

Evaluating students' evaluations of professors *

Michela Braga

Università Statale di Milano

Marco Paccagnella

Bank of Italy and Bocconi University

Michele Pellizzari

Bocconi University, IGER, IZA and FRDB

First Version: September 2010

Current version: July, 2011

Abstract

This paper contrasts measures of teacher effectiveness with the students' evaluations for the same teachers using administrative data from Bocconi University. The effectiveness measures are estimated by comparing the performance in follow-on coursework of students who are randomly assigned to teachers. We find that teacher quality matters substantially and that our measure of effectiveness is negatively correlated with the students' evaluations of professors. A simple theory rationalizes this result under the assumption that students evaluate professors based on their realized utility, an assumption that is supported by additional evidence that the evaluations respond to meteorological conditions.

Keywords: Teacher quality, Postsecondary Education.

JEL Codes: I20

*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, Giovanni Bruno, Giacomo De Giorgi, Marco Leonardi, Tommaso Monacelli, Tommy Murphy and Tommaso Nannicini for their precious comments. We would also like to thank seminar participants at the Bank of Italy, Bocconi University, London School of Economics, UC Berkeley, Università Statale di Milano and LUISS University. Davide Malacrino and Alessandro Ferrari provided excellent research assistance. 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.

1 Introduction

The use of anonymous students' evaluations of professors to measure teachers' performance has become extremely popular in many universities around the world (William E. Becker & Michael 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 levels of effort, can be observed precisely. In fact, students' evaluations are often used to inform hiring and promotion decisions (Becker & 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 (Byron W. Brown & Daniel H. Saks 1987).¹

The validity of anonymous students' evaluations as indicators of teacher ability 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. 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, Antony C. Krautmann & William Sander (1999) show that, conditional on learning, teachers who give higher grades also receive better evaluations, a finding that is confirmed by several other studies and that is thought to be a key cause of grade inflation (Bruce A. Weinberg, Belton M. Fleisher & Masanori Hashimoto 2009, Scott E. Carrell & James E. West 2010).

Measuring teaching quality is complicated also because the most common observable teachers' characteristics, such as their qualifications or experience, appear to be relatively unimportant (Alan B. Krueger 1999, Steven G. Rivkin, Eric A. Hanushek & John F. Kain 2005, Eric A. Hanushek & Steven G. Rivkin 2006). Despite such difficulties, there is also ample evidence that teachers' quality matters substantially in determining students' achievement (Rivkin, Hanushek & Kain 2005, Carrell & West 2010) and that teachers respond to incentives (David N. Figlio & Lawrence Kenny 2007, Esther Duflo, Rema Hanna & Micheal Kremer 2010, Victor Lavy 2009). Hence, understanding how professors should (or should not) be monitored and incentivized is of primary importance.

In this paper we evaluate the content of the students evaluations by contrasting them with

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

objective measures of teacher effectiveness. We construct such measures by comparing the performance in subsequent coursework of students who are randomly allocated to different teachers in their compulsory courses. For this exercise we use data about one cohort of students at Bocconi University - the 1998/1999 freshmen - who were required to take a fixed sequence of compulsory courses and who were randomly allocated to a set of teachers for each of such courses. Additionally, the data are exceptionally rich in terms of observable characteristics, in particular they include measures of cognitive ability, family income and entry wages, which are obtained from regular surveys of graduates.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 about 3% of the average grade. This effect translates into approximately 1.4% of the average entry wage or 14 euros per month (160-200 euros per year). Moreover, our measure of teaching quality appears to be negatively correlated with the students' evaluations of the professors: 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 good teachers are those who provide their students with knowledge that is useful in future learning and, at the same time, require high effort from their students. Students are heterogeneous in their disutility of effort, which is higher for the least able ones, and evaluate professors on the basis of their realized utility, which depends on grades/learning and effort. In this setting, students in the bottom part of the ability distribution may, in fact, give worse evaluations to the good teachers, who impose a high effort cost on them, than the bad teachers.

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, a piece of evidence that lends support to our specification of the learning function. Additionally, in order to support our assumption that evaluations are based on students' realized utility, we match our data with the weather conditions observed on the exact days when students filled the evaluation questionnaires. Under the assumption that the weather affects utility and not teaching quality, finding that the students' evaluations react to meteorological conditions lends support to the specification of our model.³ Our results show that students evaluate professors more negatively on rainy

²The same data are used in Giacomo De Giorgi, Michele Pellizzari & Silvia Redaelli (2010).

³One may actually think that also the mood of the professors, hence, their effectiveness in teaching is affected by the weather. However, students are asked to evaluate teachers' performance over the entire duration of the course and not exclusively on the day of the test. Moreover, it is a clear rule of the university to have students fill the questionnaires before the lecture, so that the teachers' performance on that specific day should not affect the

and cold days.

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 (Jonah E. Rockoff & Cecilia Speroni 2010, Jonah E. Rockoff 2004, Rivkin, Hanushek & Kain 2005, Figlio & Kenny 2007, Brian A. Jacob & Lars Lefgren 2008, John H. Tyler, Eric S. Taylor, Thomas J. Kane & Amy L. Wooten 2010, Thomas J. Kane & Douglas O. Staiger 2008). 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 (OECD 2010, Ina V.S. Mullis, Michael O. Martin, David F. Robitaille & Pierre Foy 2009). The large degree of heterogeneity in subjects and syllabuses in universities makes it very difficult 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.⁴

To the best of our knowledge, only three other papers investigate the role of students' evaluations in university, namely Carrell & West (2010), Florian Hoffman & Philippe Oreopoulos (2009) and Weinberg, Fleisher & 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 Hoffman & Oreopoulos (2009) and Weinberg, Fleisher & Hashimoto (2009), who cannot purge their estimates from the potential bias due to the best students selecting the courses of the best professors. Jesse Rothstein (2009) and Jesse Rothstein (2010) show that correcting such a selection bias is pivotal to producing reliable measures of teaching quality.

The study of Carrell & West (2010), a paper that was developed parallelly and independently of ours, is perhaps the most similar to ours, both in terms of methodology and results. They also document a surprising negative correlation between the students' evaluations of professors and harder measures of teaching quality, however, we improve on their analysis in at least three important dimensions. First and most important, we provide a theoretical framework for the interpretation of such a striking finding, which is absent in Carrell & West (2010). Given that our results forcefully challenge the current most popular method used by most universities around the world to measure the teaching performances of their employees, it is paramount to provide a model that can rationalize the behaviors of both students and professors which generate the observed data. Furthermore, we show that our theory is consistent with additional pieces of evidence and we use it to formulate policy proposals.

evaluations.

⁴On 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.

Second, by observing wages for our students we are able to attach a price tag to our measures of teacher quality, something that, to our knowledge, has never been possible in previous studies.⁵

Finally, Carrell & West (2010) use data from a U.S. Air Force Academy, while our empirical application is based on a more standard institution of higher education.⁶ In particular, the vast majority of the students in our sample enter a standard labor market when they graduate, whereas the cadets in Carrell & West (2010) are required to serve as officers in the U.S. Air Force for 5 years after graduation and many probably pursue a longer military career. There are many reasons why the behaviors of both teachers, students and the university/academy might vary depending on the labor market they face. For example, students may put particular effort on some exams or activities that are particularly important in the military setting - like physical activities - at the expenses of other subjects and teachers and administrators may do the same.

More generally, this paper is also related and contributes to the wider literature on performance measurement and performance pay. For example, one concern with the 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 (Bengt Holmstrom & Paul 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 (George Baker, Robert Gibbons & Kevin J. Murphy 1994, Candice Prendergast & Robert H. 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 presents our strategy to estimate teacher effectiveness and shows the results. In Section 4 we correlate teacher effectiveness with the students' evaluations of professors. Robustness checks are reported in Section 5. In Section 6 we present a simple theoretical framework that rationalizes our results, while Section 7 discusses some additional evidence that corroborates our model. Finally, Section 8 concludes.

⁵Raj Chetty, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach & Danny Yagan (2011) present some results in this same spirit but in a very different context (kindergarten) and without explicitly looking at measures of teaching quality (they rather consider teachers' experience).

⁶Bocconi 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 mid-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 track positions 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 (see the *Financial Times Business Schools Rankings*).

2 Data and institutional details

The empirical analysis in this paper is based on data for one enrollment 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.⁷

In later cohorts, the random allocation was repeated at the beginning of each academic year, so that students would take all the compulsory courses of each academic year with the same group of classmates, which only permits to identify the joint effectiveness of the entire set of teachers in each academic year.⁸ 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&Management⁹. Students in these programs were required to take a fixed sequence of compulsory courses that span the entire duration of their first two years, a good part of their third year and, in a few cases, also their last year. Table 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).¹⁰ 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 analysis.

[INSERT TABLE 1 ABOUT HERE]

Most (but not all) of the courses listed in Table 1 were taught in multiple classes (see Section 3 for details). The number of such classes varied across both degree programs and

⁷The 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.

⁸Giacomo De Giorgi, Michele Pellizzari & William G. Woolston (2011) use data for these later cohorts for a study of class size.

⁹The other degree programs were Economics and Social Disciplines, Economics and Finance, Economics and Public Administration.

¹⁰Notice 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 & 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 (see Figure A-2).

specific courses. For example, Management was the program that attracted the most students (over 70% in our cohort), who were normally divided into 8 to 10 classes. Economics and Law&Management students were much fewer and were rarely allocated to more than just two classes. Moreover, the number of classes also varied within degree programs depending on the number of available teachers in each course. For instance, in 1998/99 Bocconi did not have a law department and all law professors were contracted from other nearby universities. Hence, the number of classes in law courses were normally fewer than in other subjects. Similarly, since the management department was (and still is) much larger than the economics or the mathematics 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 were 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 (within degree program), regardless of their classes. Specifically, one of the teachers in each course (normally a senior person) acted as a coordinator, making sure 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. Other than this check, the grades were not curved, neither across nor within classes.

