DECLARATION FOR THE PhD THESIS

The undersigned

SURNAME	Paccagnella
FIRST NAME	Marco
PhD Registration Number	1197227

Thesis title:

Essays in Labor Economics

PhD in	Economics
Cycle	XXII
Candidate's tutor	Michele Pellizzari
Year of discussion	2012

DECLARES

Under his responsibility:

1) that, according to the President's decree of 28.12.2000, No. 445, mendacious declarations, falsifying records and the use of false records are punishable under the penal code and special laws, should any of these hypotheses prove

true, all benefits included in this declaration and those of the temporary embargo are automatically forfeited from the beginning;

2) that the University has the obligation, according to art. 6, par. 11, Ministerial Decree of 30th April 1999 protocol no. 224/1999, to keep copy of the thesis on deposit at the Biblioteche Nazionali Centrali di Roma e Firenze, where consultation is permitted, unless there is a temporary embargo in order to protect the rights of external bodies and industrial/commercial exploitation of the thesis;

3) that the Servizio Biblioteca Bocconi will file the thesis in its 'Archivio istituzionale ad accesso aperto' and will permit on-line consultation of the complete text (except in cases of a temporary embargo);

4) that in order keep the thesis on file at Biblioteca Bocconi, the University requires that the thesis be delivered by the candidate to Società NORMADEC (acting on behalf of the University) by online procedure the contents of which must be unalterable and that NORMADEC will indicate in each footnote the following information:

- thesis (*thesis* title) Essays in Labor Economics… ;
- by (*candidate's surname and first name)* …Paccagnella Marco…… ;
- discussed at Università Commerciale Luigi Bocconi – Milano in (year of discussion) …2012…. ;
- the thesis is protected by the regulations governing copyright (law of 22 April 1941, no. 633 and successive modifications). The exception is the right of Università Commerciale Luigi Bocconi to reproduce the same for research and teaching purposes, quoting the source;

5) that the copy of the thesis deposited with NORMADEC by online procedure is identical to those handed in/sent to the Examiners and to any other copy deposited in the University offices on paper or electronic copy and, as a consequence, the University is absolved from any responsibility regarding errors, inaccuracy or omissions in the contents of the thesis;

6) that the contents and organization of the thesis is an original work carried out by the undersigned and does not in any way compromise the rights of third parties (law of 22 April 1941, no. 633 and successive integrations and modifications), including those regarding security of personal details; therefore the University is in any case absolved from any responsibility whatsoever, civil, administrative or penal and shall be exempt from any requests or claims from third parties;

7) that the PhD thesis is not the result of work included in the regulations governing industrial property, it was not produced as part of projects financed by public or private bodies with restrictions on the diffusion of the results; it is not subject to patent or protection registrations, and therefore not subject to an embargo;

Date __27th January 2012_____

Signed (write first name and surname) ___Marco Paccagnella_____

Essays in Labor Economics

Marco Paccagnella 1197227

Universitá Commerciale Luigi Bocconi, Milano

Thesis Committee :

Michele Pellizzari, Department of Economics, Universitá Commerciale Luigi Bocconi, Milano

Tito M. Boeri, Department of Economics, Universitá Commerciale Luigi Bocconi, Milano

Marco Leonardi , Department of Economics, Universitá degli Studi di Milano

Dissertation in partial fulfillment of the requirements for the academic degree of

Doctor of Philosophy in Economics (XXII cycle)

A Paola e Anna

Acknowledgements

A large part of this thesis was completed while I was working at the Trento Branch of the Bank of Italy. I would thus like to thank my colleagues, and in particular Maria Lucia Stefani, for the constant support, encouragement and understanding.

The Bank in general has proved to be a very nice research environment. A special thank goes to Antonio Accetturo, Guglielmo Barone, Chiara Bentivogli, Raffaello Bronzini, Piero Casadio, Guido De Blasio, Elisabetta Olivieri, Sauro Mocetti, Pasqualino Montanaro, and Eliana Viviano. Needless to say, all the opinions expressed in this work are solely those of the author, and do not involve the responsibility of the Bank of Italy.

Michele Pellizzari has been a very supportive advisor (and coauthor), and I owe him a lot.

I am especially grateful to my coauthors. Antonio Filippin contributed to chapter 1, while Michela Braga and Michele Pellizzari contributed to chapter 2.

I have shared a large part of my graduate studies experience with Luna Bellani, Laura Brandimarte, Giovanna d'Adda, Michele d'Ambrosio, Lucia Esposito, Astrid Gamba, Giovanna Labartino, Elena Manzoni, Riccardo Masolo, Alessia Paccagnini, and Giovanni Vittorino: they have all been, at different points in time, great classmates, officemates, colleagues, and friends. Thank you everyone.

Contents

List of Figures

List of Tables

Preface

This dissertation consists of three essays in the field of Labor Economics. In the first chapter (which is the result of joint work with Antonio Filippin) we present a model that explains how initial differences in self-confidence can impact educational choices, in the second chapter (result of joint work with Michela Braga and Michele Pellizzari) we show that students' subjective evaluations of university professors are not a good proxy of the actual value-added provided by teachers, while in the third chapter I present some estimates of the impact of performance-related pay schemes on firms' productivity.

In the first chapter, titled *Family Background, Self-Confidence and Economic Outcomes*, we analyze the role played by self-confidence, modeled as beliefs about one's ability, in shaping task choices. We propose a model in which fully rational agents exploit all the available information to update their beliefs using Bayes' rule, eventually learning their true type. We show that when the learning process does not convergence quickly to the true ability level, even small differences in initial confidence can result in diverging patterns of human capital accumulation between otherwise identical individuals. As long as inital differences in the level of self-confidence are correlated with the socio-economic background (as a large body of empirical evidence suggests), self-confidence turns out to be a channel through which education and earnings inequalities are transmitted across generations. Our theory suggests that cognitive tests should take place as early as possible, in order to avoid that systematic differences in self-confidence among

equally talented people lead to the emergence of gaps in the accumulation of human capital.

The second chapter, titled *Evaluating Students' Evaluations of Professors*, contrasts measures of teacher effectiveness with the students' evaluations for the same teachers using administrative data from Bocconi University. The effectiveness measures are estimated by comparing the subsequent performance in follow-on coursework of students who are randomly assigned to teachers in each of their compulsory courses. We find that, even in a setting where the syllabuses are fixed, teachers still matter substantially. The average difference in subsequent performance between students who were assigned to the best and worst teachers (on the effectiveness scale) is approximately 23% of a standard deviation in the distribution of exam grades, corresponding to about 3% of the average grade. This effect translates into approximately 1.4% of the average entry wage or 160-200 euros per year. Additionally, we find that our measure of teacher effectiveness is negatively correlated with the students' evaluations of professors: in other words, teachers who are associated with better subsequent performance receive worst evaluations from their students. We rationalize these results with a simple model where teachers can either engage in real teaching or in teaching-to-the-test, the former requiring higher students' effort than the latter. Teaching-to-the-test guarantees high grades in the current course but does not improve future outcomes. Hence, if students are myopic and evaluate better teachers from whom they derive higher utility in a static framework, the model is capable of predicting our empirical finding that good teachers receive bad evaluations. The assumption that students evaluate teachers on the basis of perceived utility is supported by additional evidence that the evaluations respond to the weather conditions. Consistently with the predictions of the model, we also find that classes in which high-ability students are over-represented produce evaluations that are less at odds with estimated teacher effectiveness.

In the third chapter, titled *Performance-Related Pay and Firms' Productivity*, I estimate the productivity effect of a shift in firms' paying strategy towards collective performance-related pay schemes (PRP), using firm-level data from the Bank of Italy's Survey of Manufacturing Firms. Following the reform of industrial relations and of wage-bargaining institutions, in Italy the mid-nineties have witnessed a quick diffusion of firm-level contracts, which often contained some form of performance-related pay. Compensation schemes linked to firm's performance have also been subsidized, through different forms of fiscal incentives, which calls for a thorough assessment of the effectiveness of such practices in boosting firms' productivity. According to parametric estimates for a sample of medium and large sized Italian firms (above 50 employees), the introduction of PRP schemes has been associated with increases in productivity of 3-5%, much less than what found by previous studies. Employing more rigorous evaluation methods based on propensity-score matching I find even smaller effects, (0-2% in the short run, around 5% in the medium run), which are also generally not statistically signficant.

Chapter 1

Family Background, Self-Confidence and Economic Outcomes*

1.1 Introduction

Gaps in economic outcomes such as educational attainments and earnings tend to persist across generations and it is well-known that the socio-economic status of the parents is usually a very good predictor of the outcomes of their offspring.

Bowles, Gintis, and Osborne (2001) stress that "the advantages of the children of successful parents go considerably beyond the benefits of superior education, the inheritance of wealth, or the genetic inheritance of cognitive ability." They propose additional variables comparable to what now goes under the label of "non-cognitive skills" as factors that can supplement the ohterwise low explanatory power of the traditional variables used to fit the variance of earnings[1]. Moreover, they claim that

*This chapter is the result of joint work with Antonio Filippin

[1]In general, non-cognitive skills are defined as "personality traits that are weakly correlated with measures of intelligence" (Brunello and Schlotter, 2011). In this broad concept economists have usually investigated the so-called "Big Five" factors, following Nyhus and Pons (2005): agreeableness, conscentiousness, emotional stability, extraversion and autonomy. Other commonly used measures include the locus of control (Caraloc) and the Lawseq self-esteem score, but also attitudes toward risk and

the contribution of parental socio-economic status to earnings is in part determined by such non-cognitive skills, genetically transmitted or learned from parents that act as role models.

Since then many other authors have emphasized the role played by non-cognitive skills in explaining economic success and gaps in attainments. The current literature on the economic relevance of non-cognitive skills tends to treat these measures as inputs that enter the "black-box" of the skill production function. Cunha and Heckman (2007) propose a particular formulation of the technology of skill formation featuring self-productivity and dynamic complementarities among a multidimensional vector of cognitive and non-cognitive skills. They argue that insufficient investment in some of these skills early in life has long-lasting consequences that are very difficult or costly to revert. Heckman, Stixrud, and Urzua (2006), Cunha and Heckman (2007, 2008) and Cunha, Heckman, and Schennach (2010) have shown that gaps between children from different backgrounds open up very early in life, as soon as in pre-school age, and then tend to persist and stay roughly constant over the lifetime. Note how this finding clearly locates the rising of the problem in a period in which the role of the parents is the most important. Other recent papers that have focused on assessing the economic returns of different types of inputs (cognitive vs. non-cognitive skills) include Heineck and Anger (2010) and Lindqvist and Vestman (2011).

In this paper we want to analyze the role possibly played by a single non-cognitive skill, namely self-confidence, defined as the beliefs over one's unknown level of cognitive ability. Hence, our model entails the simplest possible multidimensional vector of skills, containing only two elements: a cognitive skill (innate ability) and a non-cognitive one (self-confidence). The use of such a framework is neither meant to deny the importance of other skills, nor the well-established fact that cognitive and non-cognitive abilities are

educational aspirations and expectations.

multidimensional in nature, nor to downplay the significance of the interaction among them. It simply reflects our goal to isolate and highlight a very specific mechanism, i.e. the role that a wrong self-confidence plays through the distortion of task choices. In other words, our purpose is to go into the "black-box" of the skill production function, identifying a precise and specific channel through which inherited differences in self-confidence can endogenously (i.e., through individual choices) explain the emergence and persistence of gaps in the accumulation of human capital.

The working idea of our model is that by acting as role-models, parents transmit to their children beliefs about their (unknown) ability. Such beliefs affect educational and task choices and, through this channel, contribute to widen the gap in the accumulation of human capital while the learning process (of acutal ability) takes place. The consequences of initially "wrong" beliefs can thus have long lasting effects, even if agents eventually learn their true level of ability.

An advantage of our approach is that the single non-cognitive skill we study has a clear and simple economic interpretation, and that we make transparent the channel through which it affects the accumulation of human capital (and thus, indirectly, earnings).

For self-confidence to have important effects we do not need to assume that agents enjoy holding a good image of themselves (i.e. that self-confidence enters directly the utility function), something that would imply that some degree of overconfidence is optimal[2]. Our theoretical framework assumes full rationality, given that agents extract all the available information from the signals received in order to update their beliefs, and this implies that they eventually learn their true type. Similarly, we exclude any other form of self-deception. The Bayesian learning mechanism is based on observing success or failure in the endeavour undertaken, given that the probability of success depends on the true level of ability as well as on the difficulty of the task, which is chosen

[2]Such an assumption is quite common in the behavioral economics literature (eg. in Köszegi, 2006 and Weinberg, 2009). We discuss this issue in more details in section 1.2.3.

endogenously in accordance with (updated) beliefs about one's ability.

Finally, we simulate the model with a bootstrapping procedure, showing that choices distorted by under-confidence (while all the other sources of heterogeneity are neutralized) lead to a significant gap in the accumulation of human capital during the learning process of the true level of ability. As long as it correlates with the family background, self-confidence constitutes therefore a channel through which gaps in educational attainments and earnings perpetuate across generations.

This finding also helps to explain why the early gaps based on the socio-economic background do not narrow when the role of the family becomes less important during life, and it suggests that policies aimed at providing early and accurate feedbacks on the cognitive skills of disadvantaged children can be important in promoting inter-generational income mobility.

The outline of the paper is as follows. In section 1.2 we survey the relevant literature comparing our theoretical approach with others in the literature. We also provide evidence sopporting both the important role played by self-confidence and the correlation between self-confidence and family background. In Section 1.3 we present a simple and parsimonious theoretical model that highlights how self-confidence can affect the accumulation of human capital via task choice. In section 1.4 we present the results of a simple simulation in order to better assess the implications of our model in terms of the emergenge of gaps in educational attainments between people from different backgrounds. Section 1.5 comments upon our results and draws some conclusions.

1.2 Motivation

In this section we survey the related theoretical and empirical literature to motivate the relevance of our work. In subsection 1.2.1 we document how incorrect beliefs about

ability are very common, and how this fact can have important practical consequences. In subsection 1.2.2 we discuss the different definitions and interpretations of confidence that have been used in the literature and why we have chosen to focus on a definition based on the *levels* of beliefs, rather than on their *precision*. In subsection 1.2.3 we discuss an assumption commonly made in the literature, i.e. that people actually care about their beliefs, and explain why decided not to make this assumption. Finally, in subsection 1.2.4 we justify one of the main assumption of our model by providing evidence that suggests the existence of a relevant link between self-confidence and the socio-economic background.

1.2.1 Imperfect knowledge of one's ability

We define self-confidence as the beliefs an agent holds about his own ability, following Bénabou and Tirole (2002), Hvide (2002), Köszegi (2006), Sjögren and Sällström (2004), and Weinberg (2009), among the others. This derives from the assumption that ability is unknown to the agent, instead of being his private information as in standard signaling models.

