DISCUSSION PAPER SERIES

No. 10283

THE ACADEMIC AND LABOR MARKET RETURNS OF UNIVERSITY PROFESSORS

Michela Braga, Marco Paccagnella
and Michele Pellizzari

LABOUR ECONOMICS

Centre for Economic Policy Research

ISSN 0265-8003

THE ACADEMIC AND LABOR MARKET RETURNS OF UNIVERSITY PROFESSORS

Michela Braga, Marco Paccagnella and Michele Pellizzari

Discussion Paper No. 10283
December 2014
Submitted 28 November 2014

Centre for Economic Policy Research
77 Bastwick Street, London EC1V 3PZ, UK
Tel: (44 20) 7183 8801
www.cepr.org

THE ACADEMIC AND LABOR MARKET RETURNS OF UNIVERSITY PROFESSORS[†]

Abstract

This paper estimates the impact of college teaching on students' academic achievement and labor market outcomes using administrative data from Bocconi University matched with Italian tax records. The estimation exploits the random allocation of students to teachers in a fixed sequence of compulsory courses. We find that the academic and labor market returns of teachers are only mildly positively correlated and that the professors who are best at improving the academic achievement of their best students are not always also the ones who boost their earnings the most, especially for the least able students.

JEL Classification: I20 and M55
Keywords: higher education and teacher quality

Michela Braga michela.braga@unibocconi.it
Università Bocconi

Marco Paccagnella marco.paccagnella@oecd.org
Banca d'Italia and OECD

Michele Pellizzari michele.pellizzari@unige.ch
University of Geneva, CEPR and IZA

[†] We would like to thank Bocconi University for granting us 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 Antonio Accetturo, Tito Boeri, Giacomo De Giorgi, Davide Dottori, Marco Leonardi, Tommaso Monacelli, Tommy Murphy, Tommaso Nannicini, and Paolo Sestito for their precious comments. 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 OECD or of the Bank of Italy. The usual disclaimer applies. Michele Pellizzari is also affiliated to CREAM, NCCR-LIVES and the Fondazione Rodolfo Debenedetti. Michela Braga and Marco Paccagnella are also affiliated to the Fondazione Rodolfo Debenedetti.

1 Introduction

The growing literature on the role of teachers in the education process has now firmly established that instructors are one of the most important determinants of student achievement (Steven G. Rivkin, Eric A. Hanushek & John F. Kain 2005, Scott E. Carrell & James E. West 2010). However, this result is based almost exclusively on studies using school performance as a measure of achievement while other indicators of success, namely employment or earnings, have been largely overlooked. In certain contexts or in some education institutions it might be preferable to hire and promote the teachers who are best at improving the labor market performance of their students rather than their school achievement but, while academic and labor market success are undoubtedly linked to each other, they do not overlap perfectly. Unfortunately, due to the scarcity of data linking professors, students and labor market outcomes, only a handful of very recent papers have been able to investigate the effect of teachers on labor market performance (Raj Chetty, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach & Danny Yagan 2011, Raj Chetty, John N. Friedman & Jonah E. Rockoff 2011, Giacomo De Giorgi, Michele Pellizzari & William G. Woolston 2012, Christian Dustmann, Patrick A. Puhani & Uta Schönberg 2012).

In this paper we estimate and compare measures of the academic and labor market returns of university professors using administrative data from Bocconi University. We construct such measures by comparing the future performance, either in subsequent coursework or in the labor market, of students who are randomly assigned to different teachers in a fixed sequence of compulsory courses. For this exercise we use administrative data that allow following students throughout their academic careers and into the labor market. The data are exceptionally rich in terms of observable characteristics of the students, the professors and the classes. For example, a very good measure of ex-ante ability is available for all the students in our data, namely their scores in an attitudinal test taken as part of the admission process. Hence, we can purge our measures of teacher's quality of most potential confounding factors and we can also document how they vary with the observables of both the students and the professors.

We find that the academic and labor market returns of teachers are indeed positively correlated but the magnitude of such a correlation is surprisingly small, around 0.4. Variation in academic returns only explains about 15% of the variation in labor market returns. Only about 35% of teachers who are in the top 20% of the distribution of academic returns also appear in the top 20% of the distribution of market returns and a good 13% are in the bottom quintile. We also split the sample of students by levels of ability and we estimate the impact of teachers on the performance of their best and worse students separately. We find that the correlation between the academic and labor market returns of professors is positive for high-ability students and negative for low-ability ones. We also show that professors who are good at teaching high-ability students are often not the best teachers for the least able ones. These results are consistent with the view that teaching is a multidimensional activity involving multiple tasks

each having different returns in the academia and in the labor market (C. Kirabo Jackson 2012).

Our paper contributes to the burgeoning literature on the role of teachers, improving over existing studies in a number of dimensions. Our focus on higher education is only one of the factors differentiating our work from the paper by Raj Chetty, John N. Friedman & Jonah E. Rockoff (2014*b*), which is probably the closest to ours in the literature. Due to the non-random allocation of teachers to pupils in their setting (3rd-8th graders in the US), Chetty, Friedman & Rockoff (2014*b*) cannot separately estimate the effect of professors on school and labor market performance. They indirectly estimate their correlation by regressing students' labor market outcomes on standard measures of teacher value-added. By exploiting the specificities of the process of class formation at Bocconi University, where students are randomly allocated to instructors for each compulsory course, we are able to produce separate estimates of the academic and labor market effects of each professor. We can then look at the joint distribution of these estimates, something that, to our knowledge, has never been done before in the literature. On the other hand, contrary to Chetty, Friedman & Rockoff (2014*b*), we cannot look at long-term labor market performance, since we only observe taxable income at one point in time, for most students around one to two years after graduation.

Jackson (2012) is also closely related to our work, particularly the theoretical framework where teaching is viewed as a multitasking activity with each task having different returns on different outcomes. However, we depart from Jackson (2012) in two respects. First, we focus on university professors rather than school teachers. Second, we consider academic and labor market returns separately, whereas Jackson (2012) contrasts alternative academic outcomes (cognitive test-scores) and longer-run non-cognitive indicators (disciplinary sanctions and intentions to enroll in college).

Most of the other papers focusing on higher education are concerned with the relative performance of professors on different academic tracks (Carrell & West 2010, Eric P. Bettinger & Bridget Terry Long 2010, Ronald G. Ehrenberg & Liang Zhang 2005, David N. Figlio, Morton O. Schapiro & Kevin B. Soter 2013). Carrell & West (2010) adopt our same methodology to compute measures of teacher effectiveness based on students' future outcomes. Their analysis, however, is limited to academic outcomes, whereas we extend it to earnings. Both we and Carrell & West (2010) depart from the most popular approach to measure teacher quality, the *value added model* (VA), which rests on the comparison of students' performance in standardized tests between two (or more) grades (Eric A. Hanushek 1971, Jonah E. Rockoff 2004, Rivkin, Hanushek & Kain 2005, Thomas J. Kane & Douglas O. Staiger 2008, Daniel Aaronson, Lisa Barrow & William Sander 2007). The VA model is commonly used in the context of primary and secondary education but it cannot be easily extended to college education, where there is no obvious definition of a grade and where syllabuses vary enormously.

The need to develop better measures of the quality of college teaching is further emphasized by a series of studies that cast serious doubts about the validity of student-reported evaluations, which are currently used in most universities around the world (William E. Becker & Michael

Watts 1999, Byron W. Brown & Daniel H. Saks 1987). For example, in a previous paper using the same data (Michela Braga, Marco Paccagnella & Michele Pellizzari 2014) we show that the teachers whose students perform better in subsequent coursework often receive worst evaluations, a finding that confirms the results of Carrell & West (2010).

Measuring teacher quality in any school level is both extremely important and extremely difficult. On the one hand, there is now ample evidence that teachers matter substantially for students' performance (Rivkin, Hanushek & Kain 2005, Carrell & West 2010). At the same time, the most common observable teachers' characteristics, such as their qualifications or experience, appear to be only mildly correlated with students' scores (Alan B. Krueger 1999, Rivkin, Hanushek & Kain 2005, Eric A. Hanushek & Steven G. Rivkin 2006), which makes it is difficult to identify good teachers ex-ante. In this setting contingent contracts based on ex-post outcomes would be the most obvious alternative to address the agency problem but writing such contracts requires appropriate measures of performance.

For this reason, value-added indicators of teacher quality have become popular in many countries. In the United States several studies advocated their use in hiring and promotion decisions (Chetty, Friedman & Rockoff 2011, Eric A. Hanushek 2009, Robert Gordon, Thomas J. Kane & Douglas O. Staiger 2006) and a few school districts have recently adopted such a practice. Despite their popularity, the validity of the VA approach has been questioned on various grounds. A common critique is based on the idea that teaching and learning are multidimensional processes and that test scores only capture a very limited set of abilities (Jackson 2012). This paper importantly contributes to this debate by showing that the academic and labor market returns of good teaching are indeed not always aligned.

From the policy viewpoint, our results suggest that performance measurement should be closely linked to the definition of the objective function of the education institution. Some schools or universities may see themselves as elite institutions and consequently target the very best students and the teachers who are best at maximizing the performance of such students. Some institutions may be more academic oriented and aim at transmitting the body of knowledge of one or several disciplines, regardless of the current market value of such knowledge. Other schools may take a more pragmatic approach and choose to teach the competencies most valued by the labor market at a specific point in time and space. The differences between community colleges and universities in the US or the dual systems of academic and vocational education that are common in countries like Germany, Switzerland or Australia are very good examples of teaching institutions with different objective functions, which should coherently adapt the way in which they evaluate their teachers.

The paper is organized as follows. Section 2 describes our data and the institutional details of Bocconi University. Section 3 discusses our strategy to estimate the academic and the labor market returns of professors. In Section 4 we present the main empirical results and we compare estimates produced using different outcomes (grades and earnings). Robustness checks are discussed in Section 4.1. Section 5 concludes.

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.[1] We select the cohort of students who enrolled as freshmen in the 1998/1999 academic year, as this is the only cohort in our data whose students were randomly allocated to teaching classes for each of their compulsory courses.[2] In later cohorts, the allocation system was modified and cannot be used for the estimation of teacher quality; however, we will use one of the later cohorts for an important robustness checks in Section 4.1. For earlier cohorts the class identifiers, which are a 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 seven different degree programs but we only consider the three of them that attracted a sufficient number of students to require splitting the lectures into multiple classes, namely Management, Economics and Law&Management.[3] Students in these programs were required to take a fixed sequence of compulsory courses spanning 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 A-1 in the Appendix 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 (management, economics, quantitative subjects, law).[4][5] In our analysis we consider only compulsory courses and we exclude elective subjects to avoid the complications due to the endogenous self-selection of the students.