[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 took 3 courses, divided into a total of 24 different classes: management I, which was 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 high school (academic or technical/vocational), 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.¹¹ Additionally, all students took a

¹¹The 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 entering cohort, is just above 10%. Notice that some of these students might have transferred to another university or still be working towards the completion of their

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 varied with 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 student 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 were enrolled in the Management program (74%), while Economics and Law&Management attracted 11% and 14%. Female students were generally under-represented in the student body (43% overall), apart from the degree program in Law&Management. About two thirds of the students came from outside the province of Milan, which is where Bocconi is located, and such a share increased to 75% in the Economics program. Family income was recorded in brackets and one quarter of the students were in the top bracket, whose lower threshold was in the order of approximately 110,000 euros at current prices. Students from such a wealthy background were 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 attracted 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 contacts were successful but we were still able to match almost 60% of the students in our cohort, a relatively good response rate for surveys.¹²

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 were asked to fill an evaluation questionnaire during one lecture. The questions gathered students' opinions about various aspects of the teaching experience, including the clarity of the lectures, the logistics of the course, the availability of the professor and so on. For each item in the questionnaire, students answered on a scale from 0 (very negative) to 10 (very positive) or 1 to 5.

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

program, whose formal duration was 4 years. In Section 5 we perform a robustness check to show that excluding the dropouts from our calculations is irrelevant for our results.

¹²The response rates are highly correlated with gender, because of compulsory military service, and with the 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.

individual student with a specific evaluation of the teacher. 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 A-1 in the Appendix shows, as an example, the first page of the evaluation questionnaire used in the academic year 1998-1999.¹³

In Table 4 we present some descriptive statistics of the answers to the evaluation questionnaires. We concentrate on a limited set of items, which we consider to be the most informative and interesting, namely overall teaching quality, lecturing clarity, the teacher's ability to generate interest in the subject, the logistic of the course and workload. These are the same items that we analyze in more details in Section 4. The exact wording and scaling of the questions are reported in Table A-4 in the Appendix.

The average evaluation of overall teaching quality is around 7, with a relatively large standard deviation of 0.9 and minor variations across degree programs. Although differences are not statistically significant, professors in the Economics program seem to receive slightly better students' evaluations than their colleagues in Management and, even more, in Law&Management. The same ranking holds for the other measures of teaching quality, namely the clarity of lecturing and the ability to generate interest in the subject. Economics compares slightly worse to the other programs in terms of course logistics

[INSERT TABLE 4 ABOUT HERE]

Some of the evaluation items are, understandably, highly correlated. For example, the correlation coefficient between overall teaching quality and lecturing clarity is 0.89. The course logistics and the ability of the teacher in generating interest for the subject are slightly less strongly correlated with the core measures of teacher quality (around 0.5-0.6). Workload is the least correlated with any other item (all correlation coefficients are below 0.2). The full correlation matrix is reported in Table A-5 in the Appendix.

Additionally, in Table 4 we also report the mean and standard deviations of the number of collected questionnaires and the number of officially enrolled students in each of class. One might actually be worried that students may drop out of a class in response to the quality of the teaching so that at the end of the course, when questionnaires are distributed only the students who liked the teacher are eventually present. Such a process would lead to a compression of the distribution of the evaluations, with good teachers being evaluated by their entire class (or by a majority of their allocated students) and bad teachers being evaluated only by a subset of students who particularly liked them.

The descriptive statistics reported in Table 4 seem to indicate that this is not a major issue, as on average the number of collected questionnaires is around 80% of the total number of enrolled students (the median is very similar). Moreover, when we correlate our measures of

¹³The 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.

teaching effectiveness with the evaluations we condition on the official size of the class and we weight observations by the number of questionnaires.

Indirectly, the relatively high number of questionnaires over students is evidence that attendance was also pretty high. An alternative measure of attendance can be extracted from a direct question of the evaluation forms which asks students what percentage of the lectures they attended. Such a self-reported measure of attendance is also around 80%.

2.1 The random allocation

In this section we present evidence that the random allocation of students into classes was successful. De Giorgi, Pellizzari & 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 adopted the policy of repeating the randomization for each course with the explicit purpose of encouraging wide interactions among the students.

[INSERT TABLE 5 ABOUT HERE]

Table 5 is based on test statistics derived from probit (columns 1,2,5,6,7) or OLS (columns 3 and 4) regressions of the observable students' characteristics (by column) on class dummies for each course in each degree program that we consider. The null hypothesis under consideration is the joint significance of the coefficients on the class dummies in each model, which amounts to testing for the equality of the means of the observable variables across classes. The table shows some descriptive statistics about the distribution of p-values for such tests.

The mean and median p-values are in all cases far from the conventional thresholds of 5% or 1% and only in a very few instances the null cannot be rejected by the data. The most notable exception is residence from outside Milan, which is abnormally low in two Management groups. Overall, Table 5 suggests that the randomization was rather successful.

[INSERT FIGURE 1 ABOUT HERE]

Testing the equality of means is not a sufficient test of randomization for continuous variables. Hence, in Figure 1 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 Kolmogorov-Smirnov tests confirm the visual impression.

Even though students were randomly assigned to classes, one may still be concerned about teachers being selectively allocated to classes. Although no explicit random algorithm was used to assign professors to classes, for obvious organizational reasons that was (and still is) done in

the Spring of the previous academic year, i.e. well before students were allowed to enroll, so that even if teachers were allowed to choose their class identifiers they would have no chance to know in advance the characteristics of the students who would be given that same identifier.

More specifically, there used to be (and still is) a very strong hysteresis in the matching of professors to class identifiers, so that, if no particular changes occurred, one kept the same class identifier of the previous academic year. It is only when some teachers needed to be replaced or the overall number of classes changed that modifications took place. Even in these instances, though, the distribution of class identifiers across professors changed only marginally. For example if one teacher dropped out, then a new teacher would take his/her class identifier and none of the others were given a different one. Similarly, if the total number of classes needed to be increases, the new classes would be added at the bottom of the list of identifiers with new teachers and no change would affect the existing classes and professors.¹⁴

About around the same time when teachers were given class identifiers (i.e. in the Spring of the previous academic year), also classrooms and time schedules were defined. On these two items, though, teachers did have some limited choice. Typically, the administration suggested a time schedule and room allocation and professors could request one or more modifications, which were accommodated only if compatible with the overall teaching schedule (e.g. a room of the required size was available at the new requested time).

In order to avoid any distortion in our estimates of teaching effectiveness due to the more or less convenient teaching times, we collected detailed information about the exact timing of the lectures in all the classes that we consider, so that we can hold this specific factor constant (see Section 3). Additionally, we also know in which exact room each class was taught and we further conditions on the characteristics of the classrooms, namely the building and the floor where they are located. There is no variation in other features of the rooms, such as the furniture (all rooms were - and still are - fitted with exactly the same equipment: projector, computer, white-board) or the orientation (all rooms face the inner part of the campus where there is very limited car traffic).¹⁵

Table 6 provides evidence of the lack of correlation between teachers and classes' characteristics, namely we show the results of regressions of teachers' observable characteristics on classes' observable characteristics. For this purpose, we estimate a system of 9 seemingly unrelated simultaneous equations, where each observation is a class in a compulsory course. The dependent variables are 9 teachers' characteristics (age, gender, h-index, average citations per year and 4 dummies for academic positions) and the regressors are the class characteristics listed in the rows of the table.¹⁶ The reported statistics test the null hypothesis that the

¹⁴As far as we know, the total number of classes for a course has never been reduced.

¹⁵In principle we could also condition on room fixed effects but there are several rooms in which only one class of the courses that we consider was taught.

¹⁶The h-index is a quality-adjusted measure of individual citations based on search results on Google Scholar. It was proposed by J. E. Hirsch (2005) and it is defined as follows: *A scientist has index h if h of his/her N_p papers have at least h citations each, and the other $(N_p - h)$ papers have no more than h citations each.*

coefficients on each class characteristic are all jointly equal to zero in all the equations of the system.¹⁷

[INSERT TABLE 6 ABOUT HERE]

Results show that only the time of the lectures is significantly correlated with the teachers' observables at conventional statistical levels. In fact, this is one of the few elements of the teaching planning over which teachers had some limited choice. More specifically, professors are given a suggested time schedule for their classes in the spring of the previous academic year (usually based on the schedule of the current year), and they can either approve it or request changes. The administration, then, accommodates such changes only if they are compatible with the other many constraints in terms of rooms availability and course overlappings. In our empirical analysis we do control for all the factors in Table 6, so that our measures of teaching effectiveness are purged from the potential confounding effect of teaching times on students' learning.

3 Estimating teacher effectiveness

We use performance data 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 enjoyed better outcomes later on. This approach is similar to the *value-added* methodology that is more commonly used in primary and secondary schools (Dan Goldhaber & Michael Hansen 2010, Eric A. Hanushek & Steven G. Rivkin 2010, Hanushek & Rivkin 2006, Rothstein 2009, Rivkin, Hanushek & Kain 2005, Eric A. Hanushek 1979) but it departs from its standard version, that uses contemporaneous outcomes and conditions on past performance, since we use future performance to infer current teaching quality.¹⁸

One most obvious 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 unobservables.

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 electives. Moreover, elective courses were usually taken by fewer students than compulsory ones and they were usually taught in one single class.

¹⁷To construct the tests we use the small sample estimate of the variance-covariance matrix of the system.

¹⁸For this reason we prefer to use the label *teacher effectiveness* for our estimates.

We compute our measures of teacher effectiveness in two steps. First, we estimate the conditional mean of the future grades (in compulsory courses) of students in each class according to the following procedure. Consider a set of students enrolled in degree program d and indexed by $i = 1, \dots, N_d$, 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 $c = 1, \dots, C_d$, 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 $s = 1, \dots, S_c$, 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 student i attends in semester $t \geq t_c$, where t_c denotes the semester in which course c is taught. Z_c is the set of compulsory courses taught in any term $t \geq t_c$.

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:

$$y_{ids\zeta} = \alpha_{dcs} + \beta X_i + \epsilon_{ids\zeta} \quad (1)$$

where X_i is a vector of student-level characteristics including a gender dummy, a dummy for whether the student is in the top income bracket, the entry test score and the high school leaving grade. The α 's are our parameters of interest and they measure the conditional means of the future grades of students in 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.¹⁹

Notice that, since in general there are several subsequent courses ζ for each course c , each student is observed multiple times and the error terms $\epsilon_{ids\zeta}$ are serially correlated within i and across ζ . We address this issue by adopting a standard random effect model to estimate all the equations 1 (we estimate one such equation for each course c). Moreover, we further allow for cross-sectional correlation among the error terms of students in the same class by clustering the standard errors at the class level.

