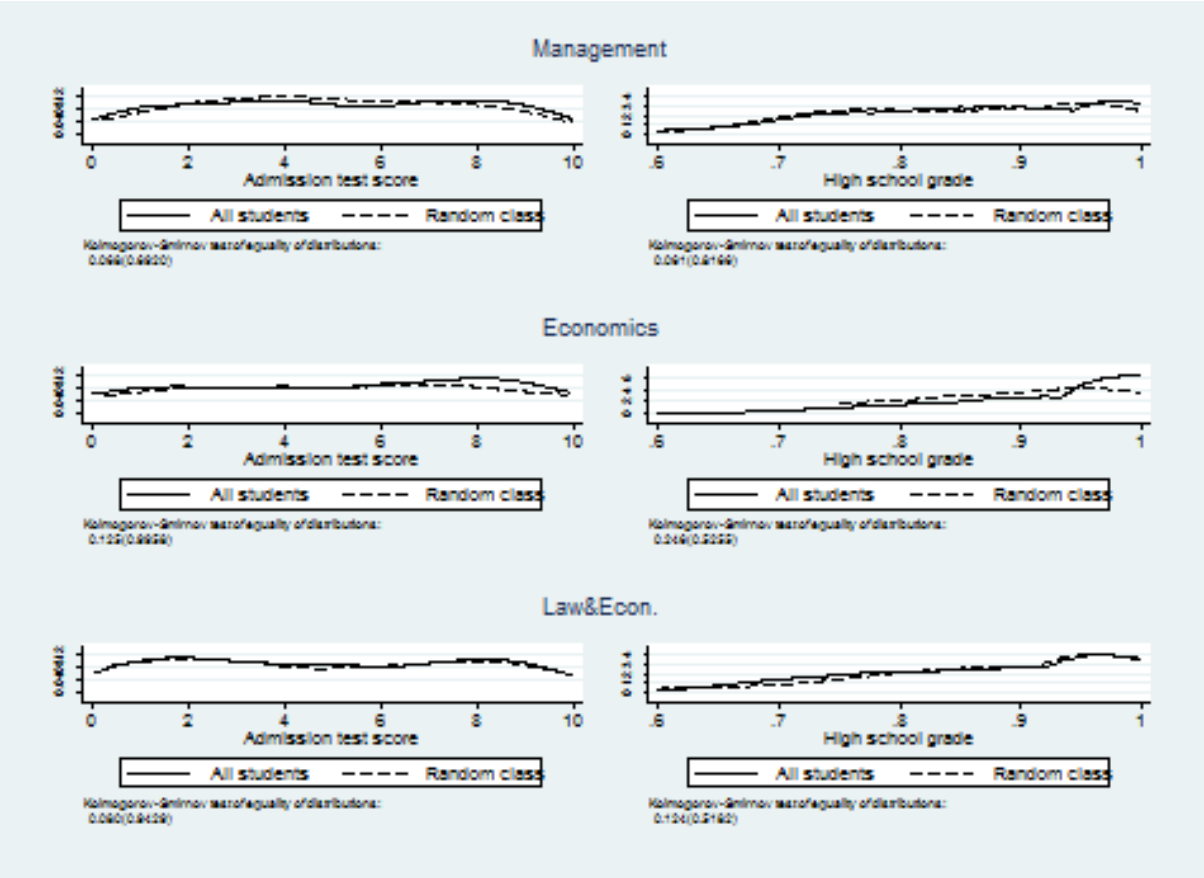


Figure 2: Evidence of random allocation - Ability variables

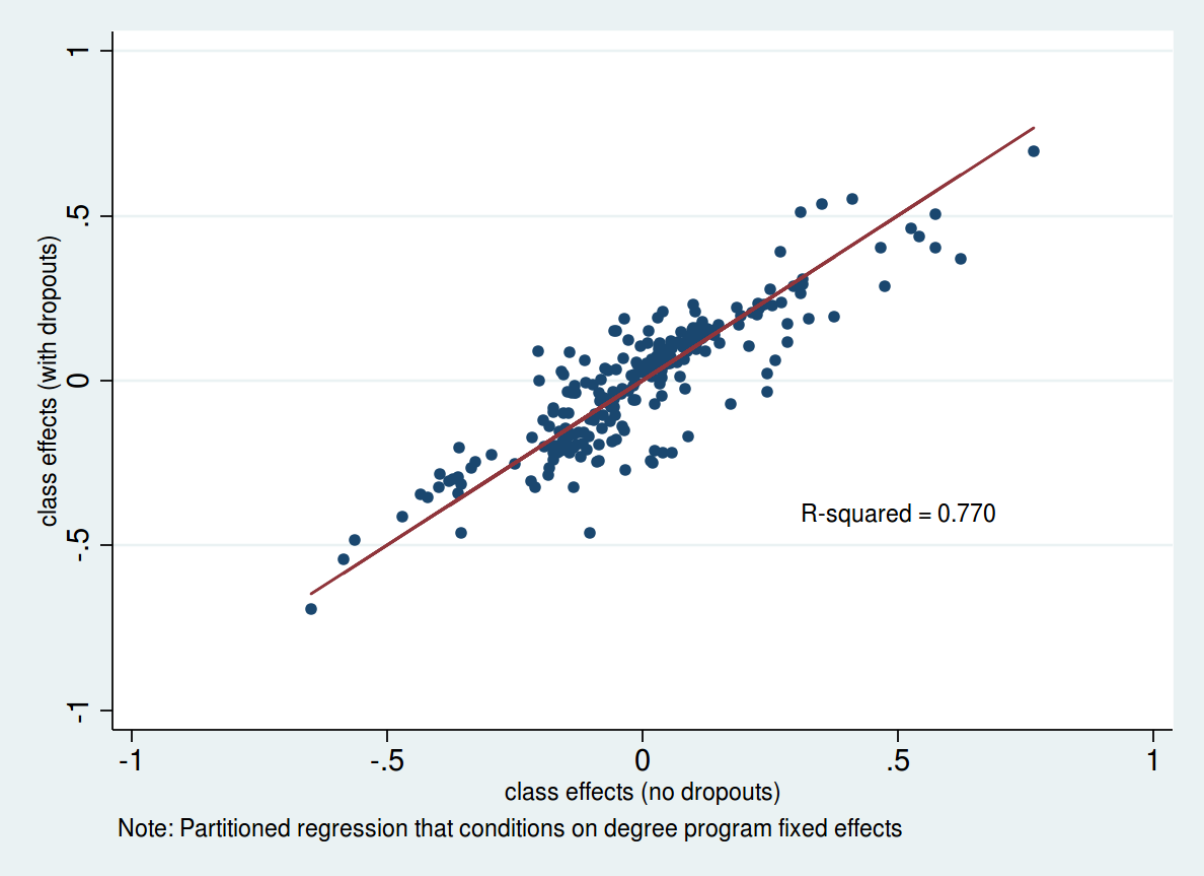


Figure 3: Robustness check for dropouts

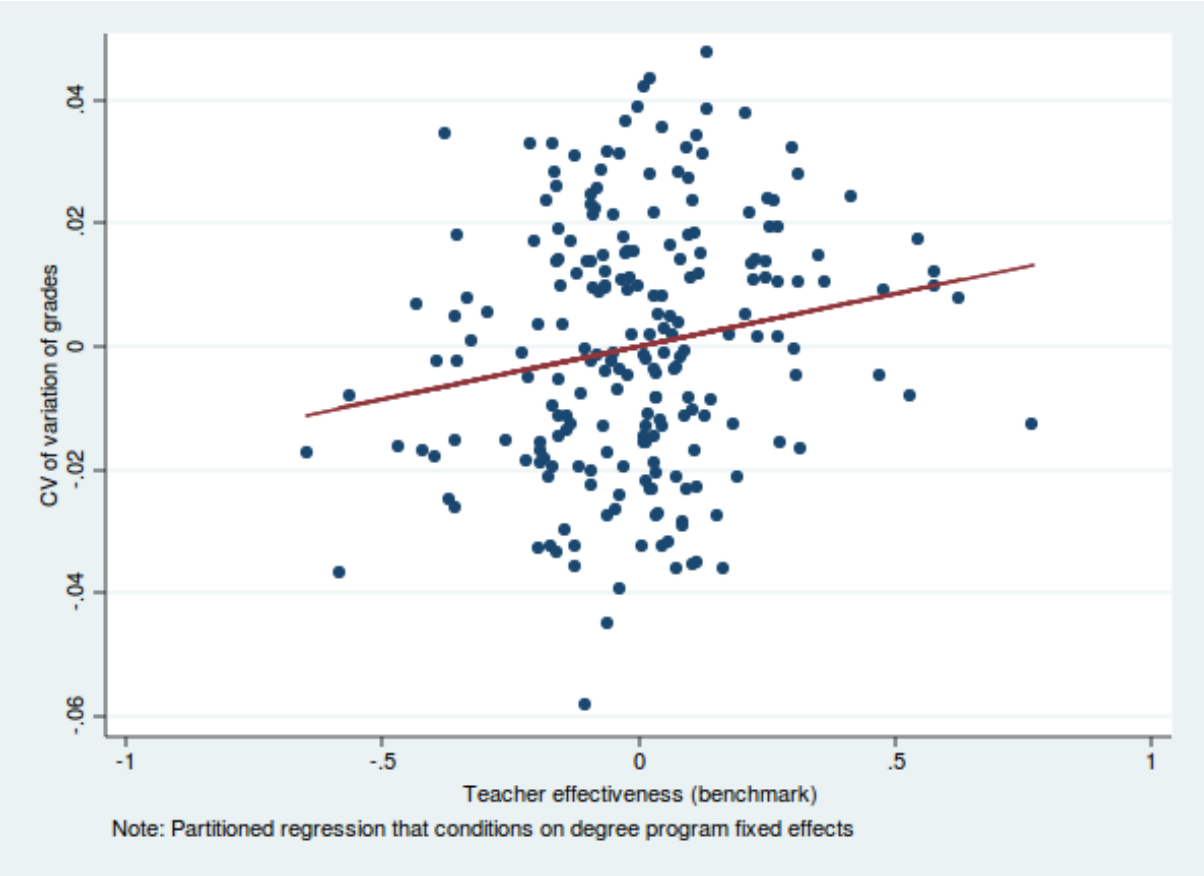


Figure 4: Teacher effectiveness and grade dispersion

Tables

Table 1: Structure of degree programs

	MANAGEMENT	ECONOMICS	LAW&MANAG.
Term I	Management I Private law Mathematics	Management I Private law Mathematics	Management I Mathematics
Term II	Microeconomics Public law Accounting	Microeconomics Public law Accounting	Accounting
Term III	Management II Macroeconomics Statistics	Management II Macroeconomics Statistics	Management II Statistics
Term IV	Business law Manag. of Public Administrations Financial mathematics Human resources management	Financial mathematics Public economics Business law	Accounting II Fiscal law Financial mathematics