Most but not all of the courses were taught in multiple classes. The number of classes into which students were split varied both across degree programs and across courses. For example, Management was the program that attracted the most students (over 70% in our cohort) and they 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, sometimes to a single one. The number of classes also varied within degree program depending on the number of

[1] This section borrows heavily from Braga, Paccagnella & Pellizzari (2014)

[2] The terms *class* and *lecture* often have different meanings in different countries and sometimes also in different schools within the same country. In most British universities, for example, *lecture* indicates a teaching session where an instructor - typically a full faculty member - presents the main material of the course; *classes* are instead practical sessions where a teacher assistant solves problem sets and applied exercises with the students. At Bocconi there was no such distinction, meaning that the same randomly allocated groups were kept for both regular lectures and applied classes. Hence, in the remainder of the paper we use the two terms interchangeably.

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

[4] Subject areas are defined according to the departments that were responsible for organizing and teaching the courses.

[5] Notice that Economics and Management share exactly the same sequence of compulsory courses in the first three terms. Indeed, students in these two programs did attend these courses together and made a final decision about their major at the end of the third term. Giacomo De Giorgi, Michele Pellizzari & Silvia 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 separate, but our results are robust to this assumption.

available teachers for each subject.

Each class was taught by one or, in a few cases, two professors. Within each course, all classes were required to follow exactly the same syllabus.[6] The exam questions were also the same for all students in the same course (and 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, deciding changes in the syllabus and addressing specific problems that might arose. 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.[7]

Our data cover in details the entire academic histories of the students, 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) and the grades in each single exam they sat at Bocconi. Graduation marks are observed for all non-dropout students.[8] Moreover, a proxy for family income is also recorded in our dataset because tuition fees depend on it. Finally, we have access to the random class identifiers that allow us to know in which class each students attended each of their courses.

A major advantage of our data is the availability of a very good measure of student ability, namely the scores obtained in a cognitive test that all students take as part of their application to the university. This test exclusively aimed at measuring cognitive ability, rather than knowledge of subjects related to economics or management and included sections on reading comprehension, verbal and visual relations, verbal patterns, mathematical reasoning. It was intended to provide a general measure of ability.

[INSERT TABLE 1 ABOUT HERE]

Table 1 reports descriptive statistics for the students in our data by degree program. The majority of them were enrolled in the Management program (74%), while Economics and Law&Management attracted 11% and 14%, respectively. Female students were generally slightly under-represented in the student body (43% overall), apart from the degree program

[6]Some variations across degree programs were allowed. For example, mathematics was taught slightly more formally to students in Economics than in Law&Management.

[7]Unfortunately, we do not know whether the coordinators ever actually intervened to adjust the distribution of grades across the classes of their courses. However, we discussed the issue with a number of Bocconi administrators and professors who were involved in the teaching and organization of courses for our cohort. According to them this instance was extremely rare and none of them recalls it happening during the years covered by our data.

[8]The dropout rate, defined as the number of students who have not graduated over the total size of the entering cohort, is around 4%. Notice that some of these students might have transferred to another university or still be working towards the completion of their program, whose formal duration was 4 years. In Section 4.1 we perform robustness checks showing that excluding the dropouts from our calculations is irrelevant for our results.

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 (gross) 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 finding that is also confirmed by university grades.

Data on earnings are obtained from tax records. We were able to merge the Bocconi data with the universe of all tax declarations submitted in Italy in 2005 (incomes earned in 2004). Since over 85% of the students in our sample graduated before May 2004, our data can be considered as a measure of initial earnings.[9] Unfortunately, only the 2004 tax declarations are currently available for research purposes and, thanks to a special agreement with Bocconi university, we have been able to merge them to the administrative records of the students. Of the 1,206 students in our sample 1,074 submitted a tax declaration in Italy in 2005, corresponding to approximately 90%. The others are likely to be either still looking for a job, or working abroad, or being out of the labor force (possibly enrolled in some post-graduate programme).[10] In our main analysis we will maintain the assumption that the students observed in the tax files are a random sub-group of the entire cohort and in Section 4.1 we present a series of robustness checks to support such an assumption.

2.1 The random allocation

In this section we present evidence that the random allocation of students into classes was successful.[11] The randomization was (and still is) performed via a simple random algorithm assigning a class identifier to all the students, who were then instructed to attend the lectures for the specific course in the class labeled with the same identifier.[12] 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 2 ABOUT HERE]

[9] Taxable income includes all earnings from employment, be it dependent or self-employment, as well as other incomes from properties (rents). Capital incomes are taxed separately and do not count towards personal taxable income.

[10] Bocconi also runs regular surveys of all alumni approximately 1 to 1.5 years since graduation and these surveys include questions on entry wages. About 60% of the students in our cohort answer the survey, a relatively good response rate for surveys, but still substantially lower than the matching we obtain with the tax records. In the subset of students that appear in both datasets, the two measures are highly correlated.

[11] De Giorgi, Pellizzari & Redaelli (2010) use data for the same cohort (although for a smaller set of courses and programs) and provide similar evidence.

[12] In fact, the allocation is not purely random as the algorithm is designed to avoid assigning too many students to certain classes and too few to others. The probability of being allocated to a given class varies with the relative number of students who were previously assigned to the class. However, the probability of being assigned to any class is never zero nor one.

Table 2 presents evidence that the students' observable characteristics are balanced across classes. More specifically, it reports 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. Hence, for each characteristic there are 20 such tests for the degree program in Management, corresponding to the 20 compulsory courses that were taught in multiple classes, 11 tests for Economics and 7 tests for Law&Management. The null hypothesis under consideration is that the coefficients on the class dummies in each model are jointly equal to zero, which amounts to testing for the equality of the means of the observable variables across classes (within courses and degree programs). The table shows descriptive statistics of the distribution of p-values for such tests.

The mean and median p-values are in all cases far from the conventional thresholds for rejection. Furthermore, the table also reports the number of tests that reject the null at the 1% and 5% levels, showing that this happens only in a very limited number of cases. The most notable exception is residence outside Milan, which is abnormally low in two Management groups. Overall, Table 2 suggests that the randomization was successful.

[INSERT FIGURE 1 ABOUT HERE]

In Figure 1 we further compare the distributions of our measures of ability (high school grades and entry test scores) for the entire student body and for one 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 enrolled, so that even if teachers were allowed to choose their class identifiers it would not be possible for them to know in advance the characteristics of the students who would be given their same identifiers.

More specifically, the matching of professors to class identifiers was (and still is) highly persistent and, if nothing changed from one academic year to the next, professors kept the same identifiers over time. Only when some teachers needed to be replaced or the overall number of classes changed, some modifications were implemented. 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 identifier and all the others kept their old ones. Similarly, if the total number of classes needed to be increases, the new classes would be added at the bottom of the list with new teachers and no change would affect the existing classes and professors.[13]

At about the same time when teachers were given class identifiers, also classrooms and time schedules were finalized. On these two items, though, teachers did have some limited

[13] As far as we know, the number of classes for a course has never been cut down.

choice. Typically, the administration suggested a time schedule and a classroom and professors could, with proper justifications, request modifications. Such demands were accommodated only when compatible with the overall teaching schedule (e.g. a room of the required size was available at the required time).

In order to avoid distortions 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. Additionally, we also know in which exact room each class was taught and we further condition on the characteristics of the classrooms, namely the buildings and the floors where they were 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).[14]

Table 3 provides evidence of the lack of correlation between teachers' and classes' characteristics by showing 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.[15] 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.

[INSERT TABLE 3 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. The test in the last row of Table 3 is useful to address concerns related to the tracking of teachers (C. Kirabo Jackson 2014), i.e. a situation in which the students of a given teacher are systematically assigned to the same or similar teachers later on. The process of repeated random allocations adopted in our setting already protects us against this concern but, in order to provide further evidence of lack of tracking, we included in the set of regressors of the simultaneous equations in Table 3 the average teacher effects of previous teachers for each class. These are the follow-on class effects that we estimate in Section 3.[16] Results show that the null of lack of correlation between past teacher quality and current teacher characteristics cannot be rejected.

[14] In principle we could also condition on room fixed effects but there are several rooms in which only one class was taught.

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

[16] For this exercise we use the follow-on academic returns of teachers but results are qualitatively similar when using any of the other measures that we construct (contemporaneous returns or labor market returns).

3 Estimating the academic and labor market returns of university professors

We use performance data for our students to measure the returns to university teaching and we do so separately for academic and labor market performance. Namely, for each compulsory course we compare the subsequent outcomes of students attending different classes, under the assumption that students who were taught by better professors enjoyed better outcomes later on. When computing the academic returns we consider the grades obtained by the students in all future compulsory courses and we look at their earnings when computing the labor market returns to teaching.

This approach is similar to the *value-added* methodology that is commonly used in primary and secondary schools (Dan Goldhaber & Michael Hansen 2010, Eric A. Hanushek & Steven G. Rivkin 2010, Hanushek & Rivkin 2006, Jesse Rothstein 2009, Rivkin, Hanushek & Kain 2005, Eric A. Hanushek 1979, Raj Chetty, John N. Friedman & Jonah E. Rockoff 2014*a*) but it departs from its standard version, that uses contemporaneous outcomes and conditions on past performance, by using future performance to infer current teaching quality. The use of future performance is meant to overcome potential distortions due to explicit or implicit collusion between the teachers and their current students. In higher education, this is a particularly serious concern given that professors are often evaluated through students' questionnaires, which have been shown to be poorly correlated with harder measures of teaching quality (Bruce A. Weinberg, Belton M. Fleisher & Masanori Hashimoto 2009, Carrell & West 2010, Antony C. Krautmann & William Sander 1999, Braga, Paccagnella & Pellizzari 2014). In fact, most of the papers that look at professors' quality in higher education use future rather than contemporaneous student performance for their calculations (Bettinger & Long 2010, Figlio, Schapiro & Soter 2013, Ehrenberg & Zhang 2005). However, for completeness and comparison with the rest of the literature we also compute teacher returns based on contemporaneous student performance.

The 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 often taught in one single class.

We compute the returns to teaching in two steps. For the sake of clarity, we first describe the computation of the academic returns based on future student performance and, then, we discuss how this procedure is adapted to compute labor market returns and academic returns

based on current performance. Our methodology is similar to Weinberg, Fleisher & Hashimoto (2009); in their setting, however, students are not randomly assigned to teachers.

In the first step, we estimate the conditional mean of the future grades of the students in each class according to the following procedure. Consider a set of students enrolled in degree program d and indexed by $i = 1, \ldots, N_d$, where N_d is the total number of students in the program. In our application there are three degree programs ($d = \{1, 2, 3\}$): Management, Economics and Law&Management. Each student i attends a fixed sequence of compulsory courses indexed by $c = 1, \ldots, C_d$, where C_d is the total number of such compulsory courses in degree program d. In each course c the student is randomly allocated to a class $s = 1, \ldots, S_c$, where S_c is the total number of classes in course c. Denote by $\zeta \in Z_c$ a generic (compulsory) course, different from c, which 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 $t \geq t_c$, excluding c itself.