More formally, we assume that the error term is composed of three additive components (all with mean equal zero):

$$\epsilon_{ids\zeta} = v_i + \omega_s + \nu_{ids\zeta} \quad (2)$$

where v_i and ω_s are, respectively, an individual and a class component, and $\nu_{ids\zeta}$ is a purely random term. Operatively, we first apply the standard random effect transformation to the original model of equation 1.²⁰

¹⁹Notice 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.

²⁰The standard random effect transformation subtracts from each variable in the model (both the dependent

In the absence of other sources of serial correlation (i.e if the variance of ω_s were zero), such a transformation would lead to a serially uncorrelated and homoskedastic variance-covariance matrix of the error terms, so that the standard random effect estimator could be produced by running simple OLS on the transformed model. In our specific case, we further cluster the transformed errors at the class level to account for the additional serial correlation induced by the term ω_s .

Overall, we are able to estimate 230 such fixed effects, the large majority of which are for Management courses.²¹ Descriptive statistics of the estimated α 's are reported in Table A-1 in the Appendix.

The second step of our approach is meant to purge the estimated α 's from the effect of other class characteristics that might affect the performance of students in later courses but are not attributable to teachers. 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 are documented to be important ingredients of the education production function, such as class size and class composition (De Giorgi, Pellizzari & Woolston 2011).

A key 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 and class composition (in terms of students' characteristics). Additionally, we also have access to the identifiers of the teachers in each class and we can recover a large set of variables like gender, tenure status, and measures of research output. We also know which of the several teachers in each course acted as coordinator. These are the same teacher characteristics that we used in Table 6. Once we condition on all these observable controls, unobservable teaching quality is likely to be the only remaining factor that generates variation in the estimated α 's. At a minimum, it should be uncontroversial that teaching quality is by far the single most important unobservable that generates variation in the $\hat{\alpha}$'s, once conditioning on the observables.

The effect of social interactions among the students might also affect the estimated $\hat{\alpha}$'s. However, notice that if such effects are related to the observable characteristics of the students, then we are able to control for those. Additionally, there might be complementarities among teachers' ability and students' interactions, as good teachers are also those who stimulate fruit-

and each of the regressors) its within-mean scaled by the factor $\theta = 1 - \sqrt{\frac{\sigma_v^2}{|Z_c|(\sigma_\omega^2 + \sigma_v^2) + \sigma_v^2}}$, where $|Z_c|$ is the cardinality of Z_c . For example, the random-effects transformed dependent variable is $y_{ids\zeta} - \theta \bar{y}_{ids}$, where $\bar{y}_{ids} = |Z_c|^{-1} \sum_{h=1}^{|Z_c|} y_{idh\zeta}$. Similarly for all the regressors. The estimates of σ_v^2 and $(\sigma_\omega^2 + \sigma_v^2)$ that we use for this transformation are the usual Swamy-Arora, also used by the command *xtreg* in Stata (P. A. V. B. Swamy & S. S. Arora 1972).

²¹We 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 $S_c = 1$. The evidence that we reported in Tables 5 and 6 also refer to the same set of 230 classes.

ful collaborations among their students. This component of the social interaction effects is certainly something that one would like to incorporate in a measure of teaching quality, as in our analysis.

[INSERT TABLE 7 ABOUT HERE]

Thus, in Table 7 we regress the estimated α 's on all observable class and teacher characteristics. In column 1 we condition only on class size and class composition, in column 2 only on information about the teachers and in column 3 we combine the two sets of controls. In all cases we weight observations by the inverse of the standard error of the estimated α 's to take into account differences in the precision of such estimates. Consistently with previous studies on the same data (De Giorgi, Pellizzari & Woolston 2011), we find that larger classes tend to be associated with worse learning outcomes, that classes with more able students, measured with either high school grades or the entry test score, also perform better and that a high concentration of high income students appears to be detrimental for learning. Overall, observable class characteristics explain about 8% of the variation in the estimated α 's within degree program, term and subject cells, where subjects are defined as in Table 1.²²

The results in column 2 show a non linear relationship between teachers' age and teaching outcomes, which might be rationalized with increasing returns to experience. Also, professors who are more productive in research seem to be less effective as teachers, when output is measured with the h-index. The effect is reversed using yearly citations but it never reaches acceptable levels of statistical significance. Finally, and consistently with the age effect, also the professor's academic position matters, with a ranking that gradually improves from assistant to associate to full professors (other academic positions, such as external or non tenured-track teachers, are the excluded group). However, as in Hanushek & Rivkin (2006) and Krueger (1999), we find that the individual traits of the teachers explain less than a tenth of the (residual) variation in students' achievement. Overall, the complete set of observable class and teachers' variables explains approximately 15% of the (residual) variation.

Our final measures of teacher effectiveness are the residuals of the regression of the estimated α 's on all the observable variables, i.e the regression reported in column 3 of Table 7. In Table 8 we present descriptive statistics of such measures.

[INSERT TABLE 8 ABOUT HERE]

The overall standard deviation of teacher effectiveness is 0.086.²³ This average is the composition of a larger variation among the courses of the program in Economics (0.159) and a more limited variation in Management (0.069) and Law & Management (0.019). Recall

²²The Partial R-squared reported at the bottom of the table refer to the R-squared of a partitioned regression where the dummies for the degree program, the term and the subject are partialled out.

²³The standard deviation that we consider is the OLS estimate of the residuals of the regression in column 3 of Table 7

that grades are normalized so that the distributions of the class effects are comparable across courses. Hence, these results can be directly interpreted in terms of changes in outcomes. In other words, the overall effect of increasing teacher effectiveness by one standard deviation is an increase in the average grade of subsequent courses by 0.086 standard deviations, roughly 0.3 of a grade point or 1.1% over the average grade of approximately 26.²⁴ Given an estimated conditional elasticity of entry wages to GPA of 0.45, such an effect would cost students slightly more than 0.5% of their average entry monthly wage of 967 euros, or about 60 euros per year.²⁵ Since in our data we only observe entry wages, it might well be that the long term effects of teaching quality are even larger.

In Table 8 we also report the standard deviations of teacher effectiveness of the courses with the least and the most variation to show that there is substantial heterogeneity across courses. Overall, we find that in the course with the highest variation (management I in the Economics program) the standard deviation of our measure of effectiveness is approximately a quarter of a standard deviation in grades. This compares to a standard deviation of essentially zero (0.010) in the course with the lowest variation (mathematics in the Law&Management program).

In the lower panel of Table 8 we show the mean (across courses) of the difference between the largest and the smallest indicators of teacher effectiveness, which allows us to compute the effect of attending a course in the class of the best versus the worst teacher. On average, this effect amounts to 0.230 of a standard deviation, that is almost 0.8 grade points or 3% over the average grade. As already noted above, this average effect masks a large degree of heterogeneity across subjects ranging from almost 80% to a mere 4% of a standard deviation.

To further understand the importance of these effects, we can also compare particularly lucky students, who are assigned to good teachers (defined as those in the top 5% of the distribution of effectiveness) throughout their sequence of compulsory courses, to particularly unlucky students, who are always assigned to bad teachers (defined as those in the bottom 5% of the distribution of effectiveness). The average grades of these two groups of students are 1.8 grade points apart, corresponding to over 7% of the average grade. Based on our estimate of the wage elasticity, this difference translates into a sizable 300-400 euros per year (30.45 euros/month) or 3.15% over the average.

For robustness and comparison, we estimate the class effects in two alternative ways. First, 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

²⁴In 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.

²⁵In Italy wages are normally paid either 13 or 14 times over the year, once every month plus one additional payment around mid December (*tredicesima*) and around mid June (*quattordicesima*).

improving students performance in financial mathematics than in business law). The subject areas are defined by the colors in Table 1 and correspond to the department that was responsible for the organization and teaching of the course. 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 and using fewer observation, which is the reason why we do not take them as our benchmark.

Next, rather than using performance in subsequent courses, we run equation 1 with the grade in the same course c as the dependent variable. We label these estimates *contemporaneous* effects²⁶. We do not consider these contemporaneous effects as alternative and equivalent measures of teacher effectiveness, but we will use them to show that they correlate very differently with the students' evaluations. Descriptive statistics for the subject and contemporaneous effects are reported in Tables A-3 and A-2 in the Appendix.

[INSERT TABLE 9 ABOUT HERE]

In Table 9 we investigate the correlation between these alternative estimates of teacher effectiveness. Specifically, we report results from weighted OLS regressions with our benchmark estimates as the dependent variable and, in turn, the subject and the contemporaneous effects on the right hand side, together with dummies for degree program, term and subject area.²⁷

Reassuringly, the subject effects are positively and significantly correlated with our benchmark, while the contemporaneous effects are negatively and significantly correlated with our benchmark, a result that is consistent with previous findings (Carrell & West 2010, Weinberg, Fleisher & Hashimoto 2009, Krautmann & Sander 1999) and to which we will return 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 teachers receives from their students. We concentrate on two core items from the evaluation questionnaires, namely overall teaching quality and the overall clarity of the lectures. Additionally, we also look at other items: the teacher's ability in generating interest for the subject, the logistics of the course (schedule of classes, combinations of practical sessions and traditional lectures) and the total workload compared to other courses.

Formally, we estimate the following equation:

$$q_{dtcs}^k = \lambda_0 + \lambda_1 \hat{\alpha}_{dtcs} + \lambda_2 C_{dtcs} + \lambda_3 T_{dtcs} + \gamma_d + \delta_t + \nu_c + \epsilon_{dtcs} \quad (3)$$

²⁶When estimating *contemporaneous* effects we include past grades in the vector of student-level characteristics of equation 1

²⁷To take into account the additional noise due to the presence of generated regressors on the right hand side of these models, the standard errors are bootstrapped. Further, each observation is weighted by the inverse of the standard error of the dependent variable, which is also a generated variable.

where q_{dts}^k is the average answer to question k in class s of course c in the degree program d (which is taught in term t), $\hat{\alpha}_{dts}$ is the estimated class fixed effect from equation 1, C_{dts} is the set of class characteristics, T_{dts} is the set of teacher characteristics. γ_d , δ_t and v_c are fixed effects for degree program, term and subject areas, respectively. ϵ_{dts} is a residual error term.