There is an extensive literature showing that agents hold a rough estimate of their cognitive skills. For instance, Dunning, Heath, and Suls (2004) survey the psychological literature documenting the presence of a weak correlation between actual and perceived performance in several domains, while Falk, Huffman, and Sunde (2006) provide experimental evidence that people are substantially uncertain about their relative ability and that this have indeed important consequences on search decisions.

Indirect evidence supporting the idea that individuals have imperfect knowledge about one's own ability can be inferred by observing that people with similar observable characteristics make different choices. In figure 1.1 we show that there is a considerable

degree of overlapping in the distribution of PISA 2006 test scores across people enrolled in different high-school tracks (which are very likely to yield very different returns on the labor market)[3].

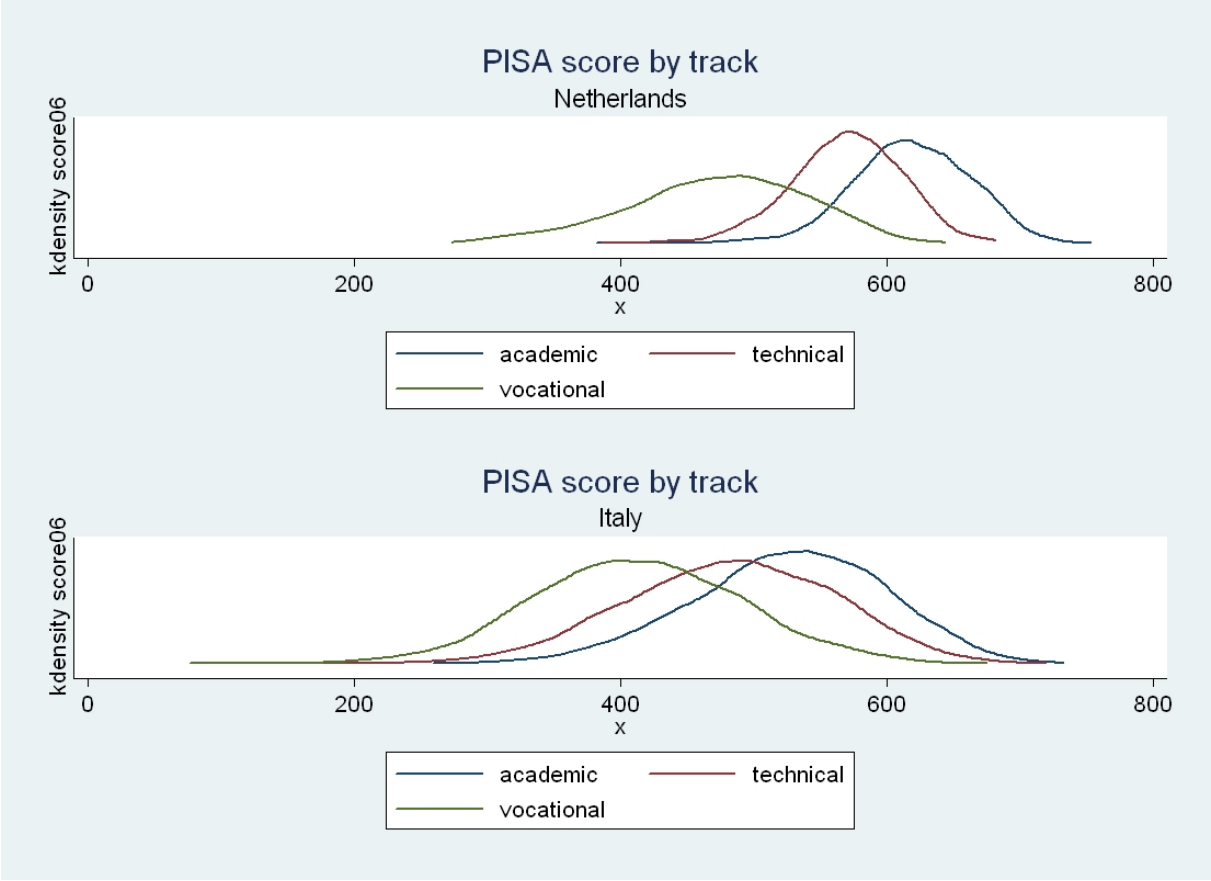


Figure 1.1: Netherlands (top), Italy (bottom)

Furthermore, comparing the top and the bottom panel (which refer respectively to the Netherlands and to Italy) we see quite a different degree of overlapping in the two countries. The main difference between the two educational systems is that while in Italy students and parents are perfectly free to choose the high-school track[4], in the Netherlands there is an aptitude test[5] administered at age 12 that, although not

[3] A similar figure appears in Checchi and Flabbi (2007).
[4] Although it is common to receive suggestions from school teachers.
[5] The so-called Cito test.

mandatory, has gained a considerable influence in recommending the secondary track most suitable for the pupil. While we recognize that there might be many other reasons to choose different tracks, the level of ability should also play an important role. Figure 1.1 shows that the degree of overlapping across tracks, which should in part be driven by ability mismatch, is much more pronounced in Italy where the uncertainty about one's ability is higher. There is also ample evidence (eg in Giuliano, 2008) that the socio-economic background has a strong influence on the choice of high-school track. These two stylized facts, examined toghether, suggests that systematic differences not related to cognitive ability do influence educational choices. Recent related literature has also investigated the role played by subjective returns to education in determining educational choices in developing countries. Jensen (2010) finds that, in the Dominican Republic, perceived returns to secondary school are extremely low, despite high measured returns, and that providing information about measured returns significantly increased enrollment rates; Nguyen (2008) finds instead that providing information on the returns to schooling improves school performance in Madagascar. Attanasio and Kaufmann (2009) and Kaufmann (2010) show that, in Mexico, subjective earning expectations and risk perceptions are important determinants of college attendance choices. However, they argue that,in that particular context, an important determinant of differential enrollment rates between poor and rich students is the presence of credit constraints. Imperfect information about ability (affected by different levels of confidence) is conceptually very similar to imperfect estimation of the market returns to education (due to incomplete information or to differences in the information set), and our model could easily be reinterpreted in this term. However, although being observationally equivalent, the two competing explanations can have very different consequences in terms of policy prescriptions.

1.2.2 Level vs. precision of beliefs

Another important issue is the definition of confidence in terms of the mean vs. the spread of the distribution of beliefs[6]. The former implies that an overconfident holds too high an estimate of his ability. The latter refers to an evaluation that is too precise, and better fits for instance an investment decision about which an agent can underestimate the variance of the future return. Confidence in terms of precision of beliefs could be adapted to a framework in which one's ability is the random variable, although it would be meaningless to talk about overconfidence as long as the true level of ability is a point estimate.

The two concepts, however, are not correlated, and their interaction can also determine counterintuitive results. For instance, it may happen that an agent that is quite confident along both dimensions could think to have a lower probability of success than another who is totally agnostic about his ability. To avoid such a possibility we assume that the probability of success is linear in ability, and we prefer to adopt a notion of confidence that refers to the level instead of the precision of beliefs. In our framework beliefs are defined as a probability distribution over the whole support of ability, so the second moment enters the picture but it can only affect the speed of convergence.

The focus on the level rather than on the precision of beliefs is one of the main difference between our model and the one proposed by Sjögren and Sällström (2004) (who also describe the endogenous evolution of self-confidence for rational agents that choose tasks and update their beliefs in a Bayesian fashion after observing the outcomes of their choices). A second major difference is that, in our framework, agents eventually discover their true type, while Sjögren and Sällström (2004) show how people can

[6]Both definitions are used in the literature. The first for instance by Hvide (2002), Bénabou and Tirole (2002), and Weinberg (2009), the second by Sjögren and Sällström (2004), while Köszegi (2006) and Belzil (2007) use both.

remain “trapped” with wrong beliefs due to insufficient experimentation and learning[7].

The probability of success that characterizes our model implies that our agents make decisions under uncertainty; our work can thus be linked to the literature that sees education as a risky investment and that investigates the role played by risk aversion. Belzil (2007) estimates both the degree of over and under estimation of labor market skills and the dispersion of the distribution of subjective beliefs. He finds evidence of frequent (but moderate) over-estimation and cases of severe under-estimation (particularly among the most able individuals). He also finds that only 25% of unobserved ability heterogeneity is perceived by the individuals as ex-ante risk and that 36% of the population act on the basis of a degenerate subjective ability distribution. Belzil and Leonardi (2007) find that risk aversion can be a deterrent to investing in education, but that differences in risk attitude account for a modest portion of the probability of entering higher education. Since the effect of underconfidence can be confounded with that of risk aversion, in order to isolate the role of confidence we assume that agents are risk neutral.

1.2.3 Self-confidence in the utility function

There is a wealth of contributions in the psychological literature showing that confidence affects the task choice (see Weinberg (2009) and literature therein). In this paper we also focus on the role that confidence plays through this channel, and more specifically on how task choice shapes human capital acquisition. In our model confidence affects utility only indirectly through the choice of task, which determines how much human capital the agent gets, and there is no direct influence like there would be in case the agent enjoys thinking that his ability is high. Examples of models in

[7] To achieve this result, they have to assume that there are non-informative task, in which the probability of succes is equal to one.

which beliefs about one's ability enter directly the utility function are Weinberg (2009) and Köszegi (2006). While such models rationalize many interesting features of human behaviour (along with the result that moderate levels of overconfidence turn out to be optimal), we decide to stick to a simpler theoretical framework in which this does not happen. The main reason is that once agents are supposed to enjoy holding a good self-image, they should also be capable of tailoring the information acquisition during their learning process in such a way to preserve it, for instance by means of beliefs that are "pragmatic" (Hvide, 2002) or more generally self-serving, as well as with selective memory (Bénabou and Tirole, 2002)[8].

Manipulating the information acquisition can only be effective *in the short run*, unless agents end up stuck in a self-confirming equilibrium in which their learning process reaches a fixed point although their beliefs are wrong. In other words, beliefs are wrong but never disconfirmed by the evidence either because further experimenation is not available or because agents continue to indefinitely self-deceive themselves[9]. Although such an outcome cannot be exclueded, we find more interesting to analyze the effect of holding a wrong self-image when the true type is eventually learned. Including beliefs in the utility function would only incentivate some form of self-deception that, even allowing the manipulation of information acquisition, would only have the transitory effect of slowing down the learning process, and therefore we prefer to avoid such a complication. Our model thus adheres to a perfectly rational framework, with agents characterized by standard preferences and that unbiasedly exploit the whole information available.

[8]Bénabou and Tirole (2002) also assume that discount rates are lower at shorter horizons than at more distant ones (time-inconsistency). Belzil (2007), however, find a predominance of the future component of intertemporal utility over the present component in schooling decision, and interpret it as evidence supportive of the standard time-consistent model.

[9]Models in Köszegi (2006) and Weinberg (2009) are characterized by a small number of periods. Hvide (2002) justifies pragmatic beliefs in the long run with a thought experiment in which "the agent takes into account what pays rather than what is true."

1.2.4 The inter-generational transmission of confidence

ecent empirical findings provide support for one of the key assumptions of our model, namely that self-confidence is correlated with the family background. Cesarini, Johannesson, Lichtenstein, and Wallace (2009), using Swedish data on a sample of twins and defining overconfidence as the difference between the perceived and actual rank in cognitive ability, argue that genetic differences explain 16-34% of the variation in overconfidence, and that common environmental differences explain an additional 5-11%. A series of studies on different longitudinal UK datasets (collected in Goodman and Gregg, 2010) find a strong intergenerational correlation not only in cognitive skills, but also in a variety of attitudes that can be considered proxies of confidence. In particular, Gregg and Washbrook (2011) find that, even after controlling for long-run family background factors and prior attainment, children are more likely to perform well in tests at age 11 if they have strong beliefs in their own ability and have a more internal locus of control[10], and they also find that children from poorer families are less likely to have these attributes. Chowdry, Crawford, and Goodman (2011) find that richer parents have higher expectations of their children's educational attainments and that young people from poorer families have lower ability beliefs, a more external locus of control and lower educational aspirations and expectations. After controlling for attainment at age 11, 15% of the socio-economic gap in attainment at age 16 is accounted for by child attitudes, and an additional 12% is accounted for by parental attitudes. Chevalier, Gibbons, Thorpe, Snell, and Hoskins (2009) find that working class undergraduates underestimate their performance relative to others, but also that working class secondary school pupils have greater confidence and a more positive self-evaluation of their math ability. This finding may be due to differences

[10] People with an external locus of control tend to think that luck or fate, rather than their own actions, are what matters in life. It is likely that this is related to low levels of confidence in own ability.

in peer-groups and to the “big fish, small pond effect”. Here we provide additional evidence about the link between socio-economic background and self-confidence using data from the OECD-PISA study. This dataset contains what we believe is a good proxy for self-confidence, namely “Science Self-Efficacy”, an index built from student’s answers to questions about the ease with which they believe they could perform eight science-related tasks. This variable is a good proxy for beliefs about academic ability because it is meant to go “beyond how good students think they are in subjects such as science. It is more concerned with the kind of confidence that is needed for them to successfully master specific learning tasks, and is therefore not simply a reflection of a student’s abilities and performance” (OECD, 2009)[11].

We thus regress our measure of confidence on family background, adding controls at the individual, school and family level; results are presented in Table 1.1.

The relationship between Self-efficacy and family background is significant and positive as expected, displaying a convex correlation. In the second column we also control for the score obtained by the student in the Science section of the test. This is a proxy for “true” ability, comparable across students in different countries and unobserved by the student at the time of filling in the questionnaire. The inclusion of PISA score captures some variance of self-efficacy, but the positive relationship with family background remains strong. Notice that controlling for the PISA score is likely to bias downward the role played by self-confidence, because if our model is correct the PISA score already encompasses the gap in the human capital accumulated up to that point also because of a different self-confidence. In other words, two students with the same innate ability but characterized by a different self-confidence should also display a different PISA score.

[11] See Ferla, Valcke, and Cai (2009) for a discussion on the differences between Self-Efficacy and Self-Concept. Since Self-Efficacy solicits goal-referenced evaluation and does not ask students to compare their ability to that of others, we believe it is a better proxy for the notion of confidence that we use in the model of Section 1.3.

Adding further controls at the student level (column 3) and at the parent and school level (column 4) does not change significantly the results, which we interpret as suggestive evidence that family background has a direct impact on self-confidence, over and above the one operating through the transmission of cognitive skills.