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

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

where X_i is a vector of student characteristics including a gender dummy, the entry test score and the high school leaving grade, a dummy for whether the student is in the top income bracket and for whether he/she enrolled later than normal or resided outside the province of Milan (which is where Bocconi is located).[17] The α's are our parameters of interest and they measure the conditional means of future grades of the students in class s: high values of α_{dcs} indicate that, on average, students attending course c in class s performed better (in subsequent courses) than students in the same degree program d taking course c in a different class.

Thanks to the random allocation the class fixed effects α_{dcs} are exogenous in equation 1 and identification is straightforward. The normalization of the dependent variable (within courses) allows interpreting the class effects in terms of standard deviation changes in the outcome.

Notice that, since in general there are several subsequent courses ζ for each course c, each student in equation 1 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 and independent

[17] Students should normally enroll at university in September of the calendar year in which they turn 19 unless they entered school early or they were retained at some point of their school career.

components (all with mean equal zero):

$$\epsilon_{ids\zeta} = \upsilon_i + \omega_s + \nu_{ids\zeta} \tag{2}$$

where υ_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.[18] In the absence of other sources of serial correlation (e.g. if the variance of ω_s were zero), such a transformation would lead to a serially uncorrelated and homoskedastic variance-covariance matrix of the transformed 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.

The second step of our approach is meant to purge the estimated α's of the effect of other class characteristics that might affect the performance of students in later courses but are not necessarily 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 2012).

Assuming linearity, the estimated class effects can be written as follows:

$$\widehat{\alpha}_{dcs} = \gamma_0 + \gamma_1 T_{dcs} + \gamma_2 C_{dcs} + \tau_{dcs} + u_{dcs} \tag{3}$$

where τ_{dcs} is the unobservable quality of teaching and T_{dcs} and C_{dcs} are other teacher and class characteristics, respectively. γ_1 and γ_2 are fixed parameters and u_{dcs} is the estimation error.

A key advantage of our data is that most of the factors that can be be thought as being included in T_{dcs} and C_{dcs} are observable. In particular, we 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 the identity of the teachers who coordinated each course. Additionally, based on our academic records we can construct measures of both class size and class composition (in terms of students' characteristics). Hence, we can estimate τ_{dcs} as the OLS residuals of equation 3, since the estimation error u_{dcs} has zero mean and converges in probability to zero (given consistency of $\widehat{\alpha}_{dcs}$). Further, in equation 3 we weight the observations by the inverse of the standard error of the estimated α's to take into account differences in the precision of such estimates.

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

Obviously, we cannot be guaranteed to observe all the relevant variables in T_{dcs} and C_{dcs}, however, given the richness of our data, it should be uncontroversial that teaching quality is by far the single most important unobservable that generates variation in the estimated residuals.[19]

Compared to other papers using data where the same professors are observed teaching different cohorts of students over time (Chetty et al. 2011, Chetty, Friedman & Rockoff 2011), we cannot estimate separately teacher and class effects. In fact, all the student cohorts following the one that we consider in this study were randomly allocated to classes only once per academic year, so that students would take all the compulsory courses of each academic year with the same group of classmates. In such a setting, only the joint effect of the entire set of teachers in each academic year can be identified.

Hence, we exploit the rich set of observables in our data to purge the estimated class effects through our two-step procedure to obtain a statistics that can be interpreted as teaching quality or the returns to teaching.[20] We believe that this approach is appropriate in our context. First of all, our data are indeed extremely rich and include information on a number of features that are normally unobservable in other studies (Chetty, Friedman & Rockoff 2011, Kane & Staiger 2008, Daniel F. McCaffrey, Tim R. Sass, J. R. Lockwood & Kata Mihaly 2009). Moreover, we consider a single institution rather than all schools in an entire region or school district as in Chetty, Friedman & Rockoff (2011) or in Kane & Staiger (2008). As a consequence, variation in class and student characteristics is limited and very likely to be captured by our rich set of controls.

In fact, one may actually be worried that we purge of too many factors rather than too few, insofar as teaching quality is itself a function of some of the teachers' observables. For this reason, we present results conditioning on all the available class and teachers' characteristics as well as conditioning only on the class characteristics. Moreover, in Section 4.1 we further show that our measures of teaching quality can predict the outcomes of students taught by the same professors in later cohorts, thus supporting the intuition that our methodology captures teacher effects as separate from class effect.

While the OLS residuals of equation 3 are consistent estimates of the τ_{dcs}s, estimating their variance requires taking into account the variance of the estimation error u_{dcs}. For this purpose we follow again Weinberg, Fleisher & Hashimoto (2009), adopting a procedure that is similar to the shrinkage models commonly used in the literature (Kane & Staiger 2008, Gordon, Kane & Staiger 2006, Thomas J. Kane, Jonah E. Rockoff & Douglas O. Staiger 2008, Rockoff 2004) but that is adapted to our peculiar framework where teachers are observed teaching only one

[19] Social interactions among the students might also be part of equation 3. However, notice that if such effects are related to the observable characteristics of the students, then we are able to control for those (up to functional form variations). Additionally, there might be complementarities between teacher's ability and students' interactions, as good teachers are also those who stimulate fruitful 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.

[20] Notice additionally that in a 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 always attached to a specific person.

class.

We randomly split in half each class in our sample and we replicate our estimation procedure for each of them, so that for each class we have two estimates of τ, say $\widehat{\tau}'_{dcs}$ and $\widehat{\tau}''_{dcs}$. Since the only source of estimation error in our setting is unobservable idiosyncratic variation in student performance, the random split of the classes guarantees that the estimation errors in $\widehat{\tau}'_{dcs}$ and $\widehat{\tau}''_{dcs}$ are orthogonal to each other.[21] Hence, the variance of τ_{dcs} can be estimated as the covariance between $\widehat{\tau}'_{dcs}$ and $\widehat{\tau}''_{dcs}$:

$$\begin{aligned} Cov\left(\widehat{\tau}'_{dcs},\, \widehat{\tau}''_{dcs}\right) &\xrightarrow{p} Cov\left(\tau_{dcs} + u'_{dcs}, \tau_{dcs} + u''_{dcs}\right) \\ &= Var\left(\tau_{dcs}\right) + Cov\left(u'_{dcs}, u''_{dcs}\right) \\ &= Var\left(\tau_{dcs}\right) \end{aligned} \tag{4}$$

In our calculations approximately 60% of the uncorrected variance of the teacher effects is due to the estimation error, depending on the specification and the outcome measure (grades or earnings).

Some of the papers in this literature also use the estimated variance to adjust the teacher effects according to a Bayesian procedure that shrinks towards zero the least precise estimates (Carrell & West 2010). Given that we plan to use the $\widehat{\tau}$s in secondary regression analyses, we prefer to adjust only the variance and avoid further complications in the derivation of correct inference results for regressions where the teacher effects are used as dependent or independent variables.

To estimate the labor market returns of teaching, we follow the same procedure described above but we replace future exam grades with earnings as a dependent variable in equation 1. This simplifies the estimation substantially, since there is only one outcome per student and no need to account for serial correlation. We still cluster the standard errors at the level of the class.

In order to compare our results with the rest of the literature we also compute a measure of teaching quality using contemporaneous grades as outcomes. Specifically, we replicate the exact same procedure described above but we replace the students future grades with their grades in course c on the left hand side of equation 1. As in the case of earnings, there is only one contemporaneous grade per student per course and inference is simplified (clustering at the class level is always maintained).

Throughout the rest of the paper we will refer to the estimates of τ_{dcs} based on future exam grades simply as the *academic returns* to teaching, to those based on earnings as the *labor market returns* and to those based on contemporaneous grades as the *contemporaneous returns*.

[21] The existence of social interactions among the students may introduce correlation between the estimation errors across the random halves of the classes. Hence, the validity of this shrinkage procedure requires the additional assumption that any effect of social interactions among the students is captured by either the observable students' characteristics or the class effects.

Our methodology differs from the most common approaches in the literature in two ways: first, we estimate equation 1 separately for each course and, second, we run two subsequent regressions to purge the class effects of a large set of observable class characteristics. Most papers, especially those focusing on primary and secondary schools, usually run one pooled model and exploit repeated observations on teachers to tease out class shocks.

The estimation of the class effects separately for each course is motivated by the fact that, contrary to most other papers using future student performance to measure teacher quality (Bettinger & Long 2010, Ehrenberg & Zhang 2005, Figlio, Schapiro & Soter 2013), in our setting the dimension of the vectors of future grades varies across courses (i.e. the number of future grades is different for each course). Alternatively we could have considered only the average future grade for each student in each course, thus reducing the dimension of all such vectors to one, but at the cost of making less efficient use of the information in our data. We could have also derived a GMM specification to estimate all class effects jointly in a fully interacted model this approach could deliver efficiency gains but at the cost of a substantially more complicated estimation strategy. Of course the dimensionality of the vector of student performance is equal to one for all equations when we estimate teacher returns based on contemporaneous grades or earnings but, for the sake of simplicity and comparison, we prefer to adopt the same methodology for all three measures of teacher quality that we construct. The motivation for the two-step procedure is also practical. In principle we could have included the class and teacher characteristics already in equation 1 but, given the peculiarity of our setting where professors are observed teaching only one class, it is particularly important to show the estimates of equation 3.

Nevertheless, in order to show that our results do not depend on the particular methodology that we adopt, Figure A-1 in the Appendix compares our preferred estimates of the (follow-on) academic returns, namely the τ_{dcs}, with similar estimates produced with a simple random effect model. In scu model we pool all courses together, we use as outcome the average future grade of the student and we control for both the student's the class's and the teacher's observables. Two sets of estimates are extremely similar to each other.[22]

4 Empirical results

Overall, we are able to estimate both the academic and the labor market returns of 230 teachers. We cannot run equation 1 for courses that have no contemporaneous nor subsequent courses.[23] For such courses, the set Z_c is empty. Additionally, some courses in Economics and in Law&Management are taught in one single class.[24] For such courses, the computation

[22]For simplicity the figure only shows results for the classes of the Management program. In the pooled random effect model the coefficients on the explanatory variables are constrained to be the same across courses.

[23]For example, Corporate Strategy for Management, Banking for Economics and Business Law for Law&Management (see Table A-1).

[24]For example Econometrics for Economics students or Statistics for Law&Management (see Table A-1).

of the academic returns of teaching based on future exam grades is impossible since $S_c = 1$.

When we consider contemporaneous academic returns or labor market returns we do not face the above constraints, as incomes and contemporaneous grades are available for all students and all courses. In fact, we can estimate these effects for slightly more teachers (242 in total) but, given that our main purpose is the comparison of different types of returns, we prefer to focus on the subsample for which we can estimate all three indicators.