Term V	Banking Corporate finance Management of industrial firms	Econometrics Economic policy	Corporate finance
Term VI	Marketing Management III Economic policy Managerial accounting	Banking	
Term VII	Corporate strategy		
Term VIII			Business law II

The colors indicate the subject area the courses belong to: red=management, black=economics, green=quantitative, blue=law. Only compulsory courses are displayed.

Table 2: Descriptive statistics of degree programs

	Term							
Variable	I	II	III	IV	V	VI	VII	VIII
				Management				
No. Courses	3	3	3	4	3	4	1	-
No. Classes	24	21	23	26	23	27	12	-
Avg. Class Size	129.00	147.42	134.61	138.6	117.5	133.5	75.1	-
SD Class Size	73.13	80.57	57.46	100.06	16.64	46.20	11.89	-
				Economics				
No. Courses	3	3	3	3	2	1	-	-
No. Classes	24	21	23	4	2	2	-	-
Avg. Class Size	129.00	147.42	134.61	98.3	131.0	65.5	-	-
SD Class Size	73.13	80.57	57.46	37.81	0	37.81	-	-
				Law & Management				
No. Courses	3	4	4	4	2	-	-	1
No. Classes	5	5	5	6	3	-	-	1
Avg. Class Size	104.4	139.2	139.2	116	116	-	-	174
SD Class Size	39.11	47.65	47.67	44.96	50.47	-	-	0

Table 3: Descriptive statistics of students

Variable	Management	Economics	Law &Management	Total
1=Female	0.408	0.427	0.523	0.427
1=Outside Milan	0.620	0.748	0.621	0.634
1=Top Income Bracket	0.239	0.153	0.368	0.248
High-school grade (0-100)	86.152	93.053	88.084	87.181
	(10.905)	(8.878)	(10.852)	(10.904)
Entry Test Score (0-100)	50.615	52.415	48.772	50.544
	(28.530)	(31.752)	(29.902)	(29.084)
University Grades (0-30)	25.684	26.945	25.618	25.785
	(3.382)	(2.978)	(3.473)	(3.380)
Graduation Mark (0-110)	101.401	107.321	101.603	102.073
	(7.716)	(5.213)	(7.653)	(7.692)
Wage (Euro)[a]	966.191	1,012.241	958.381	967.964
	(260.145)	(265.089)	(198.437)	(250.367)
Numb. of students	901	131	174	1,206

[a] Based on 391 observations for Management, 36 observations for Economics, 94 observations for Law&Management, i.e. 521 observations overall.