Table 1.1: Results: Science Self-Efficacy

	[1] Baseline	[2] Pisa score	[3] Effort	[4] Parents
Index of socio-ec. status	0.318***	0.145***	0.111***	0.119***
	[0.014]	[0.014]	[0.015]	[0.026]
Index of socio-ec. status^2	0.033***	0.022*	0.030**	0.026
	[0.009]	[0.009]	[0.009]	[0.014]
Female	-0.157***	-0.141***	-0.157***	-0.031
	[0.012]	[0.011]	[0.011]	[0.017]
PISA score in Science		0.004***	0.004***	0.003***
		[0.000]	[0.000]	[0.000]
Out of school - Science			0.112***	
			[0.009]	
Self study - Science			0.114***	
			[0.006]	
Interest in learning science				0.226***
				[0.012]
Personal value of science				0.222***
				[0.013]
Parents' value of science				0.001
				[0.013]
Science career motivation				-0.029**
				[0.009]
Science activities at age 10				0.062***
				[0.008]
School-level characteristics	NO	NO	NO	YES
R squared	0.119	0.230	0.255	0.355
Observations	225,098	225,098	216,304	29,970

BRR standard errors in brackets. * p<0.1, ** p<0.05, *** p<0.01.
All regressions include country dummies and control for immigrant status, tracking and the interaction between tracking and the socio-economic status. In column 4 we also control for school-level variables like school size, student-teacher ratio, ability sorting and a dummy for public schools.

The PISA dataset has the advantage of being a large-scale, international, representative sample, but includes extremely heterogeneous students at an early stage of

their education career. Therefore, we replicate a similar analysis using another dataset with opposite characteristics, coming from a survey of a much more homogeneous population at later stage of their academic career. This dataset has been collected by circulating a questionnaire to all second year Bocconi students in 2001, subsequently merged with administrative data. It contains information about students' expectations on occupation and wages 1 and 10 years after graduation, about their family background, as well as detailed information on their academic career[12]. We use expected wage as a proxy for self-confidence, while family-background is proxied by parents' educational levels and the students' tuition category (a function of family income). Wage expectations 10 year after graduation are probably the best measure of self-confidence, since after such a spell of time wages should be expected to reflect productivity more precisely[13]. Notice that also in this case the proxies for ability are likely to bias downward the role played by self-confidence, since they also control for the gap in the human capital accumulated up to that point. Table 1.2 reports results from regressing the log of expected wage ten years after graduation on family background variables and individual controls.

While parental education does not seem to have a significant impact on expected wages, the effect of family income (proxied by tuition category) is significant and J-shaped, with a minimum in the third category[14]. Results are almost unchanged when

[12]The same data are used in Filippin and Ichino (2005), to which we refer for further details on the characteristics of the dataset.

[13]For the sake of brevity we only report results using expected wages 10 years after graduation. Results using short-term expectations are not significantly different, and are available upon request.

[14]At that time, there were 6 brackets, and more than 60% of the students in our sample were in the top three categories (with 35% of students in the top bracket). Our results imply that students in the lowest income category are more confident than those from the middle class (third income bracket). A possible explanation is that Bocconi is a very expensive university where rich families are over-represented. Students from poor families are instead under-represented because they could not afford the tuition fees without financial help, which is awarded only if strict requirements in terms of academic performance are fulfilled. Therefore, the subsample of students in lower income brackets is likely to suffer a stronger self-selection problem because only particularly good and strongly motivated students are able to enroll.

both measures of family background are included.

Table 1.2: Expected Wage 10 Years After Graduation

	[1] Parental Ed.	[2] Income	[3] Income squared	[4] Full
Parent graduate	0.056			0.030
	[0.030]			[0.031]
Parent primary ed.	0.001			-0.006
	[0.057]			[0.057]
Income bracket		0.030***	-0.138**	-0.142***
		[0.009]	[0.042]	[0.043]
Income bracket^2			0.022***	0.023***
			[0.005]	[0.005]
Female	-0.085**	-0.094**	-0.094***	-0.091**
	[0.029]	[0.029]	[0.028]	[0.028]
Family firm	0.212***	0.174**	0.164**	0.165**
	[0.057]	[0.058]	[0.057]	[0.057]
Average grade	0.022**	0.021**	0.019*	0.019*
	[0.008]	[0.008]	[0.008]	[0.008]
High School grade	-0.395*	-0.342	-0.316	-0.319
	[0.199]	[0.198]	[0.197]	[0.197]
R squared	0.117	0.127	0.146	0.147
Observations	764	764	764	764

Standard errors in brackets. * p<0.1, ** p<0.05, *** p<0.01.
All regressions include dummies for degree program, type of high school, region of residence and expected sector of employment.

Bocconi is recognized as an elite university in Italy, widely known to attract very good students and well recognized in the labour market. Hence, one should expect that the signal provided by graduating at Bocconi is strong enough to more than counterbalance the effect of any other difference in students' former endowments. In contrast, we find that the different socio-economic background still shapes wage expectations. Hence, the same observed (and observable) signals have a different impact on different people. Our interpretations is that inherited beliefs about one's own ability survive a string of commonly-believed-to-be very good signals. Unfortunately, we cannot attribute a causal interpretation to this result, because such a correlation could be a spurious spillover of

different networking abilities or different preferences correlated to the family background. However, the same correlation appears in the wage realizations of a similar but richer survey of Bocconi graduates in which a larger set of controls is available[15]. Moreover, our results are similar to what has been recently found by Delaney, Harmon, and Redmond (2011), who use a dataset collected from seven Irish universities (and thus certainly more representative of the population of Irish undergraduate students), and that also include many different measures of non-cognitive skills such as risk attitudes, time preferences and personality traits.

1.3 The Model

In this section we present in more details a multi-period model based on the assumptions already discussed in sections 1.2.1-1.2.4, in which agents choose a task on the basis of their beliefs, which are updated in a Bayesian manner after observing the outcome of every choice. Our purpose is to highlight the role played by confidence in explaining educational attainments via task choice.

As already explained, we assume that children do not know their own ability a and hold a belief represented by the density function $\mu(a)$. We define confidence the perceived ability $\hat{\mu}(a) = \int a\mu(a)da$ and underconfident a student who underestimates her ability: $\hat{\mu}(a) < a$. Similarly, the overconfident is characterized by $\hat{\mu}(a) > a$. Students make educational choices by choosing "tracks" (ψ). We think of tracks as a rather general concept, encompassing either "real" school tracks (eg. academic vs. vocational high schools) or any goal that the student sets herself. In the latter sense a track could well be interpreted as the amount of knowledge encompassed in a concept. More difficult tracks in both interpretations are more costly in terms of effort, but they also yield

[15]Results are not displayed to save space but they are available upon request.

higher payoffs in case of success. A failure could be interpreted either as a true failure in a real track (eg. the student drops out or must repeat a grade) or as the chance that, in trying to deeply understand some difficult material, the student wastes energy and time, ending up learning less than she would have done had she been less ambitious.

We assume that the probability of success is given by

$$p(s) = f(a, \psi) \tag{1.1}$$

where ψ represents how difficult is the track chosen. The probability of success is assumed to be increasing in ability ($f'(a) > 0$) and decreasing in the difficulty of the track ($f'(\psi) < 0$).

Students have then the possibility of updating their beliefs using Bayes' rule, when additional information can be derived from the outcome of their choice. Given a generic density of prior beliefs $\mu(a)$, posterior beliefs after receiving the signal implicit in the outcome $o = \{s; f\}$ are equal to:

$$\mu(a|o) = \frac{p(o)\,\mu(a)}{\int p(o)\,\mu(a)\,da.} \tag{1.2}$$

Successful outcome (s) in the track chosen allows agents to add human capital $k(\psi|s)$ to working life productivity, and agents maximize their instantaneous utility by choosing the track that optimally balances their expected acquisition of human capital with a convex cost of acquiring it $U[p(s)\,k(\psi) - \psi^2]$, given their confidence about unobserved ability.

If the track chosen is totally uninformative (e.g. $p(s) = 1$) the student does not gather evidence that contradicts his/her wrong beliefs. For instance, this may happen when there is a discrete set of tracks and the less able students self-select into the easiest track characterized by no probability of failure. This is admittedly a limit situation, and

therefore we prefer to concentrate on what happens to the gap in the accumulation of human capital when agents do learn from observed outcomes and proceed with Bayesian updating of their beliefs until their confidence eventually converges towards the true value of ability.

To achieve this goal we make some simplifying assumptions. First, we assume that the probability of success is linear in ability. The reason is that, as anticipated in section 1.2.2, we concentrate on the role played by the *level* of one's perceived ability, and not by the *precision* of such belief. This is a major difference for instance with respect to the model in Sjögren and Sällström (2004), who assume that the probability of succesfully acquiring skills of type c_1 is $p(s) = a^{c_1}$, where $a \in [0,1]$ is the agent's unknown ability, while $c_1 > 1$ measures the ability elasticity of success. In such a framework the precision of the signal is crucial, because uncertainty about ability makes riskier options more or less attractive depending on whether the probability of success is convex or concave in ability. For instance, what could happen with a convex probability of success is that a totally uncertain agent could think to have more chances of succeeding than an agent characterized by quite a precise belief of being above the average. In contrast, we choose to remove such discontinuities by assuming linearity in ability in equation (1.1) and to focus on the effect of the level of confidence[16]. Hence, we assume the following functional form of the probability of success:

$$p(s) = \psi a + (1 - \psi). \tag{1.3}$$

This specification implies that the importance of ability is proportional to the difficulty of the track. Notice that for the probability of success to be properly defined we need ability to have a finite support, and for the sake of simplicity we assume both $a \in [0,1]$

[16]Note that in order to neutralize the effect of the precision of beliefs it is not enough to assume the same variance of prior beliefs, because at different level of confidence the impact of the variance would be different as long as the probability of success is not linear in ability.

and $\psi \in (0, 1]$. The extreme value $\psi = 0$ would correspond to the uninformative case mentioned above in which ability does not matter and the signal is totally uninformative (see Figure 1.2).

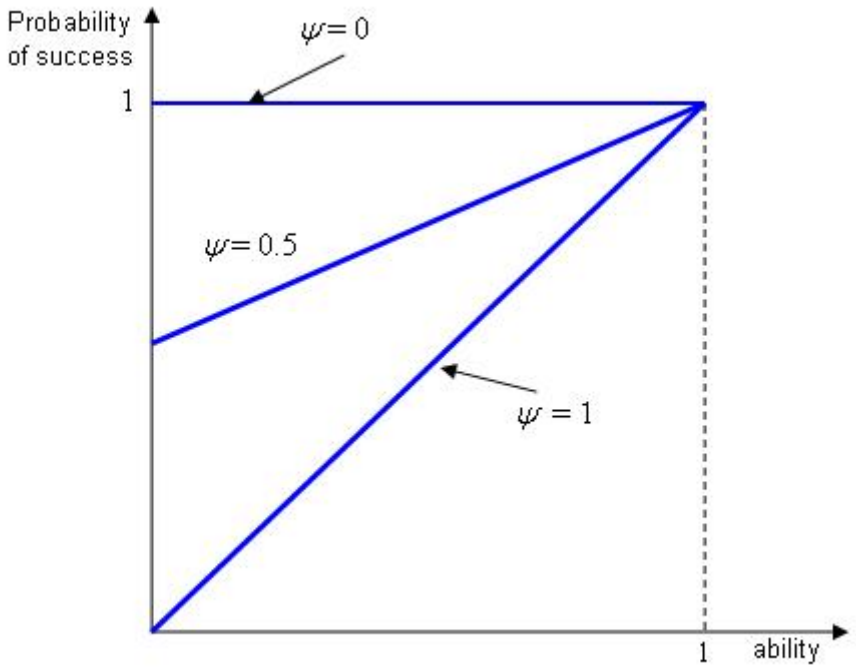


Figure 1.2: Different tracks in terms of importance of ability

We also assume that more difficult tracks allow students to acquire more human capital if successful, and in particular that the level of capital is equal to

$$k(\psi, \mu(a), a|s) = \frac{\psi}{1 + f(m)}, \tag{1.4}$$

where $m = (a - \hat{a})$ represents the ability mismatch. $\hat{a}$ represents the optimal level of ability for that track, i.e. the level of ability that characterizes the student that maximizes her utility by choosing exactly that track. We assume that $f(0) = 0$, i.e. that human capital concides with the difficulty of the track when ability perfectly fits, otherwise ψ

is corrected, with the shape of $f(m)$ when $m \neq 0$ crucially affecting the results. In particular, we assume that $f'(|m|) \geq 0$ meaning that neither under- nor overconfidence can increase human capital beyond ψ. This assumption might appear counterintuitive at first glance, but it has the great advantage of preventing self-deception. Consider the case in which in the same track the human capital is lower only for the overconfident successful students, because their ability is lower than what optimal for such a track, while the opposite happens for the underconfident successful students. In this case, the possibility of supplmenting the human capital provided by the chosen track with an ability higher than $\hat{a}$ implies that there is room for self-deception, i.e. that systematically underestimating one's ability might become an optimal solution, with a consequent bias in the choice of the track that we want to avoid for the same reasons outlined in section 1.2.3. Of course, the effect of the mistake in evaluating ability does not need to be symmetric. In the simulation below we will assume that underconfidence has no effect ($f(m) = 0$ when $m < 0$), while overconfidence has a negative impact ($f'(m) > 0$ when $m > 0$). To complete the picture, we assume that a failure leaves the stock of human capital unchanged, i.e. $k(\psi, \mu(a), a|f) = 0$[17].

Students are free to self-select into different tracks given the best estimate of their ability, trading off a lower human capital in case of success with a higher probability of acquiring it. If ability was known, the first-order conditions would imply[18]:

[17] This assumption is made without loss of generality as compared to the case in which the human capital accumulated in case of failure is positive but strictly lower: $k(\psi, \mu(a), a|s) > k(\psi, \mu(a), a|f)$.

[18] To analyze the role played by self-confidence in shaping the gap in educational attainments when agents are eventually learning their true level of ability we need to iterate this choice for several periods. In principle, we should compute the optimal track choice by maximizing a lifetime utility function. Since additional information about one's ability is valuable per se as long as it helps making better choices in the future, agents could be willing to pay a price to receive a more informative signal, by choosing a track slightly different than what would be optimal in a static framework. However, such an effect is of a second order magnitude and it does not determine appreciable changes in the results (see footnote 24 below), thereby not justifying the corresponding increase in the complication of the model. Hence, we assume that agents are myopic and that they maximize their expected utility period by period.

$$\psi^* = \frac{1}{2}\frac{1}{2-\hat{a}} \tag{1.5}$$

Given that $f(m)$ implies to truthfully self-report one's unknown ability, i.e. to set the mistake $\mu(a) - a = 0$, the optimal choice of track becomes an increasing function of confidence. However, even removing any bias in the self-evaluation of ability, $\mu(a)$ and $\hat{a}$ may still differ due to insufficient information. Equation 1.5 therefore implies that both under- and over-confidence determine a suboptimal track choice and a loss of utility due to the mismatch $\mu(a) \neq \hat{a}$.