Table 4 shows test statistics from the second-step equation 3 for the three indicators of teaching quality that we construct. Given the large set of regressors used in these models (25 in total), we have grouped them into categories: three categories for the characteristics of the classes (class size, class composition and class time and room) and four categories for the characteristics of the teachers (coordinator, demographics, citations and academic rank). The exact variables included in each category are listed in the notes to the table. Two categories only include one variable, namely class size and coordinator (a dummy equal to one when the teacher of the class is also the coordinator of the course) because we deem these variables particularly interesting. For each category the table shows the F-test and p-values for the null of joint significance of the coefficients of all the variables in the category. The table also reports the partial R-squared obtained from a partitioned regression where the full set of dummies for degree program, term and subject area are partialled out.[25] All results are shown for three different specifications, one with only the class characteristics as explanatory variables, one with only the teacher characteristics and one with both.

[INSERT TABLE 4 ABOUT HERE]

Results indicate that all the explanatory variables included in these second-step regressions have very limited explanatory power. Although some individual regressors are statistically significant in some specifications, none of the reported F-test allows rejecting the null. This is consistent with both the random allocation and the common finding in the literature about the lack of correlation between student outcomes and teachers' observables (Hanushek & Rivkin 2006, Krueger 1999). Overall, the class characteristics alone explain slightly less than 10% of the variation in student future academic performance (panel A), slightly less than 3% of the variation in earnings (panel B) and slightly less that 6% of the variation in contemporaneous grades (panel C). Teacher observables have even less explanatory power: 3.6% for academic performance, 2.5% for earnings and about 1% for contemporaneous grades. Interestingly, professors who are more productive in research do not seem to be better at teaching.[26] When conditioning on both teacher and class characteristics the (partial) R-squared remain low.[27]

[25] The full results are available upon request.

[26] See Marta De Philippis (2013) for a formal evaluation of research incentives on teaching performance using our same data.

[27] The 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 area are partialled out.

Our final measure of the returns to teaching are the residuals of the regressions of the estimated αs on all the observable variables, i.e the regressions reported in columns 3 of Table 4. In Table 5 we present descriptive statistics of such measures. For completeness, the lower panel of the table (panel B) also reports the same results computed without conditioning on the teachers' observable characteristics (i.e. residuals of the regressions in columns 1 of Table 4).

[INSERT TABLE 5 ABOUT HERE]

The average standard deviation of the academic returns is 0.048.[28] As discussed in Section 3, this number can be readily interpreted in terms of standard deviations of the distribution of students' grades. In other words, assigning students to a teacher whose academic effectiveness is one standard deviation higher than their current professor would improve grades by 4.3% of a standard deviation, corresponding to approximately 0.6% over the average.

This effect is comparable to the findings in Carrell & West (2010), who estimate an increase in GPA of approximately 0.052 of a standard deviation for a one standard deviation increase in teaching quality. To further put the magnitude of our estimates into perspective, it is useful to also consider the effect of a reduction in class size, which has been estimated by numerous papers in the literature (Joshua D. Angrist & Victor Lavy 1999, Krueger 1999, Oriana Bandiera, Valentino Larcinese & Imran Rasul 2010) and also on the same data used for this study (De Giorgi, Pellizzari & Woolston 2012). The estimates in most of these papers are in the range of 0.1 to 0.15 of a standard deviation increase in achievement for a one standard deviation reduction in class size, thus about two to three times the effect of teachers that we estimate here. Notice, however, that one of the obvious mechanisms through which reducing the size of the class affects performance is the possibility for the professors to tailor their teaching styles to their students in small classes, that is an improvement in the quality of teaching. In other words, our estimates of teaching quality are computed holding constant the size of the class whereas the usual class size effect allows the quality of teaching to vary.

In Table 5 we also report the standard deviations of the academic returns to teachers in the courses with the least and the most variation. Overall, we find that in the course with the highest variation (macroeconomics in the Economics program) the standard deviation of our measure of academic teaching quality is 0.14, approximately 3.3 times the average. This compares to a standard deviation of essentially zero in the course with the lowest variation (accounting in the Law&Management program).

The second column in Table 5 reports similar statistics for the labor market returns of professors, measured by the conditional average earnings of one's randomly assigned students, as explained in Section 3. A one standard deviation better professor leads to an increase in earnings by 5.4% of a standard deviation on average. This translates in an annual increase of gross income of about 1,000 Euros, slightly more than 5.5% over the average. Also the labor market returns are vastly heterogeneous across subjects, with the variation reaching 18% of a standard

[28]The standard deviation is computed on the basis of the *shrinkage* method described in Section 3.

deviation in earnings for mathematics in the Economics program and being close to zero for management III in the Management program.

Given that most of the existing literature uses contemporaneous achievement to construct value-added measures of teacher quality, the third column of Table 5 shows results based on such student outcomes. We are very much in line with previous findings: most existing papers estimate effects in the order of 10% of a standard deviation for a 1 standard deviation change in teacher quality and we report an average effect ranging between 7.1% and 11.6% (when excluding teacher attributes from the set of controls).

In the lower panel (panel B) of Table 5 we report the same descriptive statistics for our measures of professors' quality that do not purge the effect of the observable characteristics of the teachers. Consistent with the finding that such characteristics bear little explanatory power for students' performances (see Table 4), the results in panels A and B of Table 5 are extremely similar.

By restricting the set of students to those of high or low ability, defined as those whose performance in the attitudinal entry test is above or below the median, it is possible to replicate the procedure described in Section 3 to produce professor effects for each of these two categories of students. The descriptive statistics of these indicators are reported in Table 6 and their analysis allows understanding whether it is the best or the worst students who benefit the most from good teachers and in what dimension.[29]

[INSERT TABLE 6 ABOUT HERE]

When considering academic performance the dispersion in teachers' returns appears to be rather homogeneous across student types, with an average standard deviation of about 0.066-0.067 in both cases (panel A) and similarly for the contemporaneous returns (panel C). Larger differences emerge when teaching quality is measured with students' earnings (panel B). In this case, the low-ability students seem to benefit from effective teaching substantially more than their high ability peers, the average standard deviations being 0.178 and 0.097, respectively. When looking at the minimum and maximum effects it appears that the entire distribution of labor market returns is shifted to the right for the low-ability students.

These results appear consistent with the view that teaching is a multidimensional activity as in Jackson (2012). Furthermore, the comparison between students of different ability levels suggests that the degree of complementarity between teacher and students skills varies both across teaching activities and student types.

One obvious question that one can ask with these data is whether the professors who are best at improving the academic performance of their students are also the ones who boost their earnings the most. In Table 7 we estimate the correlations between our alternative measures of the returns to teaching, conditional on degree program, term and subject area effects. In these

[29] Notice that the effects reported in Table 5 cannot be simply derived as averages of the effects for high- and low-ability students in Table 6.

regressions we weight each observation by the inverse of the standard error of the the dependent variable and we bootstrap the covariance matrix.

[INSERT TABLE 7 ABOUT HERE]

Results show a positive and significant correlation between the academic and the labor market returns of professors when using data on all the students in each class, a finding that is remarkably consistent with Chetty, Friedman & Rockoff (2014*a*) and Chetty, Friedman & Rockoff (2014*b*), despite the stark differences in the settings and the methodologies. Chetty, Friedman & Rockoff (2014*b*) estimate that a one standard deviation increase in teacher value-added is associated with an increase in total earnings (the long-run outcome that is more similar to our taxable income) of 353 USD per year, corresponding to approximately 0.015 of a standard deviation (see Table 3 on page 2655 in Chetty, Friedman & Rockoff (2014*b*)). Based on our estimates we can calculate a similar effect of the order of 0.018 of a standard deviation.[30]

Despite being significant, the magnitude of the correlation between the academic and labor market returns of the teacher is surprisingly low: a one standard deviation increase in the first measure is associated with approximately one tenth of a standard deviation increase in the academic returns of the same teacher. When the analysis is replicated for low- and high-ability students separately, the positive association of academic and labor market returns to teaching is confirmed for high ability students but it turns negative for the low ability ones.[31]

Consistent with previous findings (Braga, Paccagnella & Pellizzari 2014, Carrell & West 2010), we also document a negative correlation between the academic returns computed on the basis of contemporaneous and subsequent achievement. This result is most likely driven by grade leniency induced by the system of teacher evaluations based on student opinions, a system that is common in most universities around the world and that was (and still is) applied also at Bocconi. Coherent with this interpretation, we also find that the contemporaneous effects are negatively correlated with the labor market effects.

[INSERT TABLE 8 ABOUT HERE]

In Table 8 we also estimate the cross-correlation between the academic and labor market returns for high- and low-ability students. In other words, we ask whether the professors who are the most effective for the good students are so also for the least able ones. As for the results of Table 7, in these regressions we condition on degree program, term and subject areas fixed effect, we weight observations by the inverse of the standard error of the estimated dependent variable and we bootstrap the covariance matrix of the estimates. Interestingly, we find a very

[30] An increase of one standard deviation in our academic returns is associated to a 0.34 increase in the labor market returns (the reverse regression of the one reported in Table 7, column 1 in panel A) and a one standard deviation increase in labor market returns is associated to 0.055 standard deviation increase in taxable earnings (Table 5), hence: $0.34 \times 0.055 = 0.0187$.

[31] Notice that there is no sense in which the estimates in panel A of Table 7 can be seen as averages of those in panels B and C.

low and insignificant correlation for academic returns and a positive and significant one for the labor market returns of professors.

The findings in Tables 7 and 8 can be rationalized under the view that the complementarity between teacher and student abilities is stronger in the production of academic achievement than earnings (Jackson 2012). A very good student learns easily and does not need to be explained things several times nor particularly clearly. Low ability students, on the other hand, really need a good teacher to understand the material. Hence, it is easier to raise both grades and wages for the high ability students that the low ability ones.

[INSERT TABLE 9 ABOUT HERE]

These results suggest that teaching is better viewed as a multidimensional activity involving a variety of tasks each of which has potentially different returns on the academic and the labor market performance of the students. Hence, it is problematic to evaluate teachers on one single dimensions, as it is often done in practice. Table 9 further emphasizes this important point by showing the joint distribution (by quintiles) of the academic and labor market returns of professors. Despite the general positive correlation documented in Table 7, the two distributions overlap only very partially. Only abut one third of the professors are in the same quintile in both distributions and approximately 35% of teachers who are in the top 20% of the distribution of academic returns also appear in the top 20% of the distribution of market returns. A sizable fraction of them (approximately 13%) are in the bottom quintile. Consistent with our previous findings, the overlap is even less substantial if one compares the distribution of the contemporaneous returns with either of the other two measures. In fact, the academic returns of teaching explain only about 16% of the variation in the labor market returns, once program, course and area fixed effects are partialled out. If one were to combine both the academic and the contemporaneous effects to predict the labor market effects one would still be able to explain less than 17% of the (residual) variation.[32]

4.1 Robustness checks

In this section we provide a number of robustness checks for the main results of our analysis.