Table 4: Randomness checks

	Female	Academic High School	High School Grade	Entry Test Score	Top Income Bracket	Outside Milan	Late Enrollers
			PANEL A: by degree program[a]				
Management							
Avg. F-stat	0.142	0.143	0.16	0.106	0.134	0.312	0.293
% pval<0.05	4.76	4.76	4.76	0	14.28	28.57	14.28
Economics							
Avg. F-stat	0.249	0.19	0.201	0.15	0.16	0.117	0.129
% pval<0.05	9.09	9.09	9.09	0	0	0	0
Law & Management							
Avg. F-stat	1.12	0.437	0.284	0.227	0.27	0.282	0.445
% pval<0.05	28.57	0	0	0	0	0	0
			PANEL B: total[b]				
Avg. F-stat	0.348	0.209	0.193	0.14	0.165	0.252	0.275
% pval<0.05	10.26	5.13	5.13	0	7.69	15.38	7.69

The tests are obtained from regressing, for each course, observable student's characteristics on class dummies. The null hypothesis is that all estimated class dummies are equal to each other.

[a] Management: 21 courses, 156 classes; Economics: 11 courses, 72 classes; Law & Management: 7 courses, 14 classes.

[b] Total: 39 courses, 242 classes.

Table 5: Descriptive statistics of estimated class effects

	Management	Economics	Law & Management	Total
Std. dev. of estimated class effects (across classes within courses):				
mean	0.055	0.156	0.034	0.080
minimum	0.028	0.020	0.006	0.006
maximum	0.085	0.232	0.087	0.242
Largest minus smallest class effect (within courses):				
mean	0.157	0.427	0.048	0.215
minimum	0.040	0.029	0.009	0.009
maximum	0.246	0.757	0.123	0.757
Statistical significance:[a]				
Avg. F	4.422	4.365	1.263	4.212
% pval<0.05	22.9	34.7	0.00	25.2
No. of courses	20	11	7	38
No. of classes	144	72	14	230

[a] Test that each estimated class effect is different from the average estimated class effect in the course.

Table 6: Alternative estimates of class effects

Dependent variable: Benchmark class effects				
Subject	0.102***			
	(0.017)			
Graduation		0.115***		
		(0.034)		
Wages			0.007	
			(0.014)	
Contemp.				-0.051***
				(0.011)
Program fixed effects	yes	yes	yes	yes
Term fixed effects	yes	yes	yes	yes
Subject fixed effects	yes	yes	yes	yes
Observations	212	230	223	230

Weighted OLS by the inverse of the standard error of the estimated benchmark class effects. Bootstrapped standard errors in parentheses. * p<0.1, ** p<0.05,***p<0.01

Table 7: Teacher effectiveness and students' evaluations - Core items

	Overall teaching quality		Lecturing clarity	
	[1]	[2]	[3]	[4]
Teacher effectiveness				
Benchmark	-0.411*	-	-0.251*	-
	(0.218)		(0.135)	
Contemporaneous	-	0.195***	-	0.144***
		(0.04)		(0.030)
Teacher observable characteristics				
1=female	-0.189*	-0.177*	-0.144**	-0.132**
	(0.113)	(0.099)	(0.066)	(0.060)
1=course coordinator	-0.168	-0.189*	-0.155	-0.165**
	(0.148)	(0.113)	(0.096)	(0.074)
Class characteristics				
Class size[a]	0.000	0.000	0.001***	0.001**
	(0.001)	(0.001)	(0.000)	(0.000)
Attendance[b]	0.615***	0.636***	0.846***	0.873***
	(0.196)	(0.203)	(0.107)	(0.157)
Average HS grade	9.883**	8.004*	5.393	4.031
	(4.820)	(4.491)	(3.288)	(3.239)
Average entry test score	-0.270	-0.282	-0.110	-0.113
	(0.365)	(0.347)	(0.205)	(0.203)
Share of high ability[c]	-0.352	-0.681	0.274	0.032
	(2.028)	(1.916)	(1.310)	(1.194)
Share of females	-2.133*	-1.894	-0.472	-0.312
	(1.267)	(1.313)	(0.704)	(0.593)
Share from outside Milan	-0.435	-0.028	-0.733	-0.449
	(0.887)	(1.005)	(0.584)	(0.545)
Share of top-income[d]	-1.330	-0.951	-0.565	-0.311
	(1.496)	(1.411)	(0.802)	(0.802)
Degree program dummies	yes	yes	yes	yes
Subject area dummies	yes	yes	yes	yes
Term dummies	yes	yes	yes	yes
Observations	215	215	215	215