Notice that the class and teacher characteristics are exactly the same as in Table 7, so that equation 3 is equivalent to a partitioned regression model of the evaluations q_{dts} on our measures of teacher effectiveness, i.e. the residuals of the regressions in Table 7, where all the observables and the fixed effects are partialled out.

Since the dependent variable in equation 3 is an average, we use 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 to take into account the presence of generated regressors (the $\hat{\alpha}$'s).

[INSERT TABLE 10 ABOUT HERE]

The first four columns of Table 10 reports the estimates of equation 3 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 all the items that we consider. In other words, teachers who are more effective in promoting future performance receive worst evaluations from their students. This relationship is statistically significant for all items (but logistics), and are of sizable magnitude. For example, one standard deviation increase in teacher effectiveness reduces the students evaluations of overall teaching quality by about 50% of a standard deviation. Such an effect could move a teacher who would otherwise receive a median evaluation down to the 31st percentile of the distribution. Effects of slightly smaller magnitude can be computed for lecturing clarity. Consistently with the findings of other studies (Carrell & West 2010, Weinberg, Fleisher & Hashimoto 2009, Krautmann & Sander 1999), when we use the contemporaneous effects (even columns) the estimated coefficients turn positive and highly significant for all items (but workload). In other words, the teachers of classes that are associated with higher grades in their own exam receive better evaluations from their students. The magnitudes of these effects is smaller than those estimated for our benchmark measures: one standard deviation change in the contemporaneous teacher effect increases the evaluation of overall teaching quality by 24% of a standard deviation and the evaluation of lecturing clarity by 11%.

The results in Table 10 clearly challenge the validity of students' evaluations of professors 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 & 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 students' 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.

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 transferred 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 2 we compare our benchmark measure of teacher effectiveness 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, term and subject fixed effects) yields an R^2 of over 88%. 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 2 ABOUT HERE]

Second, one might be worried that students might not comply with the random assignment to the classes. For various reasons they may decide to attend one or more courses in a different class from the one to which they were formally allocated. For example, they may desire to stay with their friends, who might have been assigned to 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.²⁸ Hence, we cannot exclude a priori that some students switch classes.

²⁸Possible motivations for such requests could be 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 to a class taught on the ground floor.

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

Courses in which students concentrate in the class of one or few 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 a difference is very large should be the ones that are more affected by switching behaviors.

[INSERT TABLE 11 ABOUT HERE]

In Table 11 we replicate our benchmark estimates for the two core evaluation items (overall teaching quality and lecturing clarity) 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). For comparison we also report the original estimates from Table 10 in Panel A and we find that results change only marginally. Next, in Panel C and D we exclude from the sample also the second most switched course (human resource management) and the five most switched courses, respectively.²⁹ Again, the estimated coefficients are only mildly affected, although the significance levels are reduced according with the smaller sample sizes. Overall, this exercise suggests that course switching should not affect our estimates in any major direction.

Finally, one might be worried that our results may be generated by some endogenous reaction of students to the quality of their past teachers. For example, as one meets a bad teacher in one course one might be induced to exert higher effort in the future to catch up, especially if bad teaching resulted in a lower (contemporaneous) grade. Hence, the students evaluations may reflect real teaching quality and our measure of teacher effectiveness would be biased by such a process of mean reversion, leading to a negative correlation with real teaching quality and, consequently, also with the evaluations of the students.

[INSERT FIGURE 3 ABOUT HERE]

To control for this potential feedback effect on students' effort, we recompute our benchmark measures of teacher effectiveness adding the student average grade in all previous courses to the set of controls. Figure 3 compares our benchmark teacher effectiveness with this augmented version, conditioning on the usual fixed effects for degree program, term and subject area and shows that the two are strongly correlated (even accounting for the outliers).

²⁹The five most switched courses are marketing, human resource management, mathematics for Economics and Management, financial mathematics and managerial accounting.

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 of 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.

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.

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

$$x_i = \mu_j h(e_i) + (1 - \mu_j) \bar{x} \quad (4)$$

which is a linear combination of a function $h(\cdot)$ of student's effort e_i and a constant \bar{x} , weighted 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 \bar{x} , regardless of effort. This strong assumption can obviously be relaxed and all our implications will be maintained as long as the gradient of grades to effort increases with μ_j .

The parameter \bar{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 \bar{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 material. Of course, there are costs and benefits to each type of exam and multiple-choice tests are often adopted because they can be marked quickly, easily and uncontroversially. For the sake of simplicity, however, we abstract from cost-benefit considerations.

Furthermore, equation 4 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 selecting 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 than what we present in this section, its qualitative results would be unchanged and the limited information on teachers in our data would make its additional empirical content redundant in our setting.

More specifically, one could model μ_j as an endogenous choice of the teacher and generate heterogeneity by assuming that different activities (real teaching or teaching-to-the-test) require different efforts from the professors, who face heterogeneous marginal disutilities. Such an alternative model would feature both adverse selection and moral hazard and proper measurement of teaching quality could help addressing both issues, by facilitating the identification of low quality agents (high disutility of effort) and by incentivizing effort. In our simplified framework, only adverse selection of professors takes place, but the general intuition holds also in a more complicated setting.

In all cases, a key assumption is 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:

$$U_i = x_i - \frac{1}{2} \frac{e_i^2}{\eta_i} \quad (5)$$

where η_i is a measure of student's ability.

For simplicity, we assume that students are perfectly informed about the production function of grades, i.e. they know the type of their teacher, they know the return to their effort and there is no additional stochastic component to equation 4. This assumption can be easily relaxed by introducing either imperfect information about the teacher's type or about the exact specification of the production function and, consequently, by rewriting the utility function in equation 5 in expected terms. The main intuition of our results would be unchanged. Although the perfect information assumption is obviously a modeling device and does not correspond to reality, we do believe that students know a lot about their professors, either through conversations with older students or by observation through the duration of the course.

The utility function in equation 5 implicitly assumes that students are myopic, in the sense that they care only about grades and not about real learning. The main implications of the simple theory in this section would remain unchanged also with a different utility function that incorporates real learning, as long as students of different abilities care equally about it (just like they are equally myopic in the current specification).

The quasi-linearity of equation 5 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 in the utility functions 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:

$$\mu_j \frac{\partial h(e)}{\partial e_i}(e_i^*) = \frac{e_i^*}{\eta_i} \quad (6)$$

Using equation 6 it is easy to derive the following results:

$$\frac{de_i^*}{d\eta_i} > 0 \quad (7)$$

$$\frac{de_i^*}{d\mu_j} > 0 \quad (8)$$

$$\frac{de_i^*}{d\mu_j d\eta_i} > 0 \quad (9)$$

Equation 7 shows that more able students exert higher effort. Equation 8 shows that more real teaching induces higher effort from the students and equation 9 indicates that such an effect is larger for the more able students

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

$$\frac{\partial U_i(e_i^*)}{\partial \mu_j} = h(e_i^*) - \bar{x} \quad (10)$$

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

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

$$\frac{\partial U_i(e_i^*)}{\partial \mu_j} < 0 \quad \text{if } \eta_i < \bar{\eta} \quad (12)$$

Equations 11 and 12 are particularly important under the assumption that, especially when answering questions about the overall quality of a course, students give a better evaluation to teachers (or classes) that are associated with a higher level of utility. Equations 11 and 12 suggest that high ability students evaluate better teachers or classes that are more focused on real learning while low ability students prefer teachers that teach to the test. Hence, if the (benchmark) teacher 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 students is higher.

7 Further evidence

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

First, in the model we assume that students evaluate professors on the basis of their realized utility from attending their courses. This might be a questionable assumption. Especially university administrators who organize and elaborate the students' questionnaires are often convinced that, when asked about the ability of the teacher in presenting the course material, students express their opinion regardless of whether the teacher has imposed a high effort cost on them in order to pass the exam. In fact, an alternative behavioral model would be one in which students observe the true type of the teacher and they truthfully communicate it in the questionnaires regardless of their individual classroom experience.

In order to provide support for our specification, in Table 12 we produce evidence that the students' evaluations respond to the weather conditions on the day when they were filled. There is ample evidence that people's utility (or welfare, happiness, satisfaction) improves with good meteorological conditions (Marie Connolly Pray 2011, Cristopher Barrington-Leigh 2008, Norbert Schwarz & Gerald L. Clore 1983, Matthew C. Keller, Barbara L. Fredrickson, Oscar Ybarra, Stéphane Coté, Kareem Johnson, Joe Mikels, Anne Conway & Tor Wager 2005, Jaap J. A. Denissen, Ligaya Butalid, Lars Penke & Marcel A.G. van Aken 2008) and finding that such conditions also affect the evaluations of professors suggests that they indeed reflect utility rather than (or together with) teaching quality.

Specifically, we find that evaluations improve with temperature, deteriorate with rain and improve on foggy days. The effects are significant for most of the items that we consider and the signs of the estimates are consistent across items and specifications.

[INSERT TABLE 12 ABOUT HERE]

Obviously, teachers might be affected by meteorological conditions as much as their students and one may wonder whether the estimated effects in the odd columns of Table 12 reflect the indirect effect of the weather on teaching effectiveness. We consider this interpretation to be very unlikely since the questionnaires are distributed and filled before the lecture so that students should not be able to incorporate in their answers the performance of the teacher in the day the evaluation forms are filled in. Moreover, students' are asked to evaluate teachers' performance over the entire duration of the course and not exclusively on the day of the test.

Nevertheless, in the even columns of Table 12, we also condition on our benchmark measure of teaching effectiveness and, as we expected, we find that the estimated effects of both the weather conditions and teacher effectiveness itself change only marginally.

Second, our specification of the production function for exam grades in equation 4 implies a positive relationship between grade dispersion and the professor's propensity to engage in real teaching (μ_j). In our empirical exercise our measures of teacher effectiveness can be interpreted

as measures of the μ_j 's in the terminology of the model. Hence, if grades were more dispersed in the classes of the worst teachers one would have to question our specification of equation 4.