The effect of under- and over-confidence can differ as far as the accumulation of human capital is concerned. Rewriting confidence as the composition of optimal ability and the evaluation mistake $\mu(a) = \hat{a} + m$ we can derive that the expected human capital is given by:

$$E(k) = -\frac{1}{4}\frac{\hat{a} + 2m - 3}{(\hat{a} + m - 2)^2(1 + f(m))}. \tag{1.6}$$

The relationship between confidence and human capital can be summarized by means of the derivative of $E(k)$ with respect to the mistake m:

$$\frac{\delta E(k)}{\delta m} = \frac{1}{2}\frac{m-1}{(\hat{a} + m - 2)^3(1 + f(m))} + \frac{1}{4}\frac{(\hat{a} + 2m - 3)f'(m)}{(\hat{a} + m - 2)^2(1 + f(m))^2}. \tag{1.7}$$

As long as a small ability mismatch has a negligible impact, i.e. as long as $f'(0)$ is sufficiently small, the derivative is positive around $m = 0$ for every value of $a \in [0, 1]$. This means that a small degree of overconfidence $(m > 0)$ increases the amount of expected human capital, although at a price of lower utility because the increase of human capital would be acquired overestimating the expected return on the additional effort[19]. As

[19]The reason is that the probability of success depends on the true level of ability, and overconfidence would grant a higher level of human capital when successful, but a positive outcome is less likely to happen than what an overconfident agent expects.

overconfidence increases, the sign of $\delta E(k)/\delta m$ depends on the magnitude of the effect of the mismatch. In the limit case in which there is no effect, e.g. when $f(m) = 0$ in Equation 1.4, or in any case when such an effect is negligible, the human capital acquired would monotonically increase with overconfidence since the positive effect of the higher human capital acquired when successful dominates the negative effect of a lower chance that this event happens. In contrast, if the effect of overevaluating one's ability increases substantially with the size of the mistake (e.g. if $f(m) = m^2$) the relation between expected human capital and overconfidence becomes bow-shaped. As far as underconfidence is concerned, the condition that ensures that there is no incentive to self-deception is also sufficient to grant that human capital decreases monotonically as underconfidence increases.

Agents update their beliefs given the signal received (success or failure) at the end of each period[20]. In order to characterize the learning process and to investigate the effect of self-confidence on educational attainments we need to specify how beliefs about one's ability are shaped. The Beta distribution perfectly fits our assumption of a finite support of the ability distribution, necessary to ensure that the probability of success is linear in ability. At the same time the Beta distribution is sufficiently general to allow prior beliefs to represent different levels of confidence while keeping the whole domain of ability in their support, something necessary because with a Bayesian learning process agents can never assign a positive probability to events excluded by the prior.

[20]Note that was the agent receiving a perfectly informative signal like the exact amount of human capital acquired when successful he could invert $k(\psi, \mu(a), a|s)$ deriving with certainty her true ability level. However, data suggest that uncertainty about ability survives many signals, which therefore are not perfectly informative (or even if they are perfectly informative agents cannot fully exploit them). In what follows we assume that agents only observe the event success vs. failure. In other words, agents know only the potential amount of human capital ψ but not the actual amount once corrected for the mismatch of ability $1 + f(m)$. An intermediate situation in which additional information can be extracted from a noisy signal of the level of human capital actually acquired (in other words when different degrees of success are observable) could be formalized at the price of a significantly increased complication of the model without appreciable additional insights. Hence, we prefer to stick to the simplest version of the information structure.

The density function of the $Beta\left[\alpha,\beta\right]$ distribution is:

$$\mu(a)=\frac{a^{\alpha-1}(1-a)^{\beta-1}}{\int_0^1 a^{\alpha-1}(1-a)^{\beta-1}da}, \tag{1.8}$$

while the mean is given by:

$$\hat{\mu}(a)=\int_0^1 a\mu(a)da=\frac{\alpha}{\alpha+\beta}. \tag{1.9}$$

When $\alpha=\beta>1$ the distribution is symmetric and bell-shaped. The distribution is skewed to the left when $\alpha>\beta>1$, and to the right when $\beta>\alpha>1$[21]. The higher α and β, the lower the variance and therefore the more precise the beliefs. We assume that ability is distributed in the population following a $Beta\left[2.5,2.5\right]$, and that the same distribution also characterizes the beliefs of the median student. This is equivalent to assume that the median student ($a=0.5$) holds correct beliefs about his/her ability, because when $\mu\left(a\right)\sim Beta\left[2.5,2.5\right]$ confidence is $\hat{\mu}(a)=0.5$.

Before analyzing the effect of over- and underconfidence let us focus on the median student in order to describe in some details the learning process. After observing the outcome, the agent updates her beliefs using Bayes rule. In particular, her posterior beliefs after observing a success are:

$$\mu(a|s)=\frac{(\psi a+1-\psi)\mu(a)}{\int_0^1(\psi a+1-\psi)\mu(a)da} \tag{1.10}$$

By contrast, if a failure was observed:

$$\hat{\mu}(a|f)=\frac{(\psi-\psi a)\mu(a)}{\int_0^1(\psi-\psi a)\mu(a)da} \tag{1.11}$$

The mass of probability is reallocated according to the realization of the signal,

[21] The Uniform is a special case of the Beta distribution when both parameters are equal to 1.

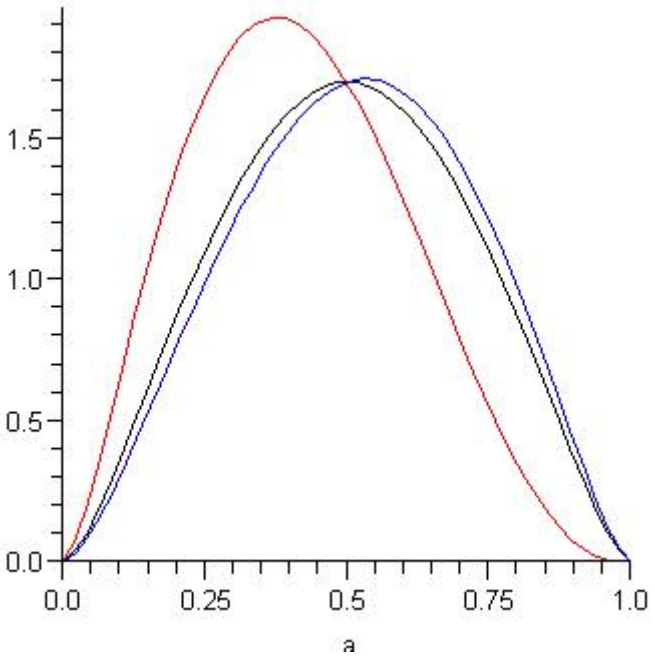


Figure 1.3: Beliefs updating of the median student after the first signal

towards the upper bound if successful (see Figure 1.3, right curve) and toward the lower bound if not (see Figure 1.3, left curve), keeping contstant the support of the density. Notice that the bad event has a stronger effect when updating beliefs[22].

The agent will then choose again the optimal track given posterior beliefs, that will be further revised after observing the outcome in the second period, and so on and so forth. The bottom line is that, within the support of initial beliefs, the distribution of beliefs changes according to the history of signals observed. Subsequent updates bring beliefs closer and closer to the true ability level as long as the agent continues receiving informative signals.

[22]The reason is that a failure is far less likely given the specification of the model. In fact, the student with correct prior beliefs will revise her confidence upward a fraction $1 - 0.5\psi$ of the times, while she will revise her confidence downward in the other 0.5ψ times. While her expected posterior confidence is always unchanged at 0.5, the upward and downward revisions would be symmetric only when $\psi = 1$, i.e. when the two events are equally likely.

1.4 Simulation

To analyze the effect of self-confidence we analyze the choices made and the human capital accumulated by an agent whose ability is always $a = 0.5$ when she holds correct prior beliefs on average $\mu(a) \sim Beta[2.5, 2.5]$, and comparing them with the counterfactuals in which she is underconfident and overconfident, respectively. In other words, we simulate the model picking up the median student and looking at the effect in her educational attainments of a wrong confidence in both directions. In fact, the higher human capital accumulated when the student is not too overconfident, i.e. when the mismatch effect does not prevail, and successful can be compensated by a probability of achieving it that is lower for two reasons. First, because the track is more difficult and therefore the same person is more likely to fail. Second, because the true ability is lower than confidence. In the utility maximization only the former is correctly internalized, and the student will therefore be successful less often than she expects. This is the engine that eventually drives her confidence towards the true level of ability.

We represent underconfidence with a distribution of prior beliefs

$$\mu(a) \sim Beta[1.5, 3] \tag{1.12}$$

skewed to the right. This implies a level of confidence $\hat{\mu}(a) = 1/3$, corresponding to the 24th percentile in the true distribution.

Similarly, overconfidence is summarized by a distribution of prior beliefs

$$\mu(a) \sim Beta[3, 1.5] \tag{1.13}$$

skewed to the left, which implies a level of confidence $\hat{\mu}(a) = 2/3$, corresponding to the 77th percentile in the true distribution. These parameters also imply that the three

distributions have roughly the same variance, and therefore that over- and underconfidence are perfectly symmetric[23]. Prior beliefs of the three different types of student are summarized in Figure 1.4. As far as the ability mismatch described in Equation 1.4 is concerned, we choose no correction in case of underconfidence ($f(m) = 0$ if $m < 0$) and a quadratic term $f(m) = 3m^2$ if $m > 0$ that implies a discount of about 7.5% in the human capital acquired in the first period by the overconfident student if successful.

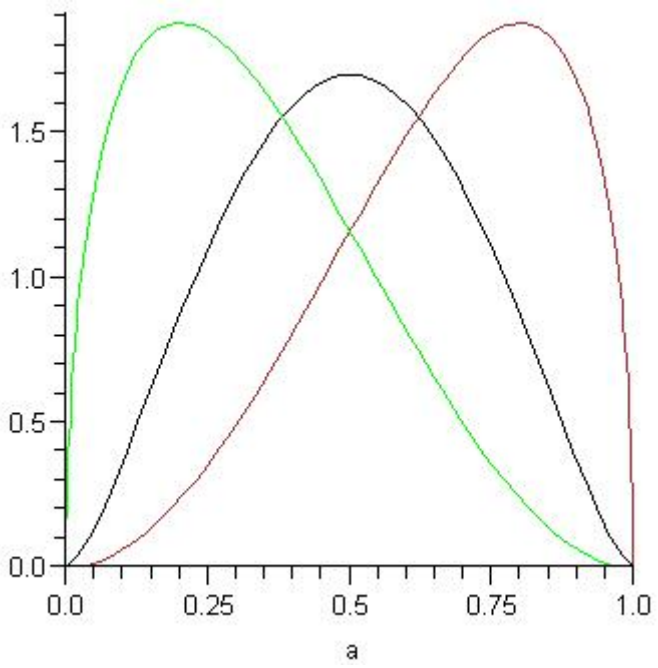


Figure 1.4: Prior beliefs given the different levels of confidence

We analize what happens to the human capital accumulated by the three types while the learning process takes place, iterating the updating of beliefs 45 times. Since the single realization of human capital relies upon a random component, we replicate the procedure 200 times.

The value of confidence slowly converges towards the true ability level for those

[23]Although the probability of success does not depend on the variance of beliefs, the latter could still affect the updating process, since the more precise the beliefs, the lower the change of confidence induced by the same signal received. We do not want the learning pattern to be affected by a different precision of beliefs, and therefore we assume the same variance in the prior distributions.

starting with a wrong prior, but the learning process is far from being completed. In fact, at the end of the 45^{th} iteration confidence is about $.425$ for the underconfident and $.558$ for the overconfident, in both cases significantly different than $.5$ ($|p| < 0.001$)[24].

Figure 1.5 displays the average gap, period by period, across repetitions, in the accumulation of human capital of the types who start with wrong priors as compared to the student starting with correct beliefs. The human capital accumulated by the underconfident is significantly lower than the human capital acquired by the student holding correct beliefs ($|p| < 0.001$), while the opposite happens for the overconfident type ($|p| < 0.001$), though the magnitude is different in absolute terms because of the cost of the mismatch $f(m)$. Notice that at the beginning, when the overconfidence is larger (and therefore also the cost of mismatching), the human capital accumulated is not much higher, while it increases as compared to the student with correct beliefs, as long as confidence converges towards the true type and the cost of mismatch decreases. Given the chosen specification of the model, the gap between the overconfident and the underconfident turns out to be about 6%.

To summarize, self-confidence can determine significant differences in the outcomes observed. When the learning process reaches the fixed point implied by discovering the true level of ability, the three types in the simulation will start making the same choices and from that moment onwards they will be observationally equivalent. However, the

[24] The speed of convergence of the two types differs a little bit. In fact, the mistake in confidence becomes significantly smaller for the over confident ($|p| = 0.038$). The reason is that the higher the track chosen, the more balanced the probability of success *given the same true level of ability* $a = 0.5$, the more informative the signal. At first glance this seems to imply that the choice of track and the educational outcomes could have been different had we internalized the different informativeness of the signals by means of dynamic optimiziation. In fact, there seems to be an additional incentive to choose a higher track thereby reducing the effect of underconfidence while increasing that of overconfidence. This is not the case, however, because such an argument holds only when the probability of success is computed holding constant the true value of ability. When choosing ψ, in contrast, agents use the best estimate of their ability $\mu(a)$. Notice that the perceived probability of success is increasing in $\mu(a)$. Hence, internalizing the different informativeness of the signal would imply a lower revision of the optimal choice at low levels of ability. In any case, maximizing utility period by period implies choices that marginally differ in terms of magnitude, and therefore a negligible mistake, particularly at low levels of ability.

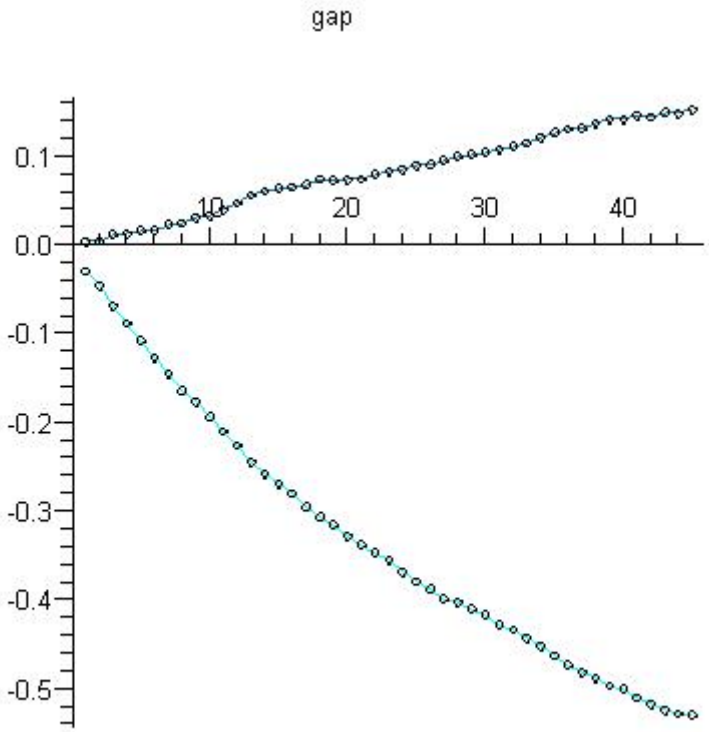


Figure 1.5: Gap in the accumulation of human capital

level of human capital acquired is and will remain significantly different. Wrong beliefs about one's ability do not need to be self-confirming to explain unequal outcomes if they lead to significantly different choices during the learning process. As long as the family background shapes children's beliefs about their ability, confidence can be a transmission mechanism that increases the intergenerational persistence of outcomes.