Weighted OLS estimates. Observations are weighted by the squared root of the number of collected questionnaires in each class. Bootstrapped standard errors in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

[a] Number or officially enrolled students in the class.

[b] Attendance is monitored by random visits of university attendants to the class.

[c] Share of students in the class who are in the top 25% of the entry test score distribution.

[d] Share of students in the class who are in the highest income bracket.

Table 8: Fixed Effects and Students' Evaluations - Other items

	Teacher ability in arousing interest		Course logistics		Course workload	
	[1]	[2]	[3]	[4]	[5]	[6]
Teacher effectiveness						
Benchmark	-0.439**	-	-0.100	-	-0.263***	-
	(0.205)		(0.087)		(0.091)	
Contemporaneous	-	0.180***	-	0.092***	-	0.017
		(0.050)		(0.021)		(0.022)
Teacher observable characteristics						
1=female	-0.150	-0.142	-0.064	-0.053	-0.084**	-0.097**
	(0.116)	(0.097)	(0.046)	(0.039)	(0.042)	(0.039)
1=course coordinator	-0.100	-0.124	-0.136**	-0.138**	-0.043	-0.064
	(0.121)	(0.099)	(0.062)	(0.058)	(0.061)	(0.061)
Class characteristics						
Class size[a]	-0.000	-0.000	0.000*	0.000	0.000	0.000
	(0.001)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)
Attendance[b]	0.019	0.026	0.180**	0.209**	0.104	0.059
	(0.202)	(0.190)	(0.079)	(0.090)	(0.074)	(0.084)
Average HS grade	10.072**	8.316**	2.963	2.118	-0.658	-0.915
	(4.680)	(4.121)	(2.015)	(1.989)	(1.974)	(1.983)
Average entry test score	-0.399	-0.417	0.022	0.027	0.073	0.048
	(0.294)	(0.312)	(0.154)	(0.132)	(0.108)	(0.103)
Share of high ability[c]	0.645	0.341	-0.867	-1.022	-0.455	-0.484
	(1.723)	(1.706)	(0.840)	(0.846)	(0.686)	(0.662)
Share of females	-2.240*	-2.002	-0.618	-0.533	-0.151	-0.064
	(1.256)	(1.390)	(0.521)	(0.407)	(0.486)	(0.406)
Share from outside Milan	-0.694	-0.302	-0.321	-0.157	0.069	0.170
	(0.909)	(0.982)	(0.382)	(0.392)	(0.352)	(0.358)
Share of top-income[d]	-1.467	-1.089	0.014	0.149	0.785	0.925*
	(1.290)	(1.445)	(0.583)	(0.581)	(0.498)	(0.533)
Degree program dummies	yes	yes	yes	yes	yes	yes
Subject area dummies	yes	yes	yes	yes	yes	yes
Term fixed dummies	yes	yes	yes	yes	yes	yes
Observations	215	215	215	215	215	215

Weighted OLS estimates. Observations are weighted by the squared root of the number of collected questionnaires in each class. Bootstrapped standard errors in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

[a] Number or officially enrolled students in the class.

[b] Attendance is monitored by random visits of university attendants to the class.

[c] Share of students in the class who are in the top 25% of the entry test score distribution.

[d] Share of students in the class who are in the highest income bracket.