[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 (on the horizontal axis). To take proper account of differences across degree programs, the variables on both axes are the residuals of weighted OLS regressions that condition on degree program, term and subject area fixed effects, as in standard partitioned regressions (the weights are the squared roots of class sizes). Consistently with equation 4 in our model, the two variables are positively correlated and such a correlation is statistically 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 0.011 with a standard error of 0.004.

[INSERT TABLE 13 ABOUT HERE]

Next, according to equations 11 and 12, 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 13. We define as high ability those students 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 investigate the relationship between the students' evaluations and teacher effectiveness by restricting the sample to classes in which high-ability students are over-represented. Results seem to suggest the presence of non linearities or threshold effects, as the estimated coefficient remains relatively stable until the fraction of high ability students in the class goes above one quarter or, more precisely, 27% which corresponds to the top 25% of the distribution of the presence of high-ability students. At that point, the estimated effect of teacher effectiveness on students' evaluations is about a quarter of the one estimated on the entire sample. The results, thus, suggest that the negative correlations reported in Table 10 are mostly due to classes with a particularly low incidence of high ability students.

8 Policies and conclusions

Using administrative archives from Bocconi University and exploiting random 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. These empirical findings can be rationalized with a simple model featuring heterogeneity in the preferences (or ability) of teachers to engage in real teaching rather than teaching-to-the-test, with the former requiring higher effort from students than the latter.

Furthermore, we also find that weather conditions on the day the questionnaires are filled are correlated with the students' evaluations of teachers. This is consistent with the assumption of our model that the evaluations reflect students' perceived utility more than teachers' ability as such. Overall, our results cast serious doubts on the validity of students' evaluations of professors as measures of teaching quality or effort.

At the same time, the strong effects of teaching quality on students' outcomes, as documented in Section 3, suggest that improving the quantity or the quality of professors' inputs in the education production function can lead to large gains. Under the interpretation offered by our model in Section 6, this could be achieved through various types of interventions.

For example, our analysis suggests that the evaluations of the best students are more aligned with actual teachers' effectiveness. Hence, the opinions of the very good students could be given more weight in the measurement of professors' performance. In order to do so, some degree of anonymity of the evaluations must be lost, as it must be possible to link the evaluations of individual students to their academic records. Of course, there is no need for the teachers to be able to make such a link and only the administration should have all the necessary information. There are certainly ways to make the separation of information between administrators and professors credible to the students so as not to bias their evaluations.

Moreover, one may think of adopting exam formats that reduce the returns to teaching-to-the-test, although this may come at larger costs due to the additional time needed to grade less standardized tests. At the same time, the extent of grade leniency could be greatly limited by making sure that teaching and grading are done by different persons. Anecdotically, we know that at Bocconi it is common practice among the teachers of the core statistics course to randomize the grading, i.e. at the end of the course the teachers of the different classes are randomly assigned the papers of another class for marking. In the only year in which this practice was abandoned, average grades increased substantially.

Another variation to the current most common use of the students' evaluations consists in postponing the collection of students' opinions, so as to give them time to appreciate the value of real teaching in subsequent learning (or even in the market). Obviously, this would also pose problems in terms of recall bias and possible retaliation for low grading.

Alternatively, one may also think of other forms of performance measurement that are more in line with the peer-review approach adopted in the evaluation of research output. It is already common practice in several departments to have colleagues sitting in some classes and observing teacher performance, especially of assistant professors. This is often done primarily with the aim of offering advice, but in principle it could also be used to measure outcomes. An obvious concern is that one could change behavior due to the presence of the observer. A slightly more sophisticated version of the same method could be based on the use of cameras to record a few teaching sessions during the course without the teacher knowing exactly which ones. The video recordings could then be viewed and evaluated by an external professor in the same field.

Obviously, these, as well as other potential alternative measurement methods, are costly but

they should be compared with the costs of the current systems of collecting students' opinions about teachers, which are often non trivial.

References

- Baker, George, Robert Gibbons, and Kevin J. Murphy.** 1994. "Subjective performance measures in optimal incentive contracts." *Quarterly Journal of Economics*, 109(4): 1125–1156.
- Barrington-Leigh, Cristopher.** 2008. "Weather as a transient influence on survey-reported satisfaction with life." Draft research paper, University of British Columbia.
- Becker, William E., and Michael Watts.** 1999. "How departments of economics should evaluate teaching." *American Economic Review (Papers and Proceedings)*, 89(2): 344–349.
- Brown, Byron W., and Daniel H. Saks.** 1987. "The microeconomics of the allocation of teachers' time and student learning." *Economics of Education Review*, 6(4): 319–332.
- Carrell, Scott E., and James E. West.** 2010. "Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors." *Journal of Political Economy*, 118(3): 409–32.
- Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan.** 2011. "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence From Project STAR." *Quarterly Journal of Economics*. forthcoming.
- De Giorgi, Giacomo, Michele Pellizzari, and Silvia Redaelli.** 2010. "Identification of Social Interactions through Partially Overlapping Peer Groups." *American Economic Journal: Applied Economics*, 2(2): 241–275.
- De Giorgi, Giacomo, Michele Pellizzari, and William G. Woolston.** 2011. "Class Size and Class Heterogeneity." *Journal of the European Economic Association*. forthcoming.
- Denissen, Jaap J. A., Ligaya Butalid, Lars Penke, and Marcel A.G. van Aken.** 2008. "The Effects of Weather on Daily Mood: A Multilevel Approach." *Emotion*, 8: 662–667.
- Duflo, Esther, Rema Hanna, and Micheal Kremer.** 2010. "Incentives Work: Getting Teachers to Come to School." MIT mimeo.
- Figlio, David N., and Lawrence Kenny.** 2007. "Individual teacher incentives and student performance." *Journal of Public Economics*, 91: 901–914.
- Goldhaber, Dan, and Michael Hansen.** 2010. "Using performance on the job to inform teacher tenure decisions." *American Economic Review (Papers and Proceedings)*, 100(2): 250–255.
- Hanushek, Eric A.** 1979. "Conceptual and empirical issues in the estimation of educational production functions." *Journal of Human Resources*, 14: 351–388.
- Hanushek, Eric A., and Steven G. Rivkin.** 2006. "Teacher quality." In *Handbook of the Economics of Education*. Vol. 1, , ed. Eric A. Hanushek and Finis Welch, 1050–1078. Amsterdam:North Holland.
- Hanushek, Eric A., and Steven G. Rivkin.** 2010. "Generalizations about using value-added measures of teacher quality." *American Economic Review (Papers and Proceedings)*, 100(2): 267–271.

- Hirsch, J. E.** 2005. “An index to quantify an individual’s scientific research output.” *Proceedings of the National Academy of Sciences of the United States of America*, 102(46): 16569–16572.
- Hoffman, Florian, and Philippe Oreopoulos.** 2009. “Professor Qualities and Student Achievement.” *The Review of Economics and Statistics*, 91(1): 83–92.
- Hogan, Timothy D.** 1981. “Faculty research activity and the quality of graduate training.” *Journal of Human Resources*, 16(3): 400–415.
- Holmstrom, Bengt, and Paul Milgrom.** 1994. “The firm as an incentive system.” *American Economic Review*, 84(4): 972–991.
- Jacob, Brian A., and Lars Lefgren.** 2008. “Can principals identify effective teachers? Evidence on subjective performance evaluation in education.” *Journal of Labor Economics*, 26: 101–136.
- Kane, Thomas J., and Douglas O. Staiger.** 2008. “Estimating teacher impacts on student achievement: an experimental evaluation.” NBER Working Paper Series 14607.
- Keller, Matthew C., Barbara L. Fredrickson, Oscar Ybarra, Stéphane Coté, Kareem Johnson, Joe Mikels, Anne Conway, and Tor Wager.** 2005. “A Warm Heart and a Clear Head. The Contingent Effects of Weather on Mood and Cognition.” *Psychological Science*, 16(9): 724–731.
- Krautmann, Antony C., and William Sander.** 1999. “Grades and student evaluations of teachers.” *Economics of Education Review*, 18: 59–63.
- Krueger, Alan B.** 1999. “Experimental estimates of education production functions.” *Quarterly Journal of Economics*, 114: 497–532.
- Lavy, Victor.** 2009. “Performance Pay and Teachers’ Effort, Productivity and Grading Ethics.” *The American Economic Review*, 95(5): 1979–2011.
- Mullis, Ina V.S., Michael O. Martin, David F. Robitaille, and Pierre Foy.** 2009. *TIMSS Advanced 2008 International Report*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.
- OECD.** 2008. *Education at a Glance*. Paris:Organization of Economic Cooperation and Development.
- OECD.** 2010. *PISA 2009 at a Glance*. OECD Publishing.
- Pray, Marie Connolly.** 2011. “Some Like It Mild and Not Too Wet: the Influence of Weather on Subjective Well-Being.” CIRPEE Cahiers de recherche 1116.
- Prendergast, Candice, and Robert H. Topel.** 1996. “Favoritism in organizations.” *Journal of Political Economy*, 104(5): 958–978.
- Rivkin, Steven G., Eric A. Hanushek, and John F. Kain.** 2005. “Teachers, Schools and Academic Achievement.” *Econometrica*, 73(2): 417–458.
- Rockoff, Jonah 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, Jonah E., and Cecilia Speroni.** 2010. "Subjective and Objective Evaluations of Teacher Effectiveness." *American Economic Review (Papers and Proceedings)*, 100(2): 261–266.
- Rothstein, Jesse.** 2009. "Student sorting and bias in value added estimation: selection on observables and unobservables." *Education Finance and Policy*, 4(4): 537–571.
- Rothstein, Jesse.** 2010. "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement." *Quarterly Journal of Economics*, 125(1): 175–214.
- Schwarz, Norbert, and Gerald L. Clore.** 1983. "Mood, Misattribution, and Judgments of Well-being: Informative and Directive Functions of Affective States." *Journal of Personality and Social Psychology*, 45(3): 513–523.
- Swamy, P. A. V. B., and S. S. Arora.** 1972. "The Exact Finite Sample Properties of the Estimators of Coefficients in the Error Components Regression Models." *Econometrica*, 40(2): pp. 261–275.
- Tyler, John H., Eric S. Taylor, Thomas J. Kane, and Amy L. Wooten.** 2010. "Using student performance data to identify effective classroom practices." *American Economic Review (Papers and Proceedings)*, 100(2): 256–260.
- Weinberg, Bruce A., Belton M. Fleisher, and Masanori Hashimoto.** 2009. "Evaluating Teaching in Higher Education." *Journal of Economic Education*, 40(3): 227–261.