Notice that in the model the probability of success increases with innate ability only, while the human capital accumulated plays no role. As already noticed in Section 1.2.4, this simplifying assumption downplays the role of nurture, since achievements are also determined by the whole history of intermediate outcomes, in turn also driven by self-confidence, as well as by the environment in which the children grow. Therefore, what found by the model is once more a lower bound of the role of self-confidence,

since the cumulative effect of the gap in the human capital accumulated during the learning process of one's ability is not taken into account. The role of nurture therefore implies that tests meant to measure students' ability are instead capturing also the gap in human capital accumulated up to that point because of a different family background. For instance, a centralized test administered at age 15 in order to select students into different tracks would probably classify as different two students characterized by the same innate ability but with a different background, thereby helping to perpetuate intergenerational inequalities. A policy implication arising from the model is therefore that cognitive tests should take place as early as possible in order to endow parents with measures of the innate level of ability of the children that are not confounded with the role that the family background can play through self-confidence among the several ways.

1.5 Conclusions

In line with some recent contributions, we claim that the socio-economic background affects not only the actual stock of cognitive skills possessed by a child (innate ability) but also the beliefs about such (unobserved) cognitive skills. There is indeed a vast literature supporting the hypothesis that people have imperfect knowledge of their ability and that many personality traits related to the concept of self-confidence are influenced by the family background in which a child grows up.

We provide further suggestive evidence about the link between confidence and family background using two very different sources: the PISA datasets, which is a representative cross-national survey of 15-year old pupils, and a very homogeneous dataset of students from Bocconi University surveyed at a later stage of their career. We show that in both samples the link between confidence and background is strong,

and survives the inclusion of good controls of unobserved and observed ability. Our proxies of ability are likely to bias downward the estimated link between confidence and background, since they capture not only innate ability but also the gap in human capital that has been accumulated up to that point.

We then propose a model in which fully rational agents, who maximize the expected acquisition of human capital, choose tasks according to their perceived ability. True ability and the difficulty of the chosen track affect the probability of success. After observing whether they succeed or not, students update their beliefs, fully exploiting the available information, following Bayes' rule. We simulate the model with a bootstrapping procedure and we show that choices distorted by over- and under-confidence lead to a significant gap in the accumulation of human capital during the process in which agents eventually learn their true level of ability.

In our model agents do not derive additional utility by holding a good self-image; the consequence of this assumption is that if a perfectly informed and benevolent planner could force individuals to choose the "right" task, the effect of wrong confidence would disappear. Nevertheless, even in a setting in which agents are fully rational and have standard preferences, a moderate degree of over-confidence can be beneficial in terms of the accumulation of human capital over the life course, although at a price of a lower utility (since overconfident and underconfident agents do not make, by construction, utility-maximizing choices). Underconfidence, on the other hand, is suboptimal in terms of both utility maximization and human capital accumulation.

The intergenerational transmission of beliefs can thus constitute a further channel through which socio-economic differences perpetuate from one generation to the other because, even if two individuals had the same innate cognitive ability, differences in beliefs would lead them to make different choices in terms of investment in education. The results of our analysis suggest that policy interventions aimed at providing early and

precise feedbacks about the cognitve skills of children from disadvantaged backgrounds can be beneficial in helping to narrow the gaps in educational attainments, by avoiding that equally talented people make different choices only because they have inherited different beliefs about their potential.

Chapter 2

Evaluating Students' Evaluations of Professors*

2.1 Introduction

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

From the point of view of the university administration, such evaluations are used to solve the agency problems related to the selection and motivation of teachers, in a context in which neither the types of teachers, nor their levels of effort, can be observed precisely. In fact, students' evaluations are often used to inform hiring and promotion decisions (Becker and Watts, 1999) and, in institutions that put a strong emphasis on

*This chapter is the result of joint work with Michela Braga and Michele Pellizzari

research, to avoid strategic behavior in the allocation of time or effort between teaching and research activities Brown and Saks (1987).[1]

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

Measuring teaching quality is complicated also because the most common observable teachers' characteristics, such as their qualifications or experience, appear to be relatively unimportant (Hanushek and Rivkin, 2006; Krueger, 1999; Rivkin, Hanushek, and Kain, 2005). Despite such difficulties, there is also ample evidence that teachers' quality matters substantially in determining students' achievement (Carrell and West, 2010; Rivkin, Hanushek, and Kain, 2005) and that teachers respond to incentives (Duflo, Hanna, and Ryan, 2010; Figlio and Kenny, 2007; Lavy, 2009). Hence, understanding how professors should (or should not) be monitored and incentivized is of primary importance.

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

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

them with objective measures of teacher effectiveness. We construct such measures by comparing the performance in subsequent coursework of students who are randomly allocated to different teachers in their compulsory courses. For this exercise we use data about one cohort of students at Bocconi University - the 1998/1999 freshmen - who were required to take a fixed sequence of compulsory courses and who where randomly allocated to a set of teachers for each of such courses. Additionally, the data are exceptionally rich in terms of observable characteristics, in particular they include measures of cognitive ability, family income and entry wages, which are obtained from regular surveys of graduates.[2]

We find that, even in a setting where the syllabuses are fixed and all teachers in the same course present exactly the same material, professors still matter substantially. The average difference in subsequent performance between students who were assigned to the best and worst teacher (on the effectiveness scale) is approximately 23% of a standard deviation in the distribution of exam grades, corresponding to about 3% of the average grade. This effect translates into approximately 1.4% of the average entry wage or 14 euros per month (160-200 euros per year). Moreover, our measure of teaching quality appears to be negatively correlated with the students' evaluations of the professors: in other words, teachers who are associated with better subsequent performance receive worst evaluations from their students. On the other hand, teachers who are associated with high grades in their own exams receive better evaluations.

We rationalize these results with a simple model, where good teachers are those who provide their students with knowledge that is useful in future learning and, at the same time, require high effort from their students. Students are heterogeneous in their disutility of effort, which is higher for the least able ones, and evaluate professors on the basis of their realized utility, which depends on grades/learning and effort. In this setting,

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

students in the bottom part of the ability distribution may, in fact, give worse evaluations to the good teachers, who impose a high effort cost on them, than the bad teachers.

Consistently with these predictions, we also find that the evaluations of classes in which high skill students (identified by their score in the cognitive admission test) are over-represented are more in line with the estimated real teacher quality. Furthermore, the distributions of grades in the classes of the most effective teachers are more dispersed, a piece of evidence that lends support to our specification of the learning function. Additionally, in order to support our assumption that evaluations are based on students' realized utility, we match our data with the weather conditions observed on the exact days when students filled the evaluation questionnaires. Under the assumption that the weather affects utility and not teaching quality, finding that the students' evaluations react to meteorological conditions lends support to the specification of our model.[3] Our results show that students evaluate professors more negatively on rainy and cold days.

There is a large literature that investigates the role of teacher quality and teacher incentives in improving educational outcomes, although most of the existing studies focus on primary and secondary schooling (Figlio and Kenny, 2007; Jacob and Lefgren, 2008; Kane and Staiger, 2008; Rivkin, Hanushek, and Kain, 2005; Rockoff, 2004; Rockoff and Speroni, 2010; Tyler, Taylor, Kane, and Wooten, 2010). The availability of standardized test scores facilitates the evaluation of teachers in primary and secondary schools and such tests are currently available in many countries and also across countries (Mullis, Martin, Robitaille, and Foy, 2009; OECD, 2010). The large degree of heterogeneity in subjects and syllabuses in universities makes it very difficult to design common tests that would allow to compare the performance of students who were

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

exposed to different teachers, especially across subjects. At the same time, the large increase in college enrollment experienced in almost all countries around the world in the past decades (OECD, 2008) calls for a specific focus on higher education, as in this study.[4]

To the best of our knowledge, only three other papers investigate the role of students' evaluations in university, namely Carrell and West (2010); Hoffman and Oreopoulos (2009); Weinberg, Fleisher, and Hashimoto (2009). Compared to these papers we improve in various directions. First of all, the random allocation of students to teachers in our setting differentiates our approach from that of Hoffman and Oreopoulos (2009) and Weinberg, Fleisher, and Hashimoto (2009), who cannot purge their estimates from the potential bias due to the best students selecting the courses of the best professors. Rothstein (2009) and Rothstein (2010) show that correcting such a selection bias is pivotal to producing reliable measures of teaching quality.

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

[4]On average in the OECD countries 56% of school-leavers enrolled in tertiary education in 2006 versus 35% in 1995. The same secular trends appear in non-OECD countries. Further, the number of students enrolled in tertiary education has increased on average in the OECD countries by almost 20% between 1998 and 2006, with the US having experienced a higher than average increase from 13 to 17 millions.

data. Furthermore, we show that our theory is consistent with additional pieces of evidence and we use it to formulate policy proposals.

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

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

More generally, this paper is also related and contributes to the wider literature on performance measurement and performance pay. For example, one concern with the students' evaluations of teachers is that they might divert professors from activities that have a higher learning content for the students (but that are more demanding in terms of students' effort) and concentrate more on classroom entertainment (popularity contests)

[5]Chetty, Friedman, Hilger, Saez, Schanzenbach, and Yagan (2011) present some results in this same spirit but in a very different context (kindergarten) and without explicitly looking at measures of teaching quality (they rather consider teachers' experience).

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

or change their grading policies. This interpretation is consistent with the view that teaching is a multi-tasking job, which makes the agency problem more difficult to solve (Holmstrom and Milgrom, 1994). Subjective evaluations, which have become more and more popular in modern human resource practices, can be seen as a mean to address such a problem and, given the very limited extant empirical evidence (Baker, Gibbons, and Murphy, 1994; Prendergast and Topel, 1996), our results can certainly inform also this area of the literature.

The paper is organized as follows. Section 2.2 describes our data and the institutional details of Bocconi University. Section 2.3 presents our strategy to estimate teacher effectiveness and shows the results. In Section 2.4 we correlate teacher effectiveness with the students' evaluations of professors. Robustness checks are reported in Section 2.5. In Section 2.6 we present a simple theoretical framework that rationalizes our results, while Section 2.7 discusses some additional evidence that corroborates our model. Finally, Section 2.8 concludes.

2.2 Data and institutional details

The empirical analysis in this paper is based on data for one enrollment cohort of undergraduate students at Bocconi university, an Italian private institution of tertiary education offering degree programs in economics, management, public policy and law. We select the cohort of the 1998/1999 freshmen for technical reasons, being the only one available in our data where students were randomly allocated to teaching classes for each of their compulsory courses.[7]

[7]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

In later cohorts, the random allocation was repeated at the beginning of each academic year, so that students would take all the compulsory courses of each academic year with the same group of classmates, which only permits to identify the joint effectiveness of the entire set of teachers in each academic year.[8] For earlier cohorts the class identifiers, which are the crucial piece of information for our study, were not recorded in the university archives.

The students entering Bocconi in the 1998/1999 academic year were offered 7 different degree programs, although only three of them attracted a sufficient number of students to require the splitting of lectures into more than one class: Management, Economics and Law&Management[9]. Students in these programs were required to take a fixed sequence of compulsory courses that span the entire duration of their first two years, a good part of their third year and, in a few cases, also their last year. Table 2.1 lists the exact sequence for each of the three programs that we consider, breaking down courses by the term (or semester) in which they were taught and by subject areas (classified with different colors: red for management, black for economics, green for quantitative subjects, blue for law).[10]. In Section 2.3 we construct measures of teacher effectiveness for the professors of these compulsory courses. We do not consider elective subjects, as the endogenous self-selection of students would complicate the analysis.

allocated groups were kept for both regular lectures and applied classes. Hence, in the remainder of the paper we use the two terms interchangeably.

[8]De Giorgi, Pellizzari, and Woolston (2011) use data for these later cohorts for a study of class size.

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

[10]Notice that Economics and Management share exactly the same sequence of compulsory courses in the first three terms. Indeed, students in these two programs did attend these courses together and made a final decision about their major at the end of the third term. De Giorgi, Pellizzari, and Redaelli (2010) study precisely this choice. In the rest of the paper we abstract from this issue and we treat the two degree programs as entirely separated. We present some robustness checks to justify this approach in Section 2.3.

Table 2.1: Structure of degree programs

	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.

Most (but not all) of the courses listed in Table 2.1 were taught in multiple classes (see Section 2.3 for details). The number of such classes varied across both degree programs and specific courses. For example, Management was the program that attracted the most students (over 70% in our cohort), who were normally divided into 8 to 10 classes. Economics and Law&Management students were much fewer and were rarely allocated to more than just two classes. Moreover, the number of classes also varied within degree programs depending on the number of available teachers in each course. For instance, in 1998/99 Bocconi did not have a law department and all law professors were contracted from other nearby universities. Hence, the number of classes in law courses were normally fewer than in other subjects. Similarly, since

the management department was (and still is) much larger than the economics or the mathematics department, courses in the management areas were normally split in more classes than courses in other subjects.

Regardless of the specific class to which students were allocated, they were all taught the same material. In other words, all professors of the same course were required to follow exactly the same syllabus, although some variations across degree programs were allowed (i.e. mathematics was taught slightly more formally to Economics students than Law&Management ones).

Additionally, the exam questions were also the same for all students (within degree program), regardless of their classes. Specifically, one of the teachers in each course (normally a senior person) acted as a coordinator, making sure that all classes progressed similarly during the term, defining changes in the syllabus and addressing specific problems that might have arisen. The coordinator also prepared the exam paper, which was administered to all classes. Grading was usually delegated to the individual teachers, each of them marking the papers of the students in his/her own class, typically with the help of one or more teaching assistants. Before communicating the marks to the students, the coordinator would check that there were no large discrepancies in the distributions across teachers. Other than this check, the grades were not curved, neither across nor within classes.

Table 2.2 reports some descriptive statistics that summarize the distributions of (compulsory) courses and their classes across terms and degree programs. For example, in the first term Management students took 3 courses, divided into a total of 24 different classes: management I, which was split into 10 classes; private law, 6 classes; mathematics, 8 classes. The table also reports basic statistics (means and standard deviations) for the size of these classes.