Table 9: Robustness check for class switching

	Overall teaching quality		Lecturing clarity	
	[1]	[2]	[3]	[4]
PANEL A: All courses				
Benchmark class effect	-0.411*	-	-0.251*	-
	(0.218)		(0.135)	
Contemporaneous class effect	-	0.195***	-	0.144***
		(0.048)		(0.030)
Observations	215	215	215	215
PANEL B: Excluding most switched course				
Benchmark class effects	-0.391*	-	-0.239*	-
	(0.213)		(0.137)	
Contemporaneous class effects	-	0.191***	-	0.142***
		(0.051)		(0.032)
Observations	207	207	207	207
PANEL C: Excluding most and second most switched course				
Benchmark class effects	-0.383**	-	-0.242*	-
	(0.183)		(0.133)	
Contemporaneous class effects	-	0.158***	-	0.129***
		(0.050)		(0.033)
Observations	199	199	199	199
PANEL D: Excluding five most switched courses				
Benchmark class effects	-0.247	-	-0.195*	-
	(0.204)		(0.119)	
Contemporaneous class effects	-	0.146	-	0.110*
		(0.079)		(0.056)
Observations	162	162	162	162

Weighted OLS estimates. Observations are weighted by the squared root of the number of collected questionnaires in each class.
Additional regressors: teacher characteristics (gender and coordinator status), class characteristics (class size, attendance, average high school grade, average entry test score, share of high ability students, share of students from outside Milan, share of top-income students), degree program dummies, term dummies, subject area dummies.
Bootstrapped standard errors in parentheses. * $p<0.1$, ** $p<0.05$, *** $p<0.01$.

Table 10: Teacher effectiveness and students evaluations by share of high ability students

	Overall teaching quality [1]	Lecturing clarity [2]
Benchmark class effects $[CE]$	-0.504**	-0.267*
	(0.270)	(0.150)
Interactions with quartiles of the distribution of high ability students		
$[CE] \times [1 = 2^{nd}\ quartile]$	0.104	0.049
	(0.212)	(0.127)
$[CE] \times [1 = 3^{rd}\ quartile]$	0.034	0.001
	(0.225)	(0.133)
$[CE] \times [1 = 4^{th}\ quartile]$	0.166	0.062
	(0.210)	(0.106)
Observations	215	215

Weighted OLS estimates. Observations are weighted by the squared root of the number of collected questionnaires in each class.
Additional regressors: teacher characteristics (gender and coordinator status), class characteristics (class size, attendance, average high school grade, average entry test score, share of high ability students, share of students from outside Milan, share of top-income students), degree program dummies, term dummies, subject area dummies.
Bootstrapped standard errors in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