Figures

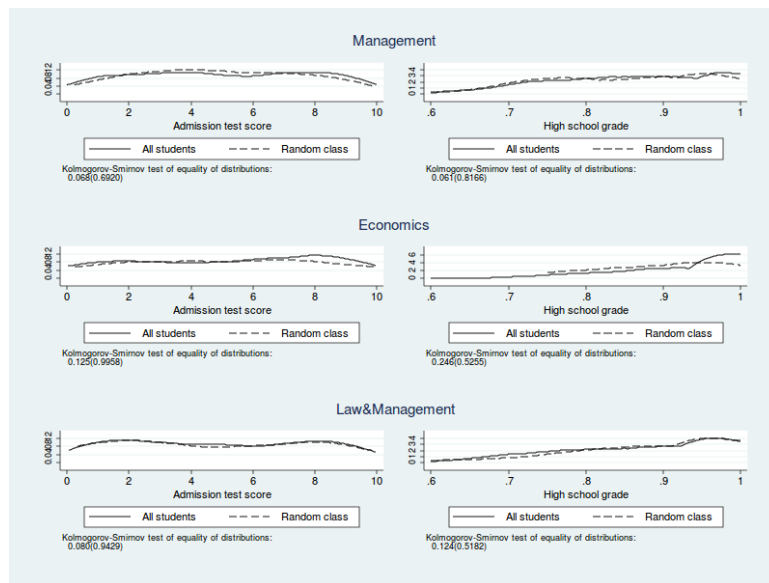


Figure 1: Evidence of random allocation - Ability variables

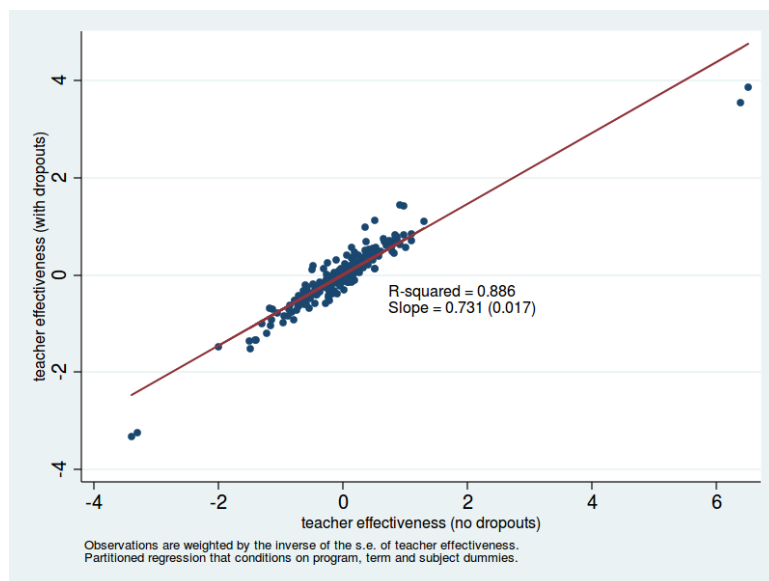


Figure 2: Robustness check for dropouts

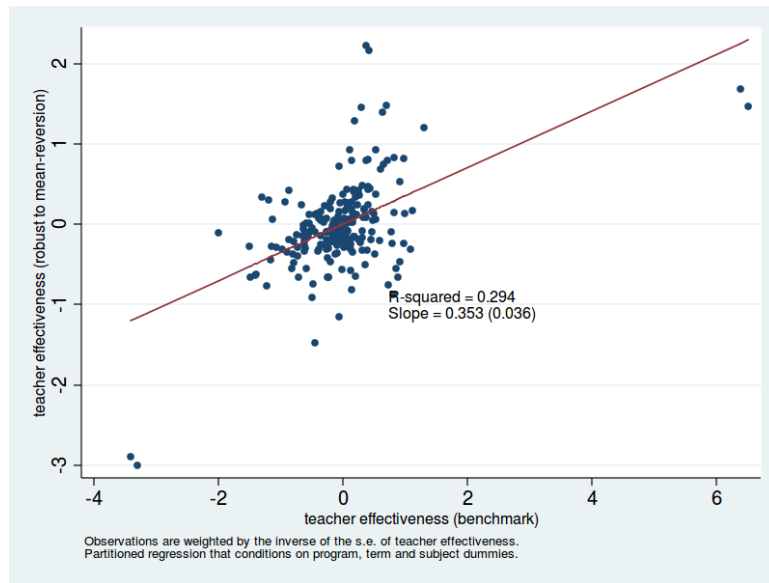


Figure 3: Robustness check for mean reversion in grades

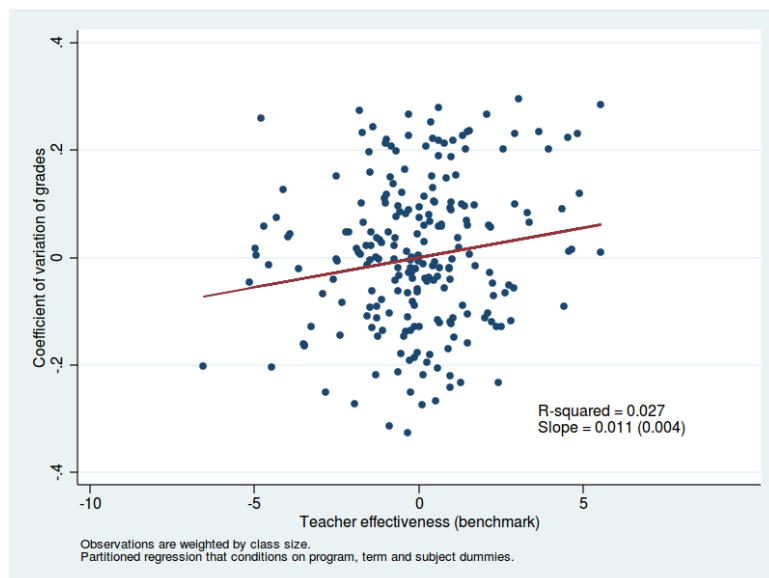


Figure 4: Teacher effectiveness and grade dispersion

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

Variable	Term							
	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	138.62	117.52	133.48	75.08	-
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	98.25	131.00	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	104.40	139.20	139.20	116.00	116.00	-	-	174.00
SD Class Size	39.11	47.65	47.67	44.96	50.47	-	-	0.00

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 ^a	0.620	0.748	0.621	0.634
1=top Income Bracket ^b	0.239	0.153	0.368	0.248
1=academic high school ^c	0.779	0.794	0.684	0.767
1=late enrollee ^d	0.014	0.015	0.011	0.014
High-school grade (0-100)	86.152 (10.905)	93.053 (8.878)	88.084 (10.852)	87.181 (10.904)
Entry Test Score (0-100)	60.422 (13.069)	63.127 (15.096)	58.894 (12.262)	60.496 (13.224)
University Grades (0-30)	25.684 (3.382)	27.032 (2.938)	25.618 (3.473)	25.799 (3.379)
Wage (Euro) ^e	966.191 (260.145)	1,012.241 (265.089)	958.381 (198.437)	967.964 (250.367)
Number of students	901	131	174	1,206

^a Dummy equal to one if the student's place of residence at the time of first enrollment is outside the province of Milan (which is where Bocconi university is located).

^b Family income is recorded in brackets and the dummy is equal to one for students who report incomes in the top bracket, whose lower threshold is in the order of approximately 110,000 euros at current prices.

^c Dummy equal to one if the student attended a academic high school, such as a lyceum, rather than professional or vocational schools.

^d Dummy equal to one if the student enrolled at Bocconi after age 19.

^e Nominal value at current (2010) prices. Based on 391 observations for Management, 36 observations for Economics, 94 observations for Law&Management, i.e. 521 observations overall.

Table 4: Descriptive statistics of students' evaluations

Variable	Management mean (std.dev.)	Economics mean (std.dev.)	Law&Manag. mean (std.dev.)	Total mean (std.dev.)
Overall teaching quality ^a	7.103 (0.956)	7.161 (0.754)	6.999 (1.048)	7.115 (0.900)
Lecturing clarity ^b	3.772 (0.476)	3.810 (0.423)	3.683 (0.599)	3.779 (0.467)
Teacher generates interest ^a	6.800 (0.905)	6.981 (0.689)	6.915 (1.208)	6.864 (0.865)
Course logistic ^b	3.683 (0.306)	3.641 (0.266)	3.617 (0.441)	3.666 (0.303)
Course workload ^b	2.709 (0.461)	2.630 (0.542)	2.887 (0.518)	2.695 (0.493)
Questionnaires/students ^c	0.777 (0.377)	0.774 (0.411)	0.864 (0.310)	0.782 (0.383)

^a Scores range from 0 to 10.

^b Scores range from 1 to 5.

^c Number of collected valid questionnaires over the number of officially enrolled students.
See Table A-4 for the exact wording of the evaluation questions.