Table 2.2: Descriptive statistics of degree programs

	Term							
Variable	I	II	III	IV	V	VI	VII	VIII
				Management				
No. Courses	3	3	3	4	3	4	1	-
No. Classes	24	21	23	26	23	27	12	-
Avg. Class Size	129.00	147.42	134.61	138.62	117.52	133.48	75.08	-
SD Class Size	73.13	80.57	57.46	100.06	16.64	46.20	11.89	-
				Economics				
No. Courses	3	3	3	3	2	1	-	-
No. Classes	24	21	23	4	2	2	-	-
Avg. Class Size	129.00	147.42	134.61	98.25	131.00	65.5	-	-
SD Class Size	73.13	80.57	57.46	37.81	0	37.81	-	-
				Law & Management				
No. Courses	3	4	4	4	2	-	-	1
No. Classes	5	5	5	6	3	-	-	1
Avg. Class Size	104.40	139.20	139.20	116.00	116.00	-	-	174.00
SD Class Size	39.11	47.65	47.67	44.96	50.47	-	-	0.00

Our data cover in details the entire academic history of the students in these programs, including their basic demographics (gender, place of residence and place of birth), high school leaving grades as well as the type of high school (academic or technical/vocational), the grades in each single exam they sat at Bocconi together with the date when the exams were sat. Graduation marks are observed for all non-dropout students.[11] Additionally, all students took a cognitive admission test as part of their application to the university and such test scores are available in our data for all the students. Moreover, since tuition fees varied with family income, this variable is also recorded in our dataset. Importantly, we also have access to the random class identifiers that allow us to identify in which class each students attended each of their courses.

Table 2.3 reports some descriptive statistics for the students in our data by degree

[11] The dropout rate, defined as the number of students who, according to our data, do not appear to have completed their programs at Bocconi over the total size of the entering cohort, is just above 10%. Notice that some of these students might have transfered to another university or still be working towards the completion of their program, whose formal duration was 4 years. In Section 2.5 we perform a robustness check to show that excluding the dropouts from our calculations is irrelevant for our results.

program. The vast majority of them were enrolled in the Management program (74%), while Economics and Law&Management attracted 11% and 14%. Female students were generally under-represented in the student body (43% overall), apart from the degree program in Law&Management. About two thirds of the students came from outside the province of Milan, which is where Bocconi is located, and such a share increased to 75% in the Economics program. Family income was recorded in brackets and one quarter of the students were in the top bracket, whose lower threshold was in the order of approximately 110,000 euros at current prices. Students from such a wealthy background were under-represented in the Economics program and over-represented in Law&Management. High school grades and entry test scores (both normalized on the scale 0-100) provide a measure of ability and suggest that Economics attracted the best students, a fact that is confirmed by looking at university grades, graduation marks and entry wages in the labor market.

Data on wages come from graduate surveys that we were able to match with the administrative records. Bocconi runs regular surveys of all alumni approximately one to one and a half years since graduation. These surveys contain a detailed set of questions on labor market experience, including employment status, occupation, and (for the employed) entry wages. As it is common with survey data, not all contacts were successful but we were still able to match almost 60% of the students in our cohort, a relatively good response rate for surveys[12].

[12]The response rates are highly correlated with gender, because of compulsory military service, and with the graduation year, given that Bocconi has improved substantially over time in its ability to track its graduates. Until the 1985 birth cohort, all Italian males were required to serve in the army for 10-12 months but were allowed to postpone the service if enrolled in full time education. For college students, it was customary to enroll right after graduation.

Table 2.3: Descriptive statistics of students

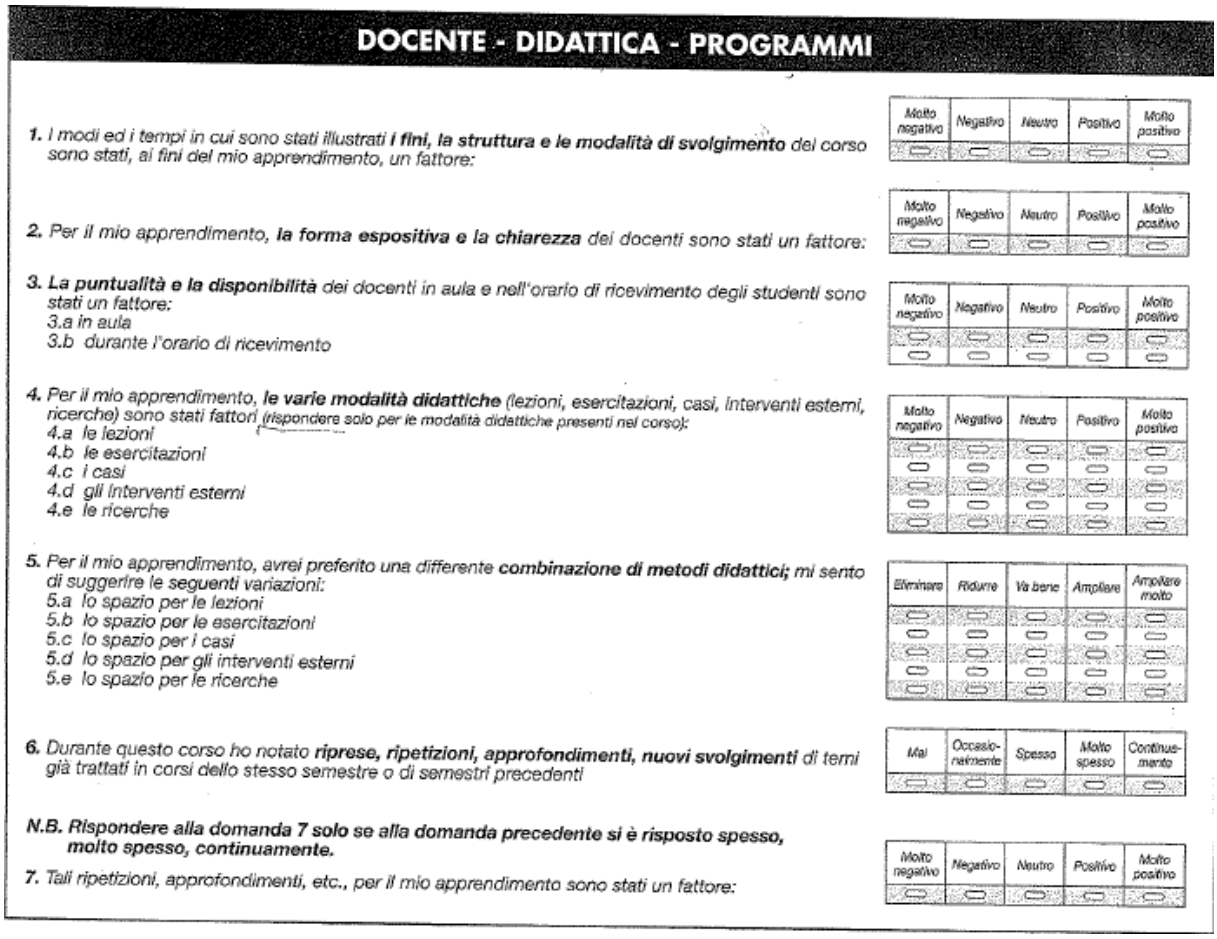
Variable	Management	Economics	Law &Manag.	Total
1=female	0.408	0.427	0.523	0.427
1=outside Milan[a]	0.620	0.748	0.621	0.634
1=top Income Bracket[b]	0.239	0.153	0.368	0.248
1=academic high school[c]	0.779	0.794	0.684	0.767
1=late enrollee[d]	0.014	0.015	0.011	0.014
High-school grade (0-100)	86.152	93.053	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]	966.191	1,012.241	958.381	967.964
	(260.145)	(265.089)	(198.437)	(250.367)
Number of students	901	131	174	1,206

[a] Dummy equal to one if the student's place of residence at the time of first enrollment is outside the province of Milan (which is where Bocconi university is located).
[b] Family income is recorded in brackets and the dummy is equal to one for students who report incomes in the top bracket, whose lower threshold is in the order of approximately 110,000 euros at current prices.
[c] Dummy equal to one if the student attended a academic high school, such as a lyceum, rather than professional or vocational schools.
[d] Dummy equal to one if the student enrolled at Bocconi after age 19.
[e] Nominal value at current (2010) prices. Based on 391 observations for Management, 36 observations for Economics, 94 observations for Law&Management, i.e. 521 observations overall.

Finally, we complement our dataset with students' evaluations of teachers. Towards the end of each term (typically in the last week), students in all classes were asked to fill an evaluation questionnaire during one lecture. The questions gathered students' opinions about various aspects of the teaching experience, including the clarity of the lectures, the logistics of the course, the availability of the professor and so on. For each item in the questionnaire, students answered on a scale from 0 (very negative) to 10 (very positive) or 1 to 5.

In order to allow students to evaluate their experience without fear of retaliation from the teachers at the exam, such questionnaires are anonymous and it is impossible to match the individual student with a specific evaluation of the teacher. However,

each questionnaire reports the name of the course and the class identifier, so that we can attach average evaluations to each class in each course. Figure 2.1 shows, as an example, the first page of the evaluation questionnaire used in the academic year 1998-1999.[13]



DOCENTE - DIDATTICA - PROGRAMMI

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

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

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

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

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

Figure 2.1: Excerpt of student questionnaire

In Table 2.4 we present some descriptive statistics of the answers to the evaluation questionnaires. We concentrate on a limited set of items, which we consider to be the most informative and interesting, namely overall teaching quality, lecturing clarity,

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

the teacher's ability to generate interest in the subject, the logistic of the course and workload. These are the same items that we analyze in more details in Section 2.4. The exact wording and scaling of the questions are reported in Table 2.6.

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

Table 2.4: Descriptive statistics of students' evaluations

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

[a] Scores range from 0 to 10.
[b] Scores range from 1 to 5.
[c] Number of collected valid questionnaires over the number of officially enrolled students.
See Table 2.6 for the exact wording of the evaluation questions.

Some of the evaluation items are, understandably, highly correlated. For example,

the correlation coefficient between overall teaching quality and lecturing clarity is 0.89. The course logistics and the ability of the teacher in generating interest for the subject are slightly less strongly correlated with the core measures of teacher quality (around 0.5-0.6). Workload is the least correlated with any other item (all correlation coefficients are below 0.2). The full correlation matrix is reported in Table 2.5.

Table 2.5: Correlations between evaluations items

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

Additionally, in Table 2.4 we also report the mean and standard deviations of the number of collected questionnaires and the number of officially enrolled students in each of class. One might actually be worried that students may drop out of a class in response to the quality of the teaching so that at the end of the course, when questionnaires are distributed only the students who liked the teacher are eventually present. Such a process would lead to a compression of the distribution of the evaluations, with good teachers being evaluated by their entire class (or by a majority of their allocated students) and bad teachers being evaluated only by a subset of students who particularly liked them. The descriptive statistics reported in Table 2.4 seem to indicate that this is not a major issue, as on average the number of collected questionnaires is around 80% of the total number of enrolled students (the median is very similar). Moreover, when we

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

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

Table 2.6: Wording of the evaluation questions

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

2.2.1 The random allocation

In this section we present evidence that the random allocation of students into classes was successful. De Giorgi, Pellizzari, and Redaelli (2010) use data for the same cohort (although for a smaller set of courses and programs) and provide similar evidence.

The randomization was (and still is) performed via a simple random algorithm that assigned a class identifier to each student, who were then instructed to attend the lectures for the specific course in the class labeled with the same identifier. The university administration adopted the policy of repeating the randomization for each course with the explicit purpose of encouraging wide interactions among the students.

Table 2.7 is based on test statistics derived from probit (columns 1,2,5,6,7) or OLS (columns 3 and 4) regressions of the observable students' characteristics (by column) on class dummies for each course in each degree program that we consider. The null hypothesis under consideration is the joint significance of the coefficients on the class dummies in each model, which amounts to testing for the equality of the means of the observable variables across classes. The table shows some descriptive statistics about the distribution of p-values for such tests.

The mean and median p-values are in all cases far from the conventional thresholds of 5% or 1% and only in a very few instances the null cannot be rejected by the data. The most notable exception is residence from outside Milan, which is abnormally low in two Management groups. Overall, Table 2.7 suggests that the randomization was rather successful.

Testing the equality of means is not a sufficient test of randomization for continuous variables. Hence, in Figure 2.2 we compare the distributions of our measures of ability (high school grades and entry test scores) for the entire student body and for a randomly

Table 2.7: 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
minimum	0.049	0.055	0.012	0.004	0.037	0.000	0.025
maximum	0.994	0.949	0.991	0.944	0.947	0.824	0.970
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
minimum	0.006	0.077	0.000	0.011	0.280	0.228	0.355
maximum	0.950	0.993	0.918	0.989	0.989	0.944	0.991
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
minimum	0.022	0.168	0.145	0.182	0.291	0.138	0.935
maximum	0.972	0.966	0.977	0.847	0.999	0.880	0.961
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 2.3.

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

selected class in each program. The figure evidently shows that the distributions are extremely similar and formal Kolmogorov-Smirnov tests confirm the visual impression.

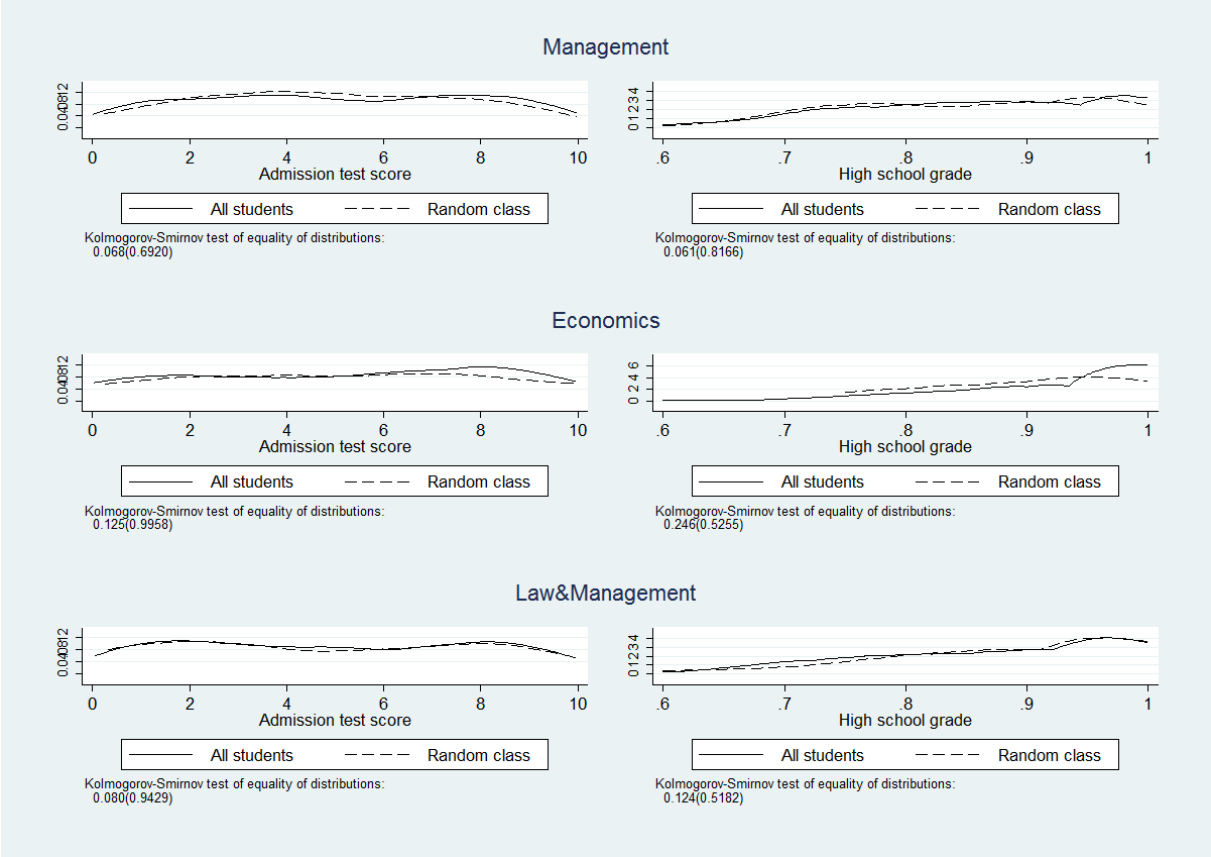


Figure 2.2: Evidence of random allocation - Ability variables

Even though students were randomly assigned to classes, one may still be concerned about teachers being selectively allocated to classes. Although no explicit random algorithm was used to assign professors to classes, for obvious organizational reasons that was (and still is) done in the Spring of the previous academic year, i.e. well before students were allowed to enroll, so that even if teachers were allowed to choose their class identifiers they would have no chance to know in advance the characteristics of the students who would be given that same identifier.