Appendix

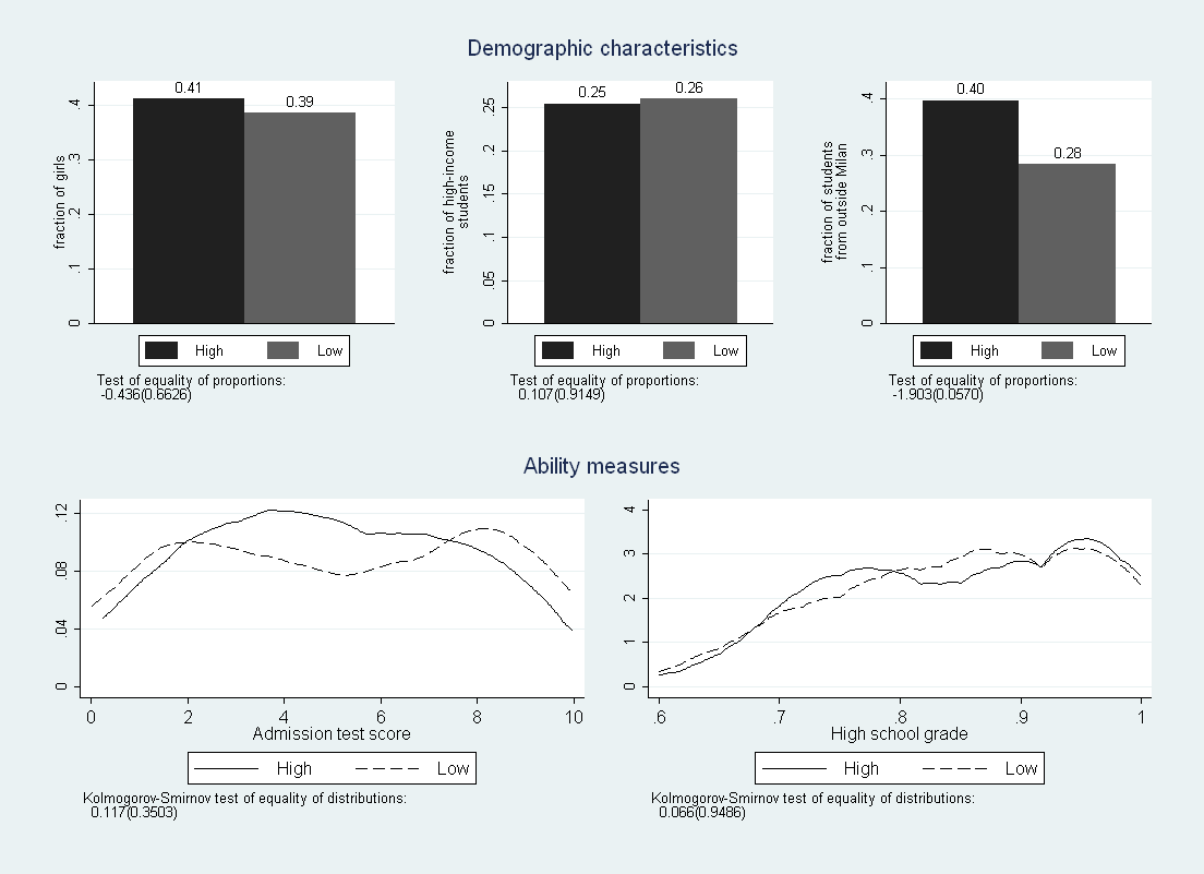


Figure A-1: Additional evidence of random allocation - Management

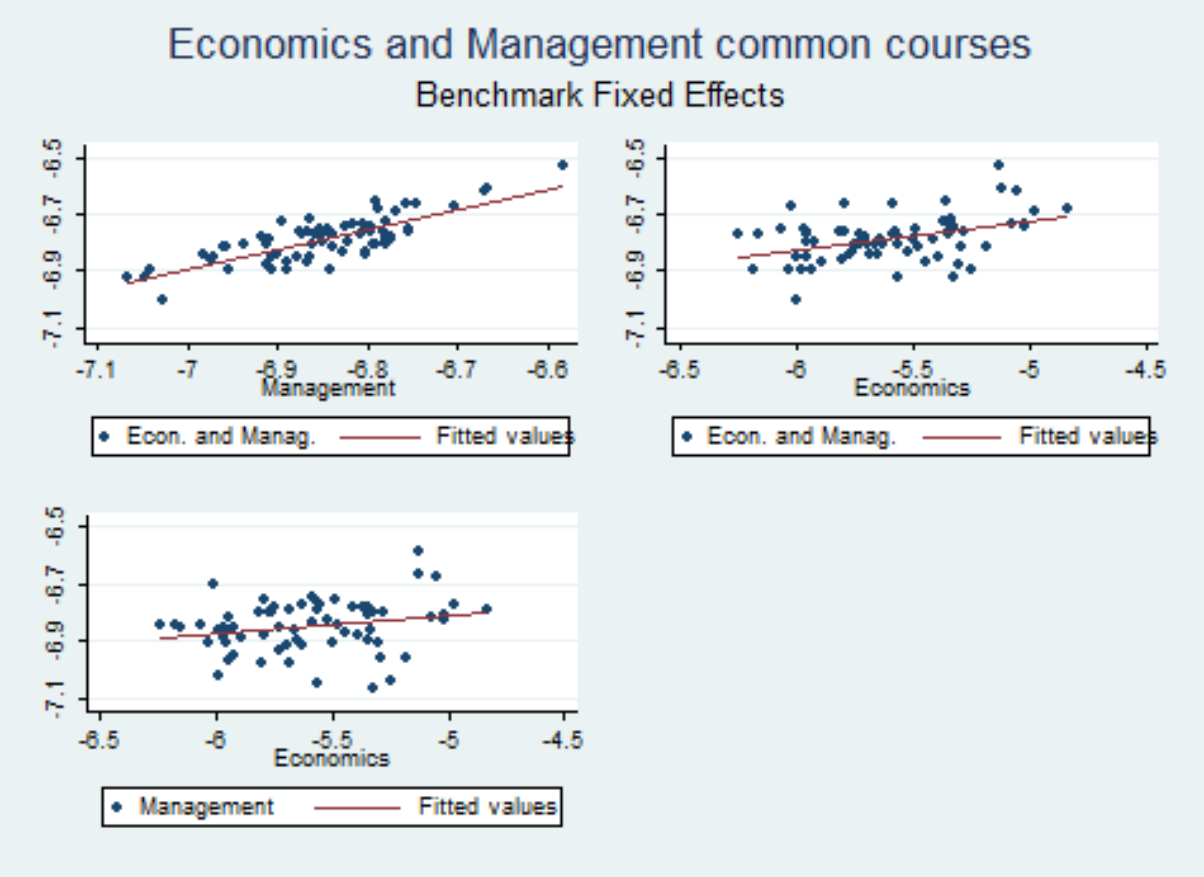


Figure A-2: Economics and Management common courses - Benchmark Fixed Effects

Table A-1: Economics and Management common courses

	Economics and Management	Management	Economics
Econ. and Manag.	-	0.900***	0.851***
		(0.008)	(0.040)
Management	1.113***	-	0.829***