A first important concern is related to the distinction between class and teacher effects. Most of the existing papers on teaching quality use data where professors are observed teaching several classes of students allowing to separately identify class and teacher effects. Our approach is necessarily different because we can only use one cohort of students for our analysis - namely the 1998/1999 freshmen - and we only observe a fixed group of professors teaching the sequence of compulsory courses for this cohort. Hence, almost all professors are observed

[32]These calculations are based on the partial R-squared of weighted OLS regressions with labor market returns as depended variable and academic and contemporaneous returns (separately or jointly) as explanatory variables. These regressions are not reported for brevity but are available from the authors upon request.

teaching only once.[33] The reason why we cannot use other cohorts is twofold. First, the class identifiers were not recorded in the university archives prior to 1998, hence we cannot use older cohorts. Second, an important reform of the degree programs was implemented at Bocconi for the 1999/2000 freshmen and, among other changes, students started being randomly allocated into classes only once every academic year (rather than once for every compulsory course as in the previous cohorts). The new freshmen took all the courses of each academic year with the same random group of peers. In such a setting it is impossible to attribute the class effect to a specific teacher.[34] Hence, we cannot use the later cohorts either.

The second step of our empirical strategy (Table 4) is meant to address precisely this problem. First we produce estimates that combine both class and teacher effects (the α_{dcs}s). Then, we exploit the exceptional richness of our data and we purge these effects of the most obvious and important sources of class variation and we interpret the residuals as reflecting exclusively variation in teacher quality. Obviously, the validity of such an interpretation rests on the quality of the set of observable class characteristics and it might be questioned.

In order to provide evidence that our empirical strategy does capture variation in teacher quality separately from class shocks, we tracked the teachers of our cohort into the following one, namely the 1999/2000 freshmen. These students were offered a completely new and different set of degree programs with different sequences of compulsory courses and, most importantly, they were randomly allocated into classes only once per academic year. Despite all these differences we can still estimate the following equation:

$$y_{idsc} = \delta_0 + \delta_1 \widehat{\tau}_{dcs} + \delta_2 X_i + \delta_3 T_{dcs} + \delta_4 C_{dcs} + \epsilon_{idsc} \tag{5}$$

where y_{idsc} is the grade of student i in degree program d, class s and course c, $\widehat{\tau}_{dcs}$ is the return to teaching of the teacher of class s estimated from the previous cohort and X_i, T_{dcs} and C_{dcs} are vectors of individual, teacher and class characteristics containing the same variables as in Section 4.

Importantly, in the new setting the class is fixed in each academic year whereas the teachers are different for each course (both within and across academic years). This allows us to estimate also a version of equation 5 that includes class fixed effects:[35]

$$y_{idsc} = \widetilde{\delta}_0 + \widetilde{\delta}_1 \widehat{\tau}_{dcs} + \widetilde{\delta}_2 X_i + \widetilde{\delta}_3 T_{dcs} + \kappa_{dsc} + \widehat{\epsilon_{idsc}} \tag{6}$$

where κ_{dst} is a fixed effect for the class.[36]

[33] There are only 3 professors who teach two classes in the sequence, too few for any meaningful analysis.

[34] The students we use for our main analysis (the freshmen of 1998/1999) were unaffected by the reform throughout their entire university curriculum. The reform only applied to the new intake of students.

[35] The class here is intended primarily as the group of students who attend courses together but it also captures other features, such as the time schedule of the courses (which were the same for all students in a group) and the physical characteristics of the classrooms (which were often the same for all the courses of the academic year).

[36] The fixed effect would be the same for all the courses taught in the same academic year of course c.

In order to facilitate the interpretation of the results, we standardize both the dependent variable y_{idsc} and the teacher effect $\widehat{\tau}_{dcs}$ so that the parameters δ_1 and $\widetilde{\delta}_1$ indicate the standard deviation effect on the outcome of a one standard deviation change in the quality of teaching and they can be directly compared to the results in Tables 5 and 6.

[INSERT TABLE 10 ABOUT HERE]

Table 10 reports estimates of equations 5 and 6 for the students who enrolled in their first year at Bocconi in September 1999, also divided by ability groups (Panels B and C). For these regressions we only consider compulsory courses. We also restrict the sample only to the degree program in Management because it is the one that remained the most consistent across the two cohorts and it is also the one for which we can track the highest number of teachers. Of the 184 professors teaching in the classes of the Management program in the old cohort, 69 of them (37%) teach also in the new cohort in the same program. Finally, we restrict the analysis to only one measure of teaching quality, namely the contemporaneous academic returns, given that for both the other measures the class shocks cumulate over outcomes, making the interpretation of the results more complex.[37] The fact that the contemporaneous returns might be affected by teacher leniency is irrelevant for this robustness check. The particular setting of the new cohort allows us to also control for individual student effects. Table 10 reports results based on both random (columns 1 and 2) and fixed (columns 3 and 4) specifications of the student effects.

The first and most important result of Table 10 is that the coefficients on the professors' returns are positive and strongly statistically significant in all reported specifications. Interpreting their magnitudes is not straightforward. The reform that took place in 1999 changed various aspects of the teaching environment and it is not completely obvious that one should expect δ_1 and $\widetilde{\delta}_1$ to equal exactly the numbers in Table 5 (and Table 6). Nevertheless, the magnitudes are reassuringly similar: in the most comparable specification (column 1 of Table 10) one standard deviation change in teaching quality is associated to approximately 6% of a standard deviation change in contemporaneous outcomes for the 1999/2000 students. This is remarkably similar to the 7.1% that we estimated for the 1998/1999 cohort (column 3 of Table 5). When we further condition on the class fixed-effects (column 2 of Table 10) the effect increases to 8.1% and the comparison is even more accurate when we restrict the attention to the low-ability students but slightly less precise for the high-ability ones.

Furthermore, comparing the specifications with and without class fixed effects is informative about the ability of our approach to isolate the variation due to teacher quality. Ideally, if our estimates of teacher quality ($\widehat{\tau}_{dcs}$) were indeed purged of all class shocks, both observable and unobservable, we should find that conditioning on class fixed effects does not change the estimated coefficients significantly. In other words δ_1 and $\widetilde{\delta}_1$ should be very similar. Of course,

[37] The class shocks of any given academic year are common to all the courses of that year. However, for both the academic and the labor market returns the estimated δ_1 and $\widetilde{\delta}_1$ are positive and significant. Results are available from the authors upon request.

given that interactions between class and teacher effects exist and that their importance might have changed with the reform, we cannot expect the two sets of estimates to be exactly identical. Table 10 shows that, first of all, the coefficients on our teacher effects remain strongly significant also when conditioning on class effects (column 2 and 4) and, additionally, that they are actually larger than those conditioning on class observables (column 1 and 3), the differences being in the order of 20% to 30%. Hence, our strategy to measure teacher quality does seem to capture teacher effects as separate from class effects.

A second obvious concern is the fact that we do not observe earnings for all the students in our sample. However, we have access to the entire population of students who enrolled at Bocconi in the academic year 1998/1999 and to the complete list of tax declarations submitted in Italy in 2005 (on incomes earned in 2004), therefore our data are more akin to census data than to representative samples. In fact, we match 1,074 out of the 1,206 students in the enrollment cohort that we consider, so the selection problem is limited to approximately 10% of the observations.

Most Bocconi students find employment within a relatively short period of time after graduation, especially when compared with other Italian universities: of the 1,206 students that we observe entering Bocconi in 1998/99, two thirds graduate before 2004 and 94% graduate before 2005 (the minimum legal duration of degree programs being four years). Hence, the few students who are not matched can only be either unemployed or enrolled in post-graduate education or working abroad. Another possibility is total tax evasion, a phenomenon that is, however, quite uncommon even in Italy. The vast majority of people report at least some income and tax evasion is particularly common among the self-employed, which represent less than 3% of our sample. Dependent employees are taxed at the source directly by their employers and have very limited chances to evade. Notice additionally that tax evasion can distort our indicators of teaching quality only under the assumption that more or less effective teachers influence their students' performance as well as their propensity to evade taxes.

[INSERT FIGURE 2 ABOUT HERE]

To show that our findings are unaffected by the imperfect matching of the university administrative records and the tax files we employ two different strategies. First, we use data from an independent source (a survey of graduates regularly run by Bocconi University) to estimate the conditional probability of employment 1 year after graduation. We estimate the model on 6,355 individuals interviewed from 2002 to 2006 and we use the estimated parameters to compute the standard Heckman correction term for our sample. We then add it as an additional regressor in the estimation of equation 1. Second, we impute missing values using the *predicted mean matching method*. Such method uses linear predictions from a standard OLS model to measure distance across selected and non-selected observations. Then, a set of nearest neighbors for each non-selected unit is identified on the basis of such distance. Finally, imputed outcomes

for the missing observations are randomly drawn from their neighbors.[38] In figure 2 we show that the labor market returns computed in either ways are extremely similar to the ones we presented in section 4. The slope coefficients are not significantly different from one in both cases and the R-squared are always above 90%.

Above and beyond the mere selection issue one might also be worried that our measure of labor market success based on initial earnings may not reflect permanent income adequately. A vast literature documents the long-lasting effects of early labor market experience and provides support for our approach (Philip Oreopoulos, Till von Wachter & Andrew Heisz 2012, Robert H Topel & Michael P Ward 1992). However, we cannot exclude a priori that in some specific cases the best teachers could somehow systematically lead their students into professions with lower starting wages and higher long-term potential, thus invalidating our strategy. For example, some professional occupations, like lawyers and accountants, require an initial period of practice during which salaries are very low or even zero. Similarly, students going into graduate school typically earn little (or nothing) from teaching or research assistance and enjoy the full returns of their occupational choice only later in their careers. Although our data do not allow us to address this concern directly, the same surveys of graduates used for Figure 2 (left panel) allow us to produce a number of simple statistics documenting that this is a limited problem. For example, the vast majority of Management students go directly into the regular job market (95%), whereas a good 19.3% of the students graduating from Law and Management go into professional services (either legal practice or accounting) and the majority of those going into graduate school are from the degree program in Economics (6.8%). Hence, the students in the Management program are the least affected by potential selection into entry jobs and our main results are virtually unchanged when we focus exclusively on such a program (which, in fact, accounts for about 75% of the students in our sample).

Another potential concern is the possible lack of compliance with the random assignment to classes. There are a number of reasons why students could choose to attend lectures in a different class from the one they were assigned to and, especially if such a choice was related to the quality of the teachers, this could be problematic for our analysis. In principle students could request to be assigned to a different class but such requests would be accommodated only under very special circumstances. Students could not simply ask to be in the same class of their friends or with a teacher of their choice. However, under special conditions reassignment could be considered. For example, a student with some disability, temporary or permanent, who would find it difficult to climb stairs could request to be in a class taught on the ground floor.