Table 5: Randomness checks - Students

	Female	Academic High School ^a	High School Grade	Entry Test Score	Top Income Bracket ^a	Outside Milan	Late Enrollees ^a
	[1]	[2]	[3]	[4]	[5]	[6]	[7]
<i>Management</i>							
<i>Test statistics:</i>	χ^2	χ^2	<i>F</i>	<i>F</i>	χ^2	χ^2	χ^2
mean	0.489	0.482	0.497	0.393	0.500	0.311	0.642
median	0.466	0.483	0.559	0.290	0.512	0.241	0.702
minimum	0.049	0.055	0.012	0.004	0.037	0.000	0.025
maximum	0.994	0.949	0.991	0.944	0.947	0.824	0.970
<i>P-value^b (total number of tests is 20)</i>							
<0.01	0	0	0	1	0	3	0
<0.05	1	0	1	1	2	6	1
<i>Economics</i>							
<i>Test statistics:</i>	χ^2	χ^2	<i>F</i>	<i>F</i>	χ^2	χ^2	χ^2
mean	0.376	0.662	0.323	0.499	0.634	0.632	0.846
median	0.292	0.715	0.241	0.601	0.616	0.643	0.911
minimum	0.006	0.077	0.000	0.011	0.280	0.228	0.355
maximum	0.950	0.993	0.918	0.989	0.989	0.944	0.991
<i>P-value^b (total number of tests is 11)</i>							
<0.01	1	0	2	0	0	0	0
<0.05	1	0	2	1	0	0	0
<i>Law & Management</i>							
<i>Test statistics:</i>	χ^2	χ^2	<i>F</i>	<i>F</i>	χ^2	χ^2	χ^2
mean	0.321	0.507	0.636	0.570	0.545	0.566	0.948
median	0.234	0.341	0.730	0.631	0.586	0.533	0.948
minimum	0.022	0.168	0.145	0.182	0.291	0.138	0.935
maximum	0.972	0.966	0.977	0.847	0.999	0.880	0.961
<i>P-value^b (total number of tests is 7)</i>							
<0.01	0	0	0	0	0	0	0
<0.05	2	0	0	0	0	0	0

The reported statistics are derived from probit (columns 1,2,5,6,7) or OLS (columns 3 and 4) regressions of the observable students' characteristics (by column) on class dummies for each course in each degree program that we consider (Management: 20 courses, 144 classes; Economics: 11 courses, 72 classes; Law & Management: 7 courses, 14 classes). The reported p-values refer to tests of the null hypothesis that the coefficients on all the class dummies in each model are all jointly equal to zero. The test statistics are either χ^2 (columns 1,2,5,6,7) or *F* (columns 3 and 4), with varying parameters depending on the model.

^a See notes to Table 3.

^b Number of courses for which the p-value of the test of joint significance of the class dummies is below 0.05 or 0.01.

Table 6: Randomness checks - Teachers

	F-test	P-value
Class size ^a	0.94	0.491
Attendance ^b	0.95	0.484
Avg. high school grade	0.73	0.678
Avg. entry test score	1.37	0.197
Share of females	1.05	0.398
Share of students from outside Milan ^c	0.25	0.987
Share of top-income students ^c	1.31	0.228
Share academic high school ^c	1.35	0.206
Share late enrollees ^c	0.82	0.597
Share of high ability ^d	0.69	0.716
Morning lectures ^e	5.24	0.000
Evening lectures ^f	1.97	0.039
Room's floor ^g	0.45	0.998
Room's building ^h	1.39	0.188

The reported statistics are derived from a system of 9 seemingly unrelated simultaneous equations, where each observation is a class in a compulsory course (184 observations in total). The dependent variables are 9 teachers' characteristics (age, gender, h-index, average citations per year and 4 dummies for academic positions) and the regressors are the class characteristics listed in the table. The reported statistics test the null hypothesis that the coefficients on each class characteristic are all jointly equal to zero in all the equations of the system. The last row tests the hypothesis that the coefficients on all regressors are all jointly zero in all equations. All tests are distributed according to a F-distribution with (9,1467) degrees of freedom, apart from the joint test in the last row, which has (108,1467) degrees of freedom.

^a Number of officially enrolled students.

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

^c See notes to Table 3.

^d Share of students in the top 25% of the entry test score distribution.

^e Share of lectures taught between 8.30 and 10.30 a.m.

^f Share of lectures taught between 4.30 and 6.30 p.m.

^g Test of the joint significance of 4 floor dummies.

^h Dummy for building A.

Table 7: Determinants of class effects

Dependent variable = $\hat{\alpha}_s$	[1]	[2] ^a	[3]
Class size ^b	-0.000** (0.000)	-	-0.000** (0.000)
Avg. HS grade	2.159** (1.039)	-	2.360** (1.070)
Avg. entry test score	-1.140 (1.392)	-	-1.530 (1.405)
Share of females	0.006 (0.237)	-	-0.094 (0.245)
Share from outside Milan	-0.080 (0.203)	-	-0.078 (0.201)
Share of top income ^b	-0.283 (0.271)	-	-0.331 (0.278)
Share from academic HS	0.059 (0.301)	-	-0.054 (0.313)
Share of late enrollees	-0.365 (0.827)	-	0.017 (0.843)
Share of high ability ^b	0.733* (0.394)	-	0.763* (0.390)
Morning lectures ^b	0.015 (0.037)	-	-0.015 (0.040)
Evening lectures ^b	-0.175 (0.452)	-	-0.170 (0.490)
1=coordinator	-	0.013 (0.038)	0.039 (0.041)
Male	-	-0.017 (0.024)	-0.014 (0.025)
Age	-	-0.013*** (0.005)	-0.013** (0.005)
Age squared	-	0.000** (0.000)	0.000* (0.000)
H-index	-	-0.008 (0.006)	-0.007 (0.006)
Citations per year	-	0.000 (0.001)	0.000 (0.001)
Full professor ^c		0.116* (0.066)	0.121* (0.072)
Associate professor ^c		0.113* (0.062)	0.118* (0.067)
Assistant professor ^c		0.109* (0.061)	0.123* (0.065)
Classroom characteristics ^d	yes	no	yes
Degree program dummies	yes	yes	yes
Subject area dummies	yes	yes	yes
Term dummies	yes	yes	yes
Partial R squared ^e	0.089	0.081	0.158
Observations	230	230	230

Observations are weighted by the inverse of the standard error of the estimated α 's. * p<0.1, ** p<0.05, ***p<0.01

^a Weighted averages of individual characteristics if there is more than one teacher per class.

^b See notes to Table 6.

^c All variables regarding the academic position refer to the main teacher of the class. The excluded dummy is a residual category (visiting prof., external experts, collaborators.)

^d Four floor dummies, one building dummy and a dummy for multi-classrooms classes.

^e R squared computed once program, term and subject fixed effects are partialled out.

Table 8: Descriptive statistics of estimated teacher effectiveness

	Management	Economics	Law & Management	Total
<i>PANEL A: Std. dev. of estimated teacher effect</i>				
mean	0.069	0.159	0.194	0.086
minimum	0.041	0.030	0.010	0.010
maximum	0.106	0.241	0.030	0.241
<i>PANEL B: Largest minus smallest class effect</i>				
mean	0.190	0.432	0.027	0.230
minimum	0.123	0.042	0.014	0.014
maximum	0.287	0.793	0.043	0.043
No. of courses	20	11	7	38
No. of classes	144	72	14	230

Teacher effectiveness is estimated by regressing the estimated class effects (α) on observable class and teacher's characteristics (see Table 7).

Table 9: Comparison of benchmark, subject and contemporaneous teacher effects

Dependent variable: Benchmark teacher effectiveness		
Subject	0.048** (0.023)	-
Contemp.	-	-0.096*** (0.019)
Program fixed effects	yes	yes
Term fixed effects	yes	yes
Subject fixed effects	yes	yes
Observations	212	230

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

Table 10: Teacher effectiveness and students' evaluations

	Teaching quality [1]	[2]	Lecturing clarity [3]	[4]	Teacher ability in generating interest [5]	[6]	Course logistics [7]	[8]	Course workload [9]	[10]
<i>Teacher effectiveness</i>										
Benchmark	-0.496** (0.236)	-	-0.249** (0.113)	-	-0.552** (0.226)	-	-0.124 (0.095)	-	-0.090 (0.104)	-
Contemporaneous	-	0.238*** (0.055)	-	0.116*** (0.029)	-	0.214*** (0.044)	-	0.078*** (0.019)	-	-0.007 (0.025)
Class characteristics	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Classroom characteristics	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Teacher's characteristics	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Degree program dummies	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Subject area dummies	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Term dummies	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Partial R2	0.019	0.078	0.020	0.075	0.037	0.098	0.013	0.087	0.006	0.001
Observations	230	230	230	230	230	230	230	230	230	230

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

Table 11: Robustness check for class switching

	Overall teaching quality		Lecturing clarity	
	[1]	[2]	[3]	[4]
<i>PANEL A: All courses</i>				
Benchmark teacher effects	-0.496** (0.236)	-	-0.249** (0.113)	-
Contemporaneous teacher effects	-	0.238*** (0.055)	-	0.116*** (0.029)
Observations	230	230	230	230
<i>PANEL B: Excluding most switched course</i>				
Benchmark teacher effects	-0.572** (0.267)	-	-0.261** (0.118)	-
Contemporaneous teacher effects	-	0.258*** (0.064)	-	0.121*** (0.030)
Observations	222	222	222	222
<i>PANEL C: Excluding most and second most switched course</i>				
Benchmark teacher effects	-0.505* (0.272)	-	-0.234* (0.128)	-
Contemporaneous teacher effects	-	0.233*** (0.062)	-	0.112*** (0.031)
Observations	214	214	214	214
<i>PANEL D: Excluding five most switched courses</i>				
Benchmark teacher effects	-0.579** (0.273)	-	-0.229* (0.122)	-
Contemporaneous teacher effects	-	0.154** (0.063)	-	0.065** (0.032)
Observations	176	176	176	176

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 12: Students' evaluations and weather conditions

	Teaching quality			Lecturing clarity			Teacher ability in generating interest			Course logistics			Course workload		
	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]					
Av. temperature	0.139* (0.074)	0.120 (0.084)	0.063* (0.036)	0.054 (0.038)	0.171*** (0.059)	0.146*** (0.054)	0.051** (0.020)	0.047** (0.019)	0.057* (0.031)	0.053* (0.029)					
1=rain	-0.882** (0.437)	-0.929** (0.417)	-0.293 (0.236)	-0.314 (0.215)	-0.653** (0.327)	-0.716** (0.287)	-0.338*** (0.104)	-0.348*** (0.108)	0.081 (0.109)	0.071 (0.128)					
1=fog	0.741** (0.373)	0.687* (0.377)	0.391** (0.191)	0.367** (0.170)	0.008 (0.251)	-0.063 (0.247)	0.303*** (0.085)	0.292*** (0.090)	-0.254*** (0.095)	-0.265*** (0.096)					
Teaching effectiveness	- (0.244)	-0.424* (0.244)	- (0.120)	-0.189 (0.120)	- (0.223)	-0.566** (0.223)	- (0.088)	-0.090 (0.088)	- (0.093)	-0.088 (0.093)					
Class characteristics	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes					
Classroom characteristics	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes					
Teacher's characteristics	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes					
Degree program dummies	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes					
Subject area dummies	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes					
Term dummies	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes					
Observations	230	230	230	230	230	230	230	230	230	230					