More specifically, there used to be (and still is) a very strong hysteresis in the

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

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

In order to avoid any distortion in our estimates of teaching effectiveness due to the more or less convenient teaching times, we collected detailed information about the exact timing of the lectures in all the classes that we consider, so that we can hold this specific factor constant (see Section 2.3). Additionally, we also know in which exact room each class was taught and we further conditions on the characteristics of the classrooms, namely the building and the floor where they are located. There is no variation in other features of the rooms, such as the furniture (all rooms were - and still are - fitted with exactly the same equipment: projector, computer, white-board) or the orientation (all rooms face the inner part of the campus where there is very limited car

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

traffic).[15]

Table 2.8 provides evidence of the lack of correlation between teachers and classes' characteristics, namely we show the results of regressions of teachers' observable characteristics on classes' observable characteristics. For this purpose, we estimate a system of 9 seemingly unrelated simultaneous equations, where each observation is a class in a compulsory course. The dependent variables are 9 teachers' characteristics (age, gender, h-index, average citations per year and 4 dummies for academic positions) and the regressors are the class characteristics listed in the rows of the table.[16] The reported statistics test the null hypothesis that the coefficients on each class characteristic are all jointly equal to zero in all the equations of the system.[17]

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

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

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

Table 2.8: Randomness checks - Teachers

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

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

[a] Number or officially enrolled students.

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

[c] See notes to Table 2.3.

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

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

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

[g] Test of the joint significance of 4 floor dummies.

[h] Dummy for building A.

Results show that only the time of the lectures is significantly correlated with the teachers' observables at conventional statistical levels. In fact, this is one of the few elements of the teaching planning over which teachers had some limited choice. More specifically, professors are given a suggested time schedule for their classes in the spring of the previous academic year (usually based on the schedule of the current year), and they can either approve it or request changes. The administration, then,

accommodates such changes only if they are compatible with the other many constraints in terms of rooms availability and course overlappings. In our empirical analysis we do control for all the factors in Table 2.8, so that our measures of teaching effectiveness are purged from the potential confounding effect of teaching times on students' learning.

2.3 Estimating teacher effectiveness

We use performance data for our students to estimate measures of teacher effectiveness. Namely, for each of the compulsory courses listed in Table 2.1 we compare the future outcomes of students that attended those courses in different classes, under the assumption that students who were taught by better professors enjoyed better outcomes later on. This approach is similar to the *value-added* methodology that is more commonly used in primary and secondary schools (Goldhaber and Hansen, 2010; Hanushek, 1979; Hanushek and Rivkin, 2006, 2010; Rivkin, Hanushek, and Kain, 2005; Rothstein, 2009) but it departs from its standard version, that uses contemporaneous outcomes and conditions on past performance, since we use future performance to infer current teaching quality.[18]

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

We avoid these complications by exploiting the random allocation of students in our cohort to different classes for each of their compulsory courses. For this same reason, we focus exclusively on compulsory courses, as self-selection is an obvious concern

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

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

We compute our measures of teacher effectiveness in two steps. First, we estimate the conditional mean of the future grades (in compulsory courses) of students in each class according to the following procedure. Consider a set of students enrolled in degree program d and indexed by $i = 1, \ldots, N_d$, 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 $t \geq t_c$.

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

$$y_{ids\zeta} = \alpha_{dcs} + \beta X_i + \epsilon_{ids\zeta} \tag{2.1}$$

where X_i is a vector of student-level characteristics including a gender dummy, a dummy for whether the student is in the top income bracket, the entry test score and the high school leaving grade. The α's are our parameters of interest and they measure the conditional means of the future grades of students in class s: high values of α indicate that, on average, students attending course c in class s performed better (in subsequent courses) than students taking course c in a different class. The random

allocation procedure guarantees that the class fixed effects α_{dcs} in equation 2.1 are purely exogenous and identification is straightforward.[19]

Notice that, since in general there are several subsequent courses ζ for each course c, each student is observed multiple times and the error terms $\epsilon_{ids\zeta}$ are serially correlated within i and across ζ. We address this issue by adopting a standard random effect model to estimate all the equations 2.1 (we estimate one such equation for each course c). Moreover, we further allow for cross-sectional correlation among the error terms of students in the same class by clustering the standard errors at the class level.

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

$$\epsilon_{ids\zeta} = \upsilon_i + \omega_s + \nu_{ids\zeta} \tag{2.2}$$

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 2.1.[20]

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

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

[20]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, also used by the command *xtreg* in Stata Swamy and Arora (1972).

the additional serial correlation induced by the term ω_s.

Overall, we are able to estimate 230 such fixed effects, the large majority of which are for Management courses.[21]. Descriptive statistics of the estimated α's are reported in Table 2.9.

Table 2.9: Descriptive statistics of estimated class effects

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

The second step of our approach is meant to purge the estimated α's from the effect of other class characteristics that might affect the performance of students in later courses but are not attributable to teachers. By definition, the class fixed effects capture all those features, both observable and unobservable, that are fixed for all students in the class. These certainly include teaching quality but also other factors that are documented to be important ingredients of the education production function, such as class size and class composition De Giorgi, Pellizzari, and Woolston (2011).

A key advantage of our data is that most of these other factors are observable. In particular, based on our academic records we can construct measures of both class

[21] We cannot run equation 2.1 for courses that have no contemporaneous nor subsequent courses, such as Corporate Strategy for Management, Banking for Economics and Business Law for Law&Management (see Table 2.1). For such courses, the set Z_c is empty. Additionally, some courses in Economics and in Law&Management are taught in one single class, for example Econometrics (for Economics students) or Statistics (for Law&Management). For such courses, we have $S_c = 1$. The evidence that we reported in Tables 2.7 and 2.8 also refer to the same set of 230 classes.

size and class composition (in terms of students' characteristics). Additionally, we also have access to the identifiers of the teachers in each class and we can recover a large set of variables like gender, tenure status, and measures of research output. We also know which of the several teachers in each course acted as coordinator. These are the same teacher characteristics that we used in Table 2.8. Once we condition on all these observable controls, unobservable teaching quality is likely to be the only remaining factor that generates variation in the estimated α's. At a minimum, it should be uncontroversial that teaching quality is by far the single most important unobservable that generates variation in the $\widehat{\alpha}$'s, once conditioning on the observables.

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

Thus, in Table 2.10 we regress the estimated α's on all observable class and teacher characteristics. In column 1 we condition only on class size and class composition, in column 2 only on information about the teachers and in column 3 we combine the two sets of controls. In all cases we weight observations by the inverse of the standard error of the estimated α's to take into account differences in the precision of such estimates. Consistently with previous studies on the same data (De Giorgi, Pellizzari, and Woolston, 2011), we find that larger classes tend to be associated with worse learning outcomes, that classes with more able students, measured with either high school grades or the entry test score, also perform better and that a high concentration

Table 2.10: Determinants of class effects

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

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

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

[b] See notes to Table 2.8.

[c] All variables regarding the academic position refer to the main teacher of the class.

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

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

of high income students appears to be detrimental for learning. Overall, observable class characteristics explain about 8% of the variation in the estimated α's within degree program, term and subject cells, where subjects are defined as in Table 2.1.[22]

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

Our final measures of teacher effectiveness are the residuals of the regression of the estimated α's on all the observable variables, i.e the regression reported in column 3 of Table 2.10. In Table 2.11 we present descriptive statistics of such measures.

The overall standard deviation of teacher effectiveness is 0.086.[23] This average is the composition of a larger variation among the courses of the program in Economics (0.159) and a more limited variation in Management (0.069) and Law & Management (0.019). Recall that grades are normalized so that the distributions of the class effects are comparable across courses. Hence, these results can be directly interpreted

[22]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 are partialled out.

[23]The standard deviation that we consider is the OLS estimate of the residuals of the regression in column 3 of Table 2.10

Table 2.11: Descriptive statistics of estimated teacher effectiveness

	Management	Economics	Law & Management	Total
PANEL A: Std. dev. of estimated teacher effect				
mean	0.069	0.159	0.019	0.086
minimum	0.041	0.030	0.010	0.010
maximum	0.106	0.241	0.030	0.241
PANEL B: Largest minus smallest class effect				
mean	0.190	0.432	0.027	0.230
minimum	0.123	0.042	0.014	0.014
maximum	0.287	0.793	0.043	0.043
No. of courses	20	11	7	38
No. of classes	144	72	14	230

Teacher effectiveness is estimated by regressing the estimated class effects (α) on observable class and teacher's characteristics (see Table 2.10).

in terms of changes in outcomes. In other words, the overall effect of increasing teacher effectiveness by one standard deviation is an increase in the average grade of subsequent courses by 0.086 standard deviations, roughly 0.3 of a grade point or 1.1% over the average grade of approximately 26.[24] Given an estimated conditional elasticity of entry wages to GPA of 0.45, such an effect would cost students slightly more than 0.5% of their average entry monthly wage of 967 euros, or about 60 euros per year.[25] Since in our data we only observe entry wages, it might well be that the long term effects of teaching quality are even larger.

In Table 2.11 we also report the standard deviations of teacher effectiveness of

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

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

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

In the lower panel of Table 2.11 we show the mean (across courses) of the difference between the largest and the smallest indicators of teacher effectiveness, which allows us to compute the effect of attending a course in the class of the best versus the worst teacher. On average, this effect amounts to 0.230 of a standard deviation, that is almost 0.8 grade points or 3% over the average grade. As already noted above, this average effect masks a large degree of heterogeneity across subjects ranging from almost 80% to a mere 4% of a standard deviation.

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

For robustness and comparison, we estimate the class effects in two alternative ways. First, we restrict the set Z_c to courses belonging to the same subject area of course c, under the assumption that good teaching in one course is likely to have a stronger effect on learning in courses of the same subject areas (e.g. a good basic mathematics teacher is more effective in improving students performance in financial

mathematics than in business law). The subject areas are defined by the colors in Table 2.1 and correspond to the department that was responsible for the organization and teaching of the course. We label these estimates *subject* effects. Given the more restrictive definition of Z_c we can only produce these estimates for a smaller set of courses and using fewer observation, which is the reason why we do not take them as our benchmark.

Next, rather than using performance in subsequent courses, we run equation 2.1 with the grade in the same course c as the dependent variable. We label these estimates *contemporaneous* effects[26]. We do not consider these contemporaneous effects as alternative and equivalent measures of teacher effectiveness, but we will use them to show that they correlate very differently with the students' evaluations. Descriptive statistics for the subject and contemporaneous effects are reported in Tables 2.12 and 2.13.

Table 2.12: Descriptive statistics of *subject* teacher effectiveness

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

In Table 2.14 we investigate the correlation between these alternative estimates of

[26]When estimating *contemporaneous* effects we include past grades in the vector of student-level characteristics of equation 2.1

Table 2.13: Descriptive statistics of *contemporaneous* teacher effectiveness

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

teacher effectiveness. Specifically, we report results from weighted OLS regressions with our benchmark estimates as the dependent variable and, in turn, the subject and the contemporaneous effects on the right hand side, together with dummies for degree program, term and subject area[27]. Reassuringly, the subject effects are positively and significantly correlated with our benchmark, while the contemporaneous effects are negatively and significantly correlated with our benchmark, a result that is consistent with previous findings (Carrell and West, 2010; Krautmann and Sander, 1999; Weinberg, Fleisher, and Hashimoto, 2009) and to which we will return in Section 2.4.

Finally, figure 2.3 show the correlation of estimated teacher effects for courses common to Economics and Management, thus justifying our decision to treat the two degree programs as entirely separated.