	(0.010)	-	(0.034)
Economics	0.669***	0.615***	-
	(0.057)	(0.056)	-
Observations	68	68	68

Weighted OLS estimates. Observations are weighted by the standard error of the dependent variable. Bootstrapped standard errors in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

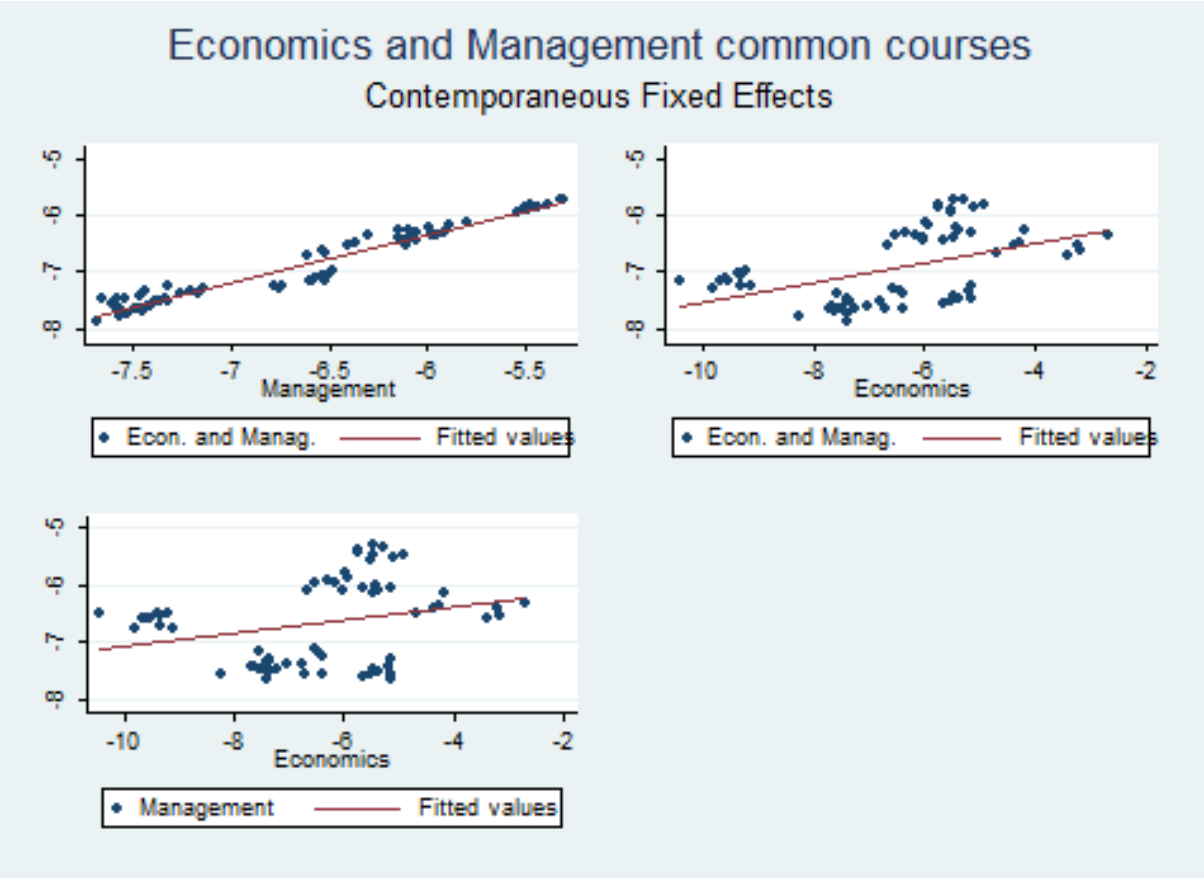


Figure A-3: Economics and Management common courses - Contemporaneous Fixed Effects