In our data we only observe the class identifier that was originally assigned to the student, thus we cannot tell who was reassigned and to which classes. However, these official reassignments are too few to have any meaningful impact on our estimates. A potentially more troublesome case is informal class switching, i.e. students attending a course in a class of their

[38] We impute 10 values from 3 neighbors.

choice without formally requesting the change. Anecdotal evidence suggests that the incidence of such movements could be substantial, especially for some subjects.

To address this concern we make use of a specific item in the students' evaluation questionnaires asking about congestion in the classroom. Specifically, the question asks whether the number of students in the classroom was detrimental to learning.[39] If non-compliance with the random allocation is orthogonal to the teachers' characteristics, then it should have no obvious effect on class congestion: it should merely result in measurement error in the estimation of our class effects, inflating the variance of the estimation error without affecting their interpretation. The most worrisome type of class switching occurs when students cluster in the class of the best or the most pleasant teachers. For example, anecdotal evidence suggests that in the most difficult quantitative courses the students tend to bunch in the class of the professors who have a reputation for being particularly clear in their explanations. The courses most affected by class switching are those in which students concentrate in one or few classes, that end up being overly congested, whereas the other classes remain half empty.

Following this intuition, we compute for each course the difference in the congestion indicator between the most and the least congested class (over the standard deviation), thus identifying the courses most likely affected by class switching behavior. In table 11 we report descriptive statistics of academic and labor market returns of professors, as in table 5, dropping the most switched course (in panel B), the two most switched courses (in panel C) and the five most switched courses (in panel D), showing that our main results are virtually unaffected (panel A reports the main results of Table 5 for comparison).

[INSERT TABLE 11 ABOUT HERE]

Finally, we show that our results are not driven by the exclusion from the estimation sample of students that, after enrolling in their first year in the academic year 1998/99, dropped out before graduating. Such students total about 4% of all individuals in our enrollment cohort. In figure 3 we compare our estimates of the academic and labor market returns with similar estimates computed including the dropouts. The two sets of estimates are very similar, for both academic and labor market returns, and there are no major discrepancies at either ends of the distributions. As for the results in Figure 2, the slope coefficients are indistinguishable from one and the R-squared are always above 90%.

[INSERT FIGURE 3 ABOUT HERE]

5 Conclusions and policy discussion

In this paper we estimate the effect of teaching quality separately on students' academic and labor market performances. To the best of our knowledge, this is the first study ever to be able

[39] The questionnaires were administered in each class during one of the last lectures of the course. See Braga, Paccagnella & Pellizzari (2014).

to analyze the joint distribution of these alternative measures of teacher quality.

Overall the academic and labor market returns of teaching are positively correlated but this result is exclusively driven by the impact professors have on the best students. The correlation is, in fact, negative when returns are computed only on the low ability students. These findings lend support to the view of teaching as a multidimensional activity and further suggest that the complementarity between teacher and student abilities is stronger in the production of academic achievement than earnings.

From the policy perspective, our results speak to the entire literature on teachers' performance and raise a number of important questions both for measurement and for the design of incentive contracts. First and above all, the very definition of teaching quality needs to be precisely clarified before any measurement can be undertaken. Being a good instructor may mean very different things depending on both the types of students and the objective of the teaching process. Before thinking about how to measure teachers' performance, any education institution should define its own objective function. Some schools or universities may see themselves as elite institutions and consequently aim at recruiting the very best students and the teachers who are best at maximizing their performance. Other schools may adopt a more egalitarian approach and decide to improve average performance by lifting the achievement of the least able students. Similarly, some institutions may be more academic oriented and decide to focus on the teaching of one or more disciplines, regardless of their current market value. Other institutions may take a more pragmatic approach and decide to endow their students with the competencies that have the highest market returns at a specific point in time and in a specific location. The differences between community colleges and universities in the US or the dual systems of academic and vocational education that are common in countries like Germany and Switzerland are very good examples of teaching institutions with different objective functions.

References

Aaronson, Daniel, Lisa Barrow, and William Sander. 2007. "Teachers and Student Achievement in the Chicago Public High Schools." *Journal of Labor Economics*, 25: 95–135.

Angrist, Joshua D., and Victor Lavy. 1999. "Using Maimonides' Rule To Estimate The Effect Of Class Size On Scholastic Achievement." *The Quarterly Journal of Economics*, 114(2): 533–575.

Bandiera, Oriana, Valentino Larcinese, and Imran Rasul. 2010. "Heterogeneous Class Size Effects: New Evidence from a Panel of University Students." *Economic Journal*, 120(549): 1365–1398.

Becker, William E., and Michael Watts. 1999. "How departments of economics should evaluate teaching." *American Economic Review (Papers and Proceedings)*, 89(2): 344–349.

Bettinger, Eric P., and Bridget Terry Long. 2010. "Does Cheaper Mean Better? The Impact of Using Adjunct Instructors on Student Outcomes." *The Review of Economics and Statistics*, 92(3): 598–613.

Braga, Michela, Marco Paccagnella, and Michele Pellizzari. 2014. "Evaluating students evaluations of professors." *Economics of Education Review*, 41(0): 71 – 88.

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, and Jonah E. Rockoff. 2011. "The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood." National Bureau of Economic Research Working Paper 17699.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014*a*. "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates." *American Economic Review*, 104(9): 2593–2632.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014*b*. "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood." *American Economic Review*, 104(9): 2633–79.

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*, 126(4): 1593–1660.

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. 2012. "Class Size and Class Heterogeneity." *Journal of the European Economic Association*, 10(4): 795–830.

De Philippis, Marta. 2013. "Research Incentives and Teaching Performance. Evidence from a Natural Experiment." mimeo.

Dustmann, Christian, Patrick A. Puhani, and Uta Schönberg. 2012. "The Long-term Effects of School Quality on Labor Market Outcomes and Educational Attainment." Centre for Research and Analysis of Migration (CReAM), Department of Economics, University College London CReAM Discussion Paper Series 1208.

Ehrenberg, Ronald G., and Liang Zhang. 2005. "Do Tenured and Tenure-Track Faculty Matter?" *Journal of Human Resources*, 40(3).

Figlio, David N., Morton O. Schapiro, and Kevin B. Soter. 2013. "Are Tenure Track Professors Better Teachers?" National Bureau of Economic Research, Inc NBER Working Papers 19406.

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.

Gordon, Robert, Thomas J. Kane, and Douglas O. Staiger. 2006. "Identifying Effective Teachers Using Performance on the Job." The Hamilton Project White Paper 2006-01.

Hanushek, Eric A. 1971. "Teacher Characteristics and Gains in Student Achievement: Estimation Using Micro Data." *American Economic Review*, 61(2): 280–88.

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. 2009. "Teacher Deselection." In *Creating a New Teaching Profession.* , ed. Dan Goldhaber and Jane Hannaway, 165–180. Urban Institute Press.

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, Jorge 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.

Jackson, C. Kirabo. 2012. "Non-Cognitive Ability, Test Scores, and Teacher Quality: Evidence from 9th Grade Teachers in North Carolina." National Bureau of Economic Research Working Paper 18624.

Jackson, C. Kirabo. 2014. "Teacher Quality at the High-School Level: The Importance of Accounting for Tracks." *Journal of Labor Economics*, 32(4).

Kane, Thomas J., and Douglas O. Staiger. 2008. "Estimating teacher impacts on student achievement: an experimental evaluation." NBER Working Paper Series 14607.

Kane, Thomas J., Jonah E. Rockoff, and Douglas O. Staiger. 2008. "What does certification tell us about teacher effectiveness? Evidence from New York City." *Economics of Education Review*, 27(6): 615–631.

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.

McCaffrey, Daniel F., Tim R. Sass, J. R. Lockwood, and Kata Mihaly. 2009. "The Intertemporal Variability of Teacher Effect Estimates." *Education Finance and Policy*, 4(4): 572–606.

Oreopoulos, Philip, Till von Wachter, and Andrew Heisz. 2012. "The Short- and Long-Term Career Effects of Graduating in a Recession." *American Economic Journal: Applied Economics*, 4(1): 1–29.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. 2005. "Teachers, Schools and Acedemic 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.

Rothstein, Jesse. 2009. "Student Sorting and Bias in Value Added Estimation: Selection on Observables and Unobservables." *Education Finance and Policy*, 4(4): 537–571.

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.

Topel, Robert H, and Michael P Ward. 1992. "Job Mobility and the Careers of Young Men." *The Quarterly Journal of Economics*, 107(2): 439–79.

Weinberg, Bruce A., Belton M. Fleisher, and Masanori Hashimoto. 2009. "Evaluating Teaching in Higher Education." *Journal of Economic Education*, 40(3): 227–261.