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

Table 13: Teacher effectiveness and students evaluations by share of high ability students

	Presence of high-ability students			
	all	>0.22 (top 75%)	>0.25 (top 50%)	>0.27 (top 25%)
	[1]	[2]	[3]	[4]
PANEL A: Overall teaching quality				
Teaching effectiveness	-0.496** (0.236)	-0.502* (0.310)	-0.543 (0.439)	-0.141*** (0.000)
PANEL B: Lecturing clarity				
Teaching effectiveness	-0.249** (0.113)	-0.240 (0.140)	-0.283 (0.191)	-0.116* (0.068)
Observations	230	171	114	56

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

DOCENTE - DIDATTICA - PROGRAMMI

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:

Molto negativo	Negativo	Neutro	Positivo	Molto positivo
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

2. Per il mio apprendimento, la forma espositiva e la chiarezza dei docenti sono stati un fattore:

Molto negativo	Negativo	Neutro	Positivo	Molto positivo
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

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

Molto negativo	Negativo	Neutro	Positivo	Molto positivo
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

3.b durante l'orario di ricevimento

Molto negativo	Negativo	Neutro	Positivo	Molto positivo
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

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

Molto negativo	Negativo	Neutro	Positivo	Molto positivo
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

4.b le esercitazioni

<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
-----------------------	-----------------------	-----------------------	-----------------------	-----------------------

4.c i casi

<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
-----------------------	-----------------------	-----------------------	-----------------------	-----------------------

4.d gli interventi esterni

<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
-----------------------	-----------------------	-----------------------	-----------------------	-----------------------

4.e le ricerche

<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
-----------------------	-----------------------	-----------------------	-----------------------	-----------------------

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

Eliminare	Ridurre	Ya bene	Ampliare	Ampliare molto
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

5.b lo spazio per le esercitazioni

<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
-----------------------	-----------------------	-----------------------	-----------------------	-----------------------

5.c lo spazio per i casi

<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
-----------------------	-----------------------	-----------------------	-----------------------	-----------------------

5.d lo spazio per gli interventi esterni

<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
-----------------------	-----------------------	-----------------------	-----------------------	-----------------------

5.e lo spazio per le ricerche

<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
-----------------------	-----------------------	-----------------------	-----------------------	-----------------------

6. Durante questo corso ho notato riprese, ripetizioni, approfondimenti, nuovi svolgimenti di temi già trattati in corsi dello stesso semestre o di semestri precedenti:

Mai	Occasionalmente	Spesso	Molto spesso	Continuamente
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

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

7. Tali ripetizioni, approfondimenti, etc., per il mio apprendimento sono stati un fattore:

Molto negativo	Negativo	Neutro	Positivo	Molto positivo
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Figure A-1: Excerpt of student questionnaire

Table A-1: Descriptive statistics of estimated class effects

	Management	Economics	Law & Management	Total
<i>Std. dev. of estimated class effects</i>				
mean	0.054	0.157	0.035	0.081
minimum	0.029	0.058	0.004	0.004
maximum	0.092	0.241	0.087	0.241
<i>Largest minus smallest class effect</i>				
mean	0.152	0.423	0.050	0.211
minimum	0.045	0.010	0.005	0.005
maximum	0.249	0.723	0.122	0.723
No. of courses	20	11	7	38
No. of classes	144	72	14	230

Table A-2: Descriptive statistics of *subject* teacher effectiveness

	Management	Economics	Law & Management	Total
<i>PANEL A: Std. dev. of estimated teacher effects</i>				
mean	0.095	0.244	0.099	0.140
minimum	0.055	0.049	0.018	0.018
maximum	0.163	0.342	0.194	0.342
<i>PANEL B: Largest minus smallest teacher effect</i>				
mean	0.266	0.733	0.140	0.377
minimum	0.175	0.069	0.026	0.026
maximum	0.428	1.171	0.275	1.171
No. of courses	17	10	7	34
No. of classes	128	70	14	212

Table A-3: Descriptive statistics of *contemporaneous* teacher effectiveness

	Management	Economics	Law & Management	Total
<i>PANEL A: Std. dev. of estimated teacher effects</i>				
mean	0.200	0.310	0.163	0.225
minimum	0.094	0.150	0.001	0.001
maximum	0.351	0.507	0.468	0.507
<i>PANEL B: Largest minus smallest teacher effect</i>				
mean	0.553	0.819	0.231	0.571
minimum	0.133	0.213	0.001	0.001
maximum	1.041	1.626	0.661	1.626
No. of courses	20	11	7	38
No. of classes	144	72	14	230

Table A-4: Wording of the evaluation questions

Overall teaching quality	<i>On a scale 0 to 10, provide your overall evaluation of the course you attended in terms of quality of the teaching.</i>
Clarity of the lectures	<i>On a scale 1 to 5, where 1 means complete disagreement and 5 complete agreement, indicate to what extent you agree with the following statement: the speech and the language of the teacher during the lectures are clear and easily understandable.</i>
Ability in generating interest for the subject	<i>On a scale 0 to 10, provide your overall evaluation about the teacher's ability in generating interest for the subject</i>
Logistics of the course	<i>On a scale 1 to 5, where 1 means complete disagreement and 5 complete agreement, indicate to what extent you agree with the following statement: the course has been carried out coherently with the objectives, the content and the schedule that were communicated to us at the beginning of the course by the teacher.</i>
Workload of the course	<i>On a scale 1 to 5, where 1 means complete disagreement and 5 complete agreement, indicate to what extent you agree with the following statement: the amount of study materials required for the preparation of the exam has been realistically adequate to the objective of learning and sitting the exams of all courses of the term.</i>

Table A-5: Correlations between evaluations items

	Overall teaching quality	Lecturing clarity	Teacher generates interest	Course logistics	Course workload
Overall teaching quality	1.000	-	-	-	-
Lecturing clarity	0.888 (0.000)	1.000	-	-	-
Teacher generates interest	0.697 (0.000)	0.536 (0.000)	1.000	-	-
Course logistics	0.742 (0.000)	0.698 (0.000)	0.506 (0.000)	1.000	-
Course workload	0.124 (0.060)	0.122 (0.064)	0.193 (0.003)	0.094 (0.153)	1.000

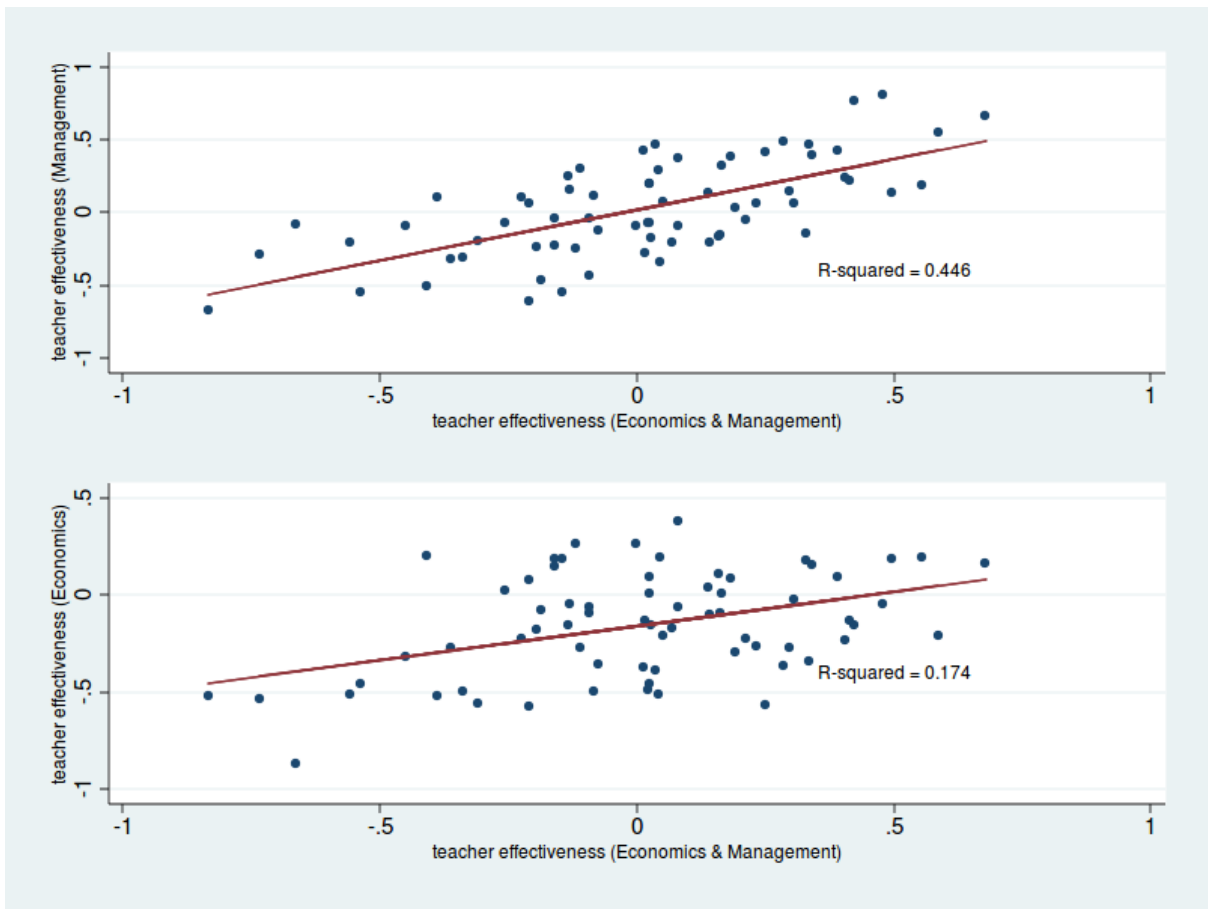


Figure A-2: Economics and Management common courses - Benchmark teacher effectiveness