Figures

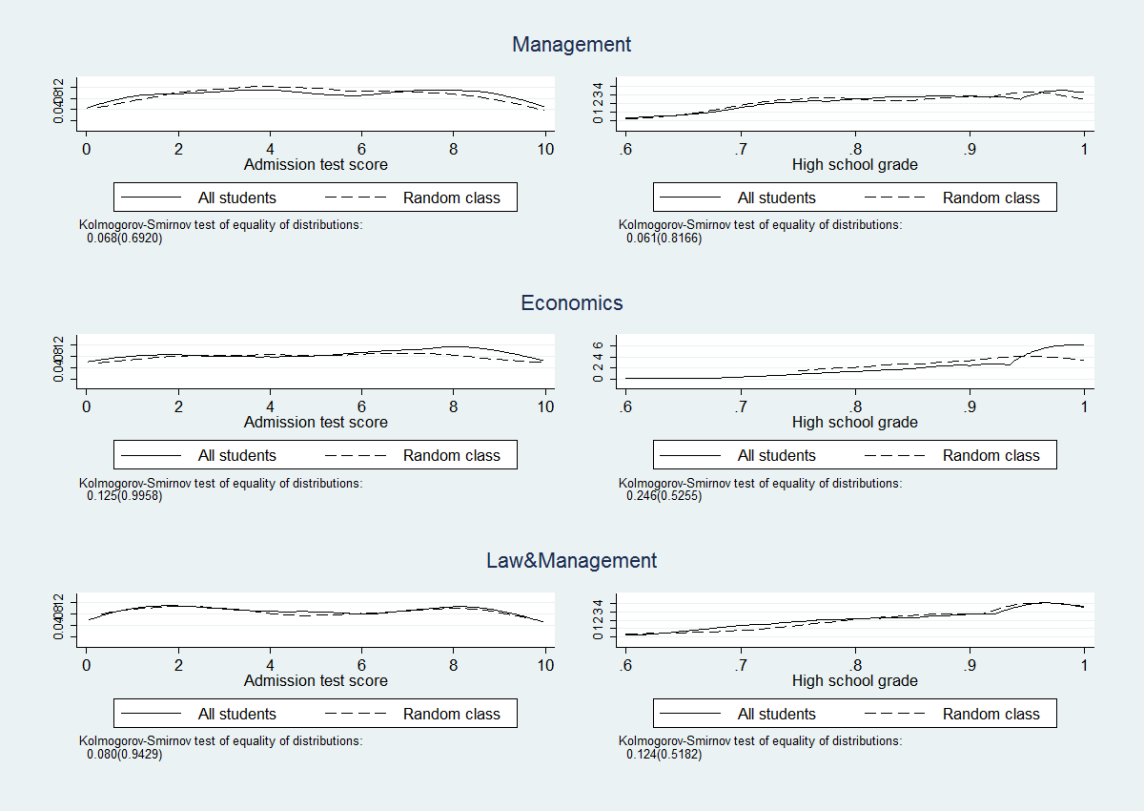

Figure 1: Evidence of random allocation - Ability variables

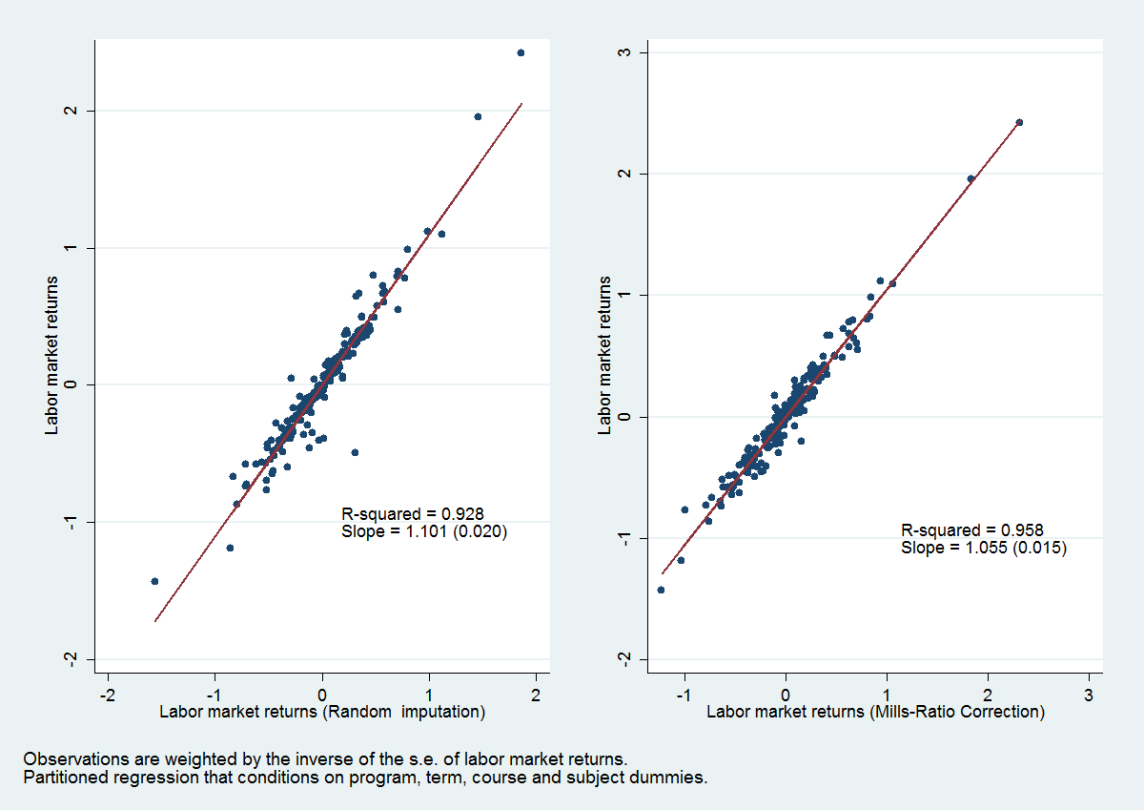

Figure 2: Robustness check for selection into employment

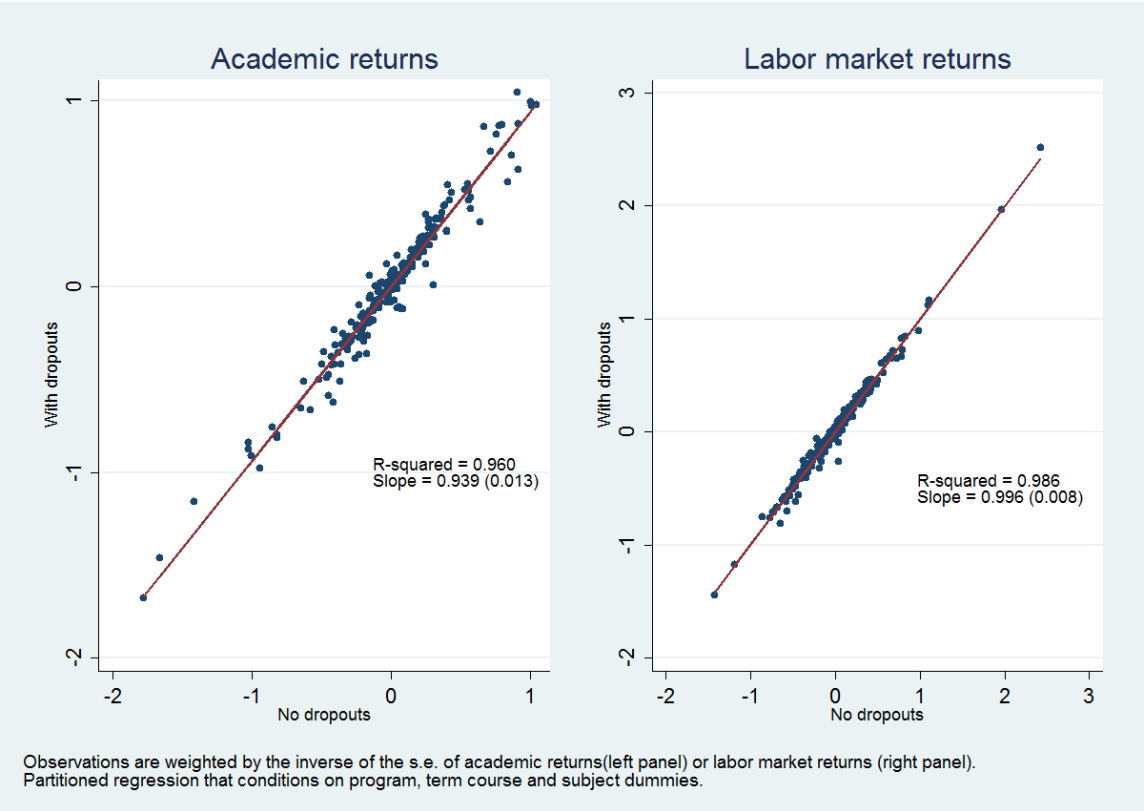

Figure 3: Robustness check for dropouts

Tables

Table 1: 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	93.053	88.084	87.181
	(10.905)	(8.878)	(10.852)	(10.904)
Entry test score (0-100)	60.422	63.127	58.894	60.496
	(13.069)	(15.096)	(12.262)	(13.224)
University grades (0-30)	25.684	27.032	25.618	25.799
	(3.382)	(2.938)	(3.473)	(3.379)
Wage (Euro)[e]	19,799.22	17,233.08	14,691.66	18,789.87
	(19,738.6)	(19,862.42)	(15,389.92)	(19,234.08)
Class size[f]	121.29	28.55	125.28	127.12
	(62.20)	(33.00)	(44.14)	(62.84)
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] Gross (before tax) annual income in 2004 at current value (2012 prices). 812 observations for Management, 100 observation for Economics, 162 observations for Law&Management (1,074 observations overall).

[f] Averages and standard deviations computed on the sample of classes rather than the sample of students.

Table 2: 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]
			Management				
Test statistics:	χ^2	χ^2	F	F	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
P-value[b] (total number of tests is 20)							
<0.01	0	0	0	1	0	3	0
<0.05	1	0	1	1	2	6	1
			Economics				
Test statistics:	χ^2	χ^2	F	F	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
P-value[b] (total number of tests is 11)							
<0.01	1	0	2	0	0	0	0
<0.05	1	0	2	1	0	0	0
			Law & Management				
Test statistics:	χ^2	χ^2	F	F	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
P-value[b] (total number of tests is 7)							
<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 1.

[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 3: Randomness checks - Teachers

	F-test	P-value
Class size[a]	1.11	0.351
Attendance[b]	0.89	0.535
Avg. high school grade	0.86	0.562
Avg. entry test score	0.99	0.446
Share of females	1.10	0.363
Share of students from outside Milan[c]	0.13	0.999
Share of top-income students[c]	1.63	0.103
Share academic high school[c]	1.80	0.065
Share late enrollees[c]	0.97	0.456
Share of high ability[d]	0.60	0.795
Morning lectures [e]	3.77	0.000
Evening lectures [f]	2.08	0.028
Room's floor[g]	0.52	0.992
Room's building[h]	3.52	0.000
Academic returns of previous teachers	1.52	0.134
Labor market returns of previous teachers	0.76	0.653

The reported statistics are derived from 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 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. All tests are distributed according to a F-distribution with (9,1215) degrees of freedom.

[a] Number or officially enrolled students.

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

[c] See notes to Table 1.

[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 4: Second-step regressions. F-stats and p-values

Dep. variable = $\widehat{\alpha}_s$	Regressors are the characteristics of					
	only the class		only the teacher		both	
PANEL A: Academic returns						
Class size[b]	1.87	(0.173)	-	-	0.70	(0.402)
Class composition[c]	0.81	(0.598)	-	-	0.87	(0.543)
Class time and room[d]	1.07	(0.381)	-	-	1.19	(0.313)
Coordinator[e]	-	-	0.34	(0.562)	0.52	(0.472)
Demographics[f]	-	-	0.20	(0.896)	0.29	(0.835)
Citations[g]	-	-	0.22	(0.801)	0.13	(0.876)
Academic rank[h]	-	-	1.52	(0.211)	1.65	(0.178)
Partial R-squared[i]	0.096		0.036		0.124	
PANEL B: Labor market returns						
Class size[b]	0.76	(0.384)	-	-	0.47	(0.492)
Class composition[c]	0.42	(0.911)	-	-	0.40	(0.921)
Class time and room[d]	0.20	(0.990)	-	-	0.12	(0.999)
Coordinator[e]	-	-	0.44	(0.510)	0.24	(0.627)
Demographics[f]	-	-	0.77	(0.510)	0.62	(0.600)
Citations[g]	-	-	0.71	(0.491)	0.69	(0.505)
Academic rank[h]	-	-	0.28	(0.842)	0.22	(0.886)
Partial R-squared[i]	0.027		0.025		0.047	
PANEL C: Contemporaneous academic returns						
Class size[b]	0.60	(0.440)	-	-	0.16	(0.688)
Class composition[c]	0.34	(0.951)	-	-	0.32	(0.959)
Class time and room[d]	1.23	(0.283)	-	-	1.46	(0.182)
Coordinator[e]	-	-	0.06	(0.810)	0.01	(0.942)
Demographics[f]	-	-	0.04	(0.990)	0.13	(0.939)
Citations[g]	-	-	0.03	(0.974)	0.04	(0.957)
Academic rank[h]	-	-	0.44	(0.721)	0.54	(0.657)
Partial R-squared[i]	0.058		0.009		0.068	

The reported statistics are F-tests (and corresponding p-values) for the null of joint significance of all the explanatory variables in the indicated category.

[a] Averages of individual characteristics if there is more than one teacher per class (weighted by the relative number of hours taught).

[b] Number of students in the class. Only one variable in this category.

[c] Average high-school leaving grade, average entry test score, % of students with entry test scores in the top quartile, % of students from academic high schools (*licei*), % of girls in the class, % of students residing outside Milan, % of students with incomes in the top bracket, % of students enrolled at university later than normal.

[d] Dummies for classes with lectures in the morning (start before 12am), lectures in the evening (start after 5pm) - afternoon is the reference category; dummies for the building and the floor where the classrooms are located; dummy for courses that are taught in different classrooms.

[e] Dummy equal to one if the teacher of the class is the coordinator of the course. Only one variable in this category.

[f] Gender (dummy) and age (squared) of the teacher.

[g] H-index and citations per year.

[h] Dummies for assistant, associate and full professor (non-tenure or tenure-track lecturers are the reference category).

[i] R squared computed once program, term and subject fixed effects are partialled out.

Table 5: Academic and labor market returns to teaching

	Academic returns	Labor market returns	Contemporaneous returns
PANEL A: Controlling for class and teachers' observables			
avg. standard deviation	0.043	0.055	0.071
min. standard deviation	0.000	0.005	0.004
max. standard deviation	0.143	0.184	0.308
PANEL B: Controlling for class observables only			
avg. standard deviation	0.043	0.054	0.116
min. standard deviation	0.000	0.006	0.008
max. standard deviation	0.142	0.181	0.428
No. of courses	38	38	38
No. of classes	230	230	230

The returns to teaching are estimated by regressing the estimated class effects (α) on observable class and teacher's characteristics (see Table 4). The standard deviations are computed as discussed in Section 3.

Table 6: Academic and labor market returns to teaching by student ability

	Low-ability	High-ability
PANEL A: Academic returns		
avg. standard deviation	0.066	0.067
min. standard deviation	0.002	0.002
max. standard deviation	0.318	0.203
PANEL B: Labor market returns		
avg. standard deviation	0.178	0.097
min. standard deviation	0.024	0.000
max. standard deviation	1.233	0.347
PANEL C: Contemporaneous returns		
avg. standard deviation	0.107	0.087
min. standard deviation	0.003	0.004
max. standard deviation	0.426	0.357
No. of courses	38	38
No. of classes	230	230

The returns to teaching by students' ability are estimated as in Table 5 (Panel A) but restricting the original sample of students to either those whose entry test scores are above the median (high ability) or below the median (low ability).

Table 7: Comparison of academic and labor market returns of teachers

Dependent variable =	Academic returns	Labor market returns
PANEL A: All students		
Labor market returns	0.098**	-
	(0.013)	
Contem. academic returns	-0.088***	-0.137***
	(0.000)	(0.036)
PANEL B: Low-ability students		
Labor market returns	-0.248***	-
	(0.043)	
Contem. academic returns	-0.001	0.002
	(0.012)	(0.025)
PANEL C: High-ability students		
Labor market returns	0.088*	-
	(0.051)	
Contem. academic returns	-0.016	-0.163***
	(0.018)	(0.024)

Each estimate reported in the table is obtained from a separate regression including the type of returns to teaching indicated in each row together with fixed effects for degree program, course, term and subject area. 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 8: Cross-comparison of academic and labor market returns to teaching by students' ability

Dependent variable: Returns for low-ability students		
	academic	labor market
Returns for high-ability students	0.103	0.453***
	(0.130)	(0.121)
Program fixed effects	yes	yes
Course fixed effects	yes	yes
Term fixed effects	yes	yes
Subject fixed effects	yes	yes

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 9: Joint distributions of the returns to teaching

			Quantiles of				
			Labor market returns				
			[1]	[2]	[3]	[4]	[5]
Quantiles of	Academic returns	[1]	39.13	19.57	13.04	17.39	10.87
		[2]	17.39	21.74	28.26	19.57	13.04
		[3]	17.39	23.91	28.26	17.39	13.04
		[4]	13.04	15.22	15.22	28.26	28.26
		[5]	13.04	19.57	15.22	17.39	34.78
			Labor market returns				
			[1]	[2]	[3]	[4]	[5]
	Contemp. returns	[1]	21.74	13.04	2.17	17.39	45.65
		[2]	17.39	19.57	21.74	23.91	17.39
		[3]	4.35	26.09	47.83	19.57	2.17
		[4]	10.87	34.78	10.87	17.39	26.09
		[5]	45.65	6.52	17.39	21.74	8.70
			Academic returns				
			[1]	[2]	[3]	[4]	[5]
	Contemp. returns	[1]	6.52	6.52	8.70	26.09	52.17
		[2]	2.17	30.43	19.57	21.74	26.09
		[3]	17.39	28.26	26.09	17.39	10.87
		[4]	36.96	10.87	23.91	21.74	6.52
		[5]	36.96	23.91	21.74	13.04	4.35

Table 10: Teacher's returns one year ahead

Dep. variable: contemporaneous exam grades	[1]	[2]	[3]	[4]
PANEL A: All students				
Contemporaneous academic returns[a]	0.058*** (0.012)	0.081*** (0.015)	0.059*** (0.012)	0.081*** (0.010)
Class effects	No	Yes	No	Yes
Student effects	random	random	fixed	fixed
Observations[b]	7,499	7,499	7,499	7,499
PANEL B: Low-ability students[c]				
Contemporaneous academic returns[a]	0.079*** (0.019)	0.105*** (0.017)	0.083*** (0.016)	0.105*** (0.019)
Class effects	No	Yes	No	Yes
Student effects	random	random	fixed	fixed
Observations[b]	3,765	3,765	3,765	3,765
PANEL C: High-ability students[c]				
Contemporaneous academic returns[a]	0.037** (0.016)	0.057*** (0.019)	0.035** (0.016)	0.057*** (0.018)
Class effects	No	Yes	No	Yes
Student effects	random	random	fixed	fixed
Observations[b]	3,734	3,734	3,734	3,734

[a] Teacher effects estimated using students' grades in the course taught by the teacher.
[b] One observation for each student in each course.
[c] Sample restricted to students with entry test below (low ability) or above (high ability) the median of the cohort.
Only students in the Management program. Both the dependent variable and the the contemporaneous academic returns of the teachers have been standardised within course.
Bootstrapped standard errors in parentheses. * p<0.1, ** p<0.05,***p<0.01

Table 11: Robustness check for class switching

	Returns computed on: grades	earnings
PANEL A: All courses		
avg. standard deviation	0.043	0.055
min. standard deviation	0.000	0.005
max. standard deviation	0.143	0.184
No. of courses	38	38
No. of classes	230	230
PANEL B: Excluding the most switched course		
avg. standard deviation	0.043	0.055
min. standard deviation	0.000	0.005
max. standard deviation	0.143	0.184
No. of courses	37	37
No. of classes	222	222
PANEL C: Excluding the two most switched courses		
avg. standard deviation	0.044	0.056
min. standard deviation	0.000	0.005
max. standard deviation	0.143	0.184
No. of courses	36	36
No. of classes	214	214
PANEL D: Excluding the five most switched courses		
avg. standard deviation	0.046	0.058
min. standard deviation	0.000	0.006
max. standard deviation	0.143	0.146
No. of courses	29	29
No. of classes	170	170

The returns to teaching are estimated by regressing the estimated class effects (α) on observable class and teacher's characteristics (see Table 4). The standard deviations are computed as discussed in Section 3.

Appendix

Table A-1: Compulsory courses by degree program

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

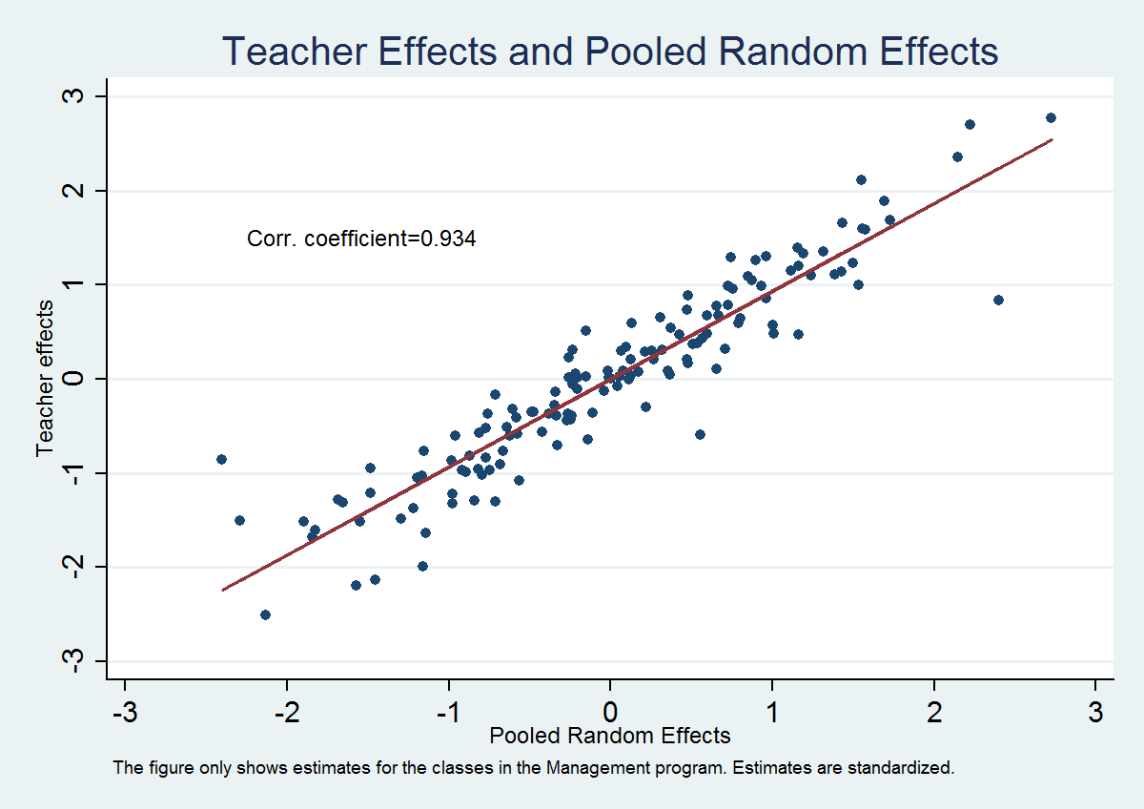

Figure A-1: Alternative estimates of the academic returns