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

Table 2.14: Comparison of benchmark, subject and contemporaneous teacher effects

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

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

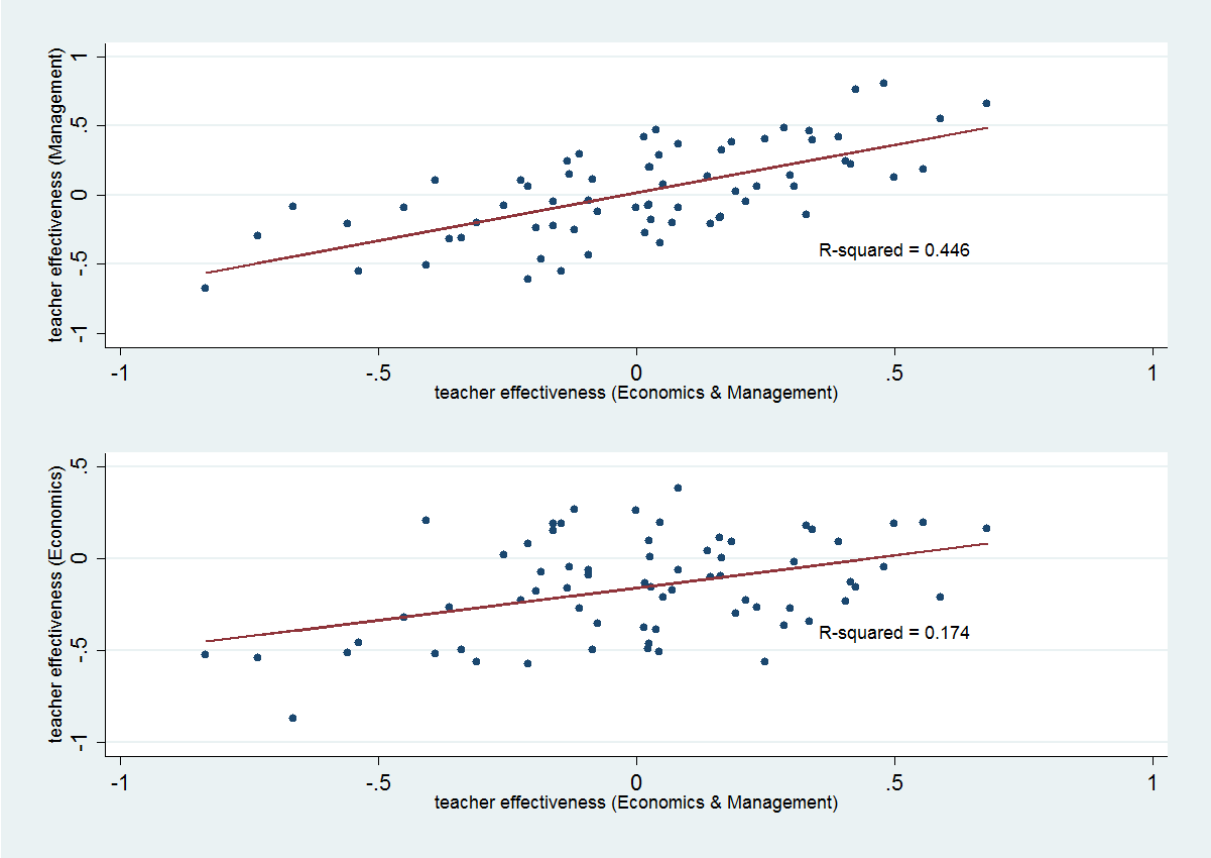


Figure 2.3: Economics and Management common courses - Benchmark teacher effectiveness

2.4 Correlating teacher effectiveness and students' evaluations

In this section we investigate the relationship between our measures of teaching effectiveness from Section 2.3 and the evaluations teachers receives from their students. We concentrate on two core items from the evaluation questionnaires, namely overall teaching quality and the overall clarity of the lectures. Additionally, we also look at other items: the teacher's ability in generating interest for the subject, the logistics of the

course (schedule of classes, combinations of practical sessions and traditional lectures) and the total workload compared to other courses.

Formally, we estimate the following equation:

$$q^k_{dtcs} = \lambda_0 + \lambda_1 \widehat{\alpha}_{dtcs} + \lambda_2 C_{dtcs} + \lambda_3 T_{dtcs} + \gamma_d + \delta_t + \upsilon_c + \epsilon_{dtcs} \tag{2.3}$$

where q^k_{dtcs} is the average answer to question k in class s of course c in the degree program d (which is taught in term t), $\widehat{\alpha}_{dtcs}$ is the estimated class fixed effect from equation 2.1, C_{dtcs} is the set of class characteristics, T_{dtcs} is the set of teacher characteristics. γ_d, δ_t and υ_c are fixed effects for degree program, term and subject areas, respectively. ϵ_{dtcs} is a residual error term.

Notice that the class and teacher characteristics are exactly the same as in Table 2.10, so that equation 2.3 is equivalent to a partitioned regression model of the evaluations q_{dtcs} on our measures of teacher effectiveness, i.e. the residuals of the regressions in Table 2.10, where all the observables and the fixed effects are partialled out.

Since the dependent variable in equation 2.3 is an average, we use weighted OLS, where each observation is weighted by the square root of the number of collected questionnaires in the class, which corresponds to the size of the sample over which the average answers are taken. Additionally, we also bootstrap the standard errors to take into account the presence of generated regressors (the $\widehat{\alpha}$'s).

The first four columns of Table 2.15 reports the estimates of equation 2.3 for a first set of core evaluation items, namely overall teaching quality and lecturing clarity. For each of these items we show results obtained using our benchmark estimates of teacher effectiveness and those obtained using the contemporaneous class effects.

Results show that our benchmark class effects are negatively associated with all the items that we consider. In other words, teachers who are more effective in promoting

Table 2.15: Teacher effectiveness and students' evaluations

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

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

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

The results in Table 2.15 clearly challenge the validity of students' evaluations of professors as a measure of teaching quality. Even abstracting from the possibility that professors strategically adjust their grades to please the students (a practice that is made difficult by the timing of the evaluations, that are always collected before the exam takes place), it might still be possible that professors who make the classroom experience more enjoyable do that at the expense of true learning or fail to encourage students to exert effort. Alternatively, students might reward teachers who prepare them for the exam, that is teachers who teach to the test, even if this is done at the expenses of true learning. This interpretation is consistent with the results in Weinberg, Fleisher, and Hashimoto (2009), who provide evidence that students are generally unaware of the value of the material they have learned in a course, and it is the interpretation that

we adopt to develop the theoretical framework of Section 2.6.

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

2.5 Robustness checks

In this section we present robustness checks for our main results in Sections 2.3 and 2.4.

First, we investigate the role of students' dropout in the estimation of our measures of teacher effectiveness. In our main empirical analysis students who do not have a complete academic record are excluded. These are students who either dropped out of higher education or have transfered to another university or are still working towards the completion of their programs, whose formal duration was 4 years. They total about 10% of all the students who enrolled in their first year in 1998-1999. In order to check that excluding them does not affect our main results, in Figure 2.4 we compare our benchmark measure of teacher effectiveness estimated in Section 2.3 with similar estimates that include such dropout students. As it is evident, the two sets of estimates are very similar and regressing one over the other (controlling for degree program, term and subject fixed effects) yields an R^2 of over 88%. Importantly, there does not seem to be larger discrepancies between the two versions of the class effects for the best or the worst teachers.

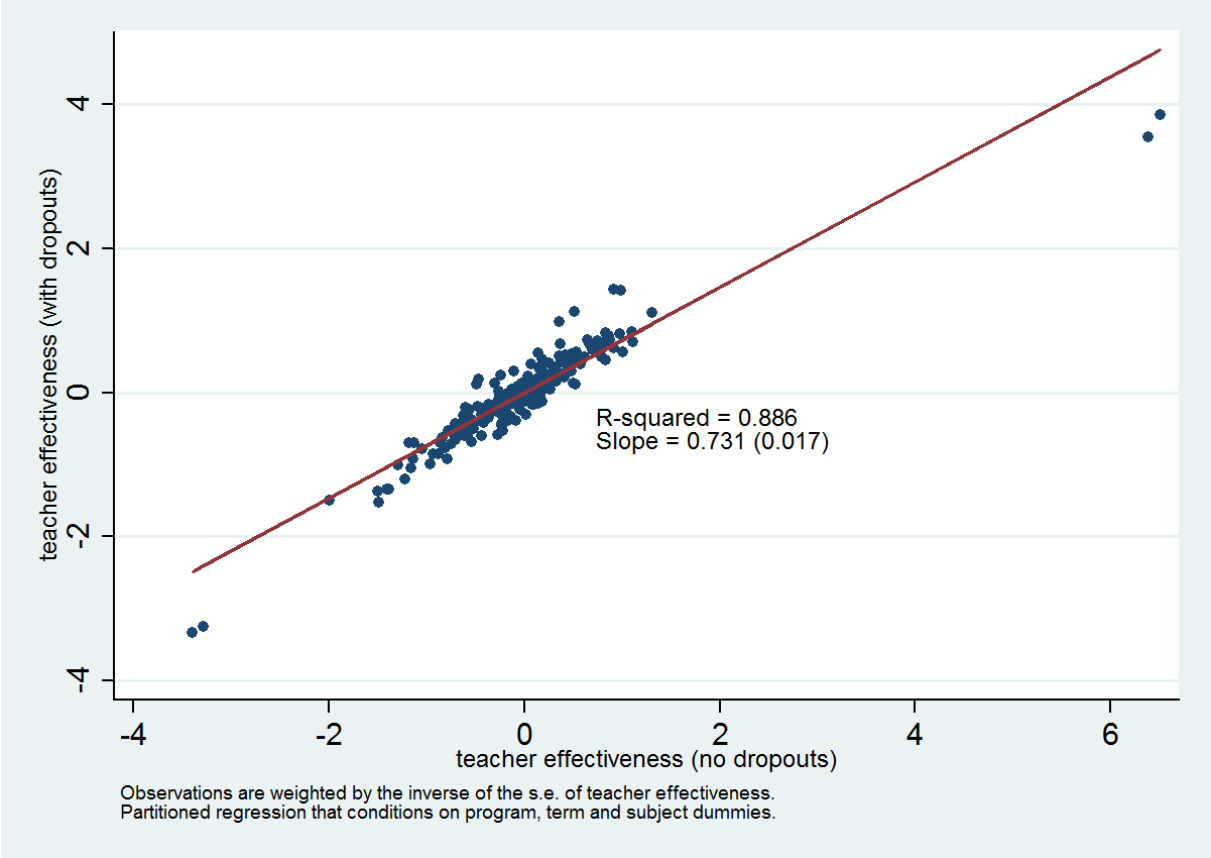


Figure 2.4: Robustness check for dropouts

Second, one might be worried that students might not comply with the random assignment to the classes. For various reasons they may decide to attend one or more courses in a different class from the one to which they were formally allocated. For example, they may desire to stay with their friends, who might have been assigned to a different class, or they may like a specific teacher, who is known to present the subject particularly clearly. Unfortunately, such changes would not be recorded in our data, unless the student formally asked to be allocated to a different class, a request that needed to be adequately motivated.[28] Hence, we cannot exclude a priori that some

[28]Possible motivations for such requests could be health reasons. For example, due to a broken leg a student might not be able to reach classrooms in the upper floors of the university buildings and could ask to be assigned to a class taught on the ground floor.

students switch classes.

If the process of class switching is unrelated to teaching quality, then it merely affects the precision of our estimated class effects, but it is very well possible that students switch in search for good or lenient lecturers. We can get some indication of the extent of this problem from the students' answers to an item of the evaluation questionnaires that asks about the congestion in the classroom. Specifically, the question asks whether the number of students in the class was detrimental to one's learning. We can, thus, identify the most congested classes from the average answer to such question in each course.

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

In Table 2.16 we replicate our benchmark estimates for the two core evaluation items (overall teaching quality and lecturing clarity) by excluding the most switched course (Panel B), i.e. the course with the largest difference between the most and the least congested classes (which is marketing). For comparison we also report the original estimates from Table 2.15 in Panel A and we find that results change only marginally. Next, in Panel C and D we exclude from the sample also the second most switched course (human resource management) and the five most switched courses, respectively.[29] Again, the estimated coefficients are only mildly affected, although the

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

Table 2.16: Robustness check for class switching

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

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

significance levels are reduced according with the smaller sample sizes. Overall, this exercise suggests that course switching should not affect our estimates in any major direction.

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

To control for this potential feedback effect on students' effort, we recompute our benchmark measures of teacher effectiveness adding the student average grade in all previous courses to the set of controls. Figure 2.5 compares our benchmark teacher effectiveness with this augmented version, conditioning on the usual fixed effects for degree program, term and subject area and shows that the two are strongly correlated (even accounting for the outliers).

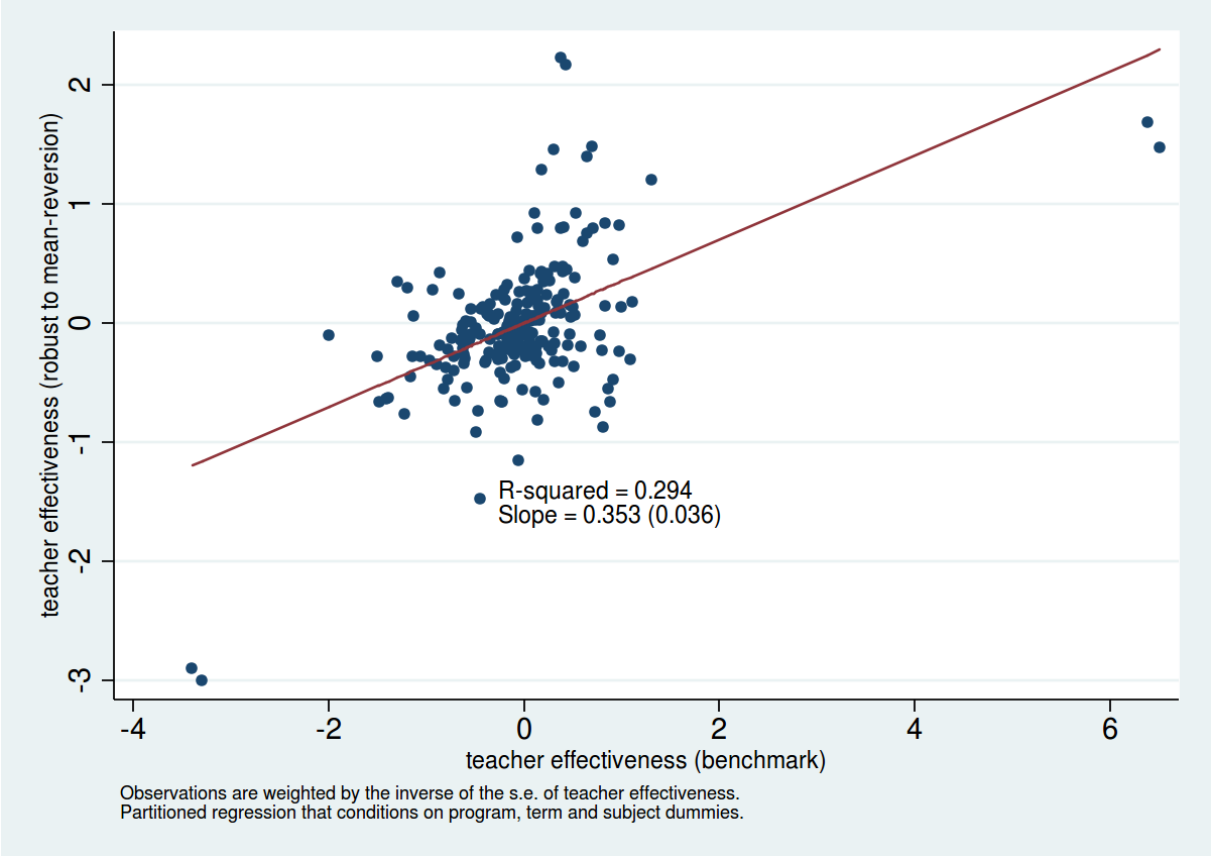


Figure 2.5: Robustness check for mean reversion in grades

2.6 Interpreting the results: a simple theoretical framework

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

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

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

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

$$x_i = \mu_j h(e_i) + (1 - \mu_j)\overline{x} \tag{2.4}$$

which is a linear combination of a function $h(\cdot)$ of student's effort e_i and a constant $\overline{x}$, weighted by the teacher's type μ_j. We assume $h(\cdot)$ to be a continuous and twice differentiable concave function. Under full real teaching ($\mu_j = 1$) grades vary with students' effort; on the other hand, if the teacher exclusively teaches to the test ($\mu_j = 0$), everyone gets the same grade $\overline{x}$, regardless of effort. This strong assumption can obviously be relaxed and all our implications will be maintained as long as the gradient of grades to effort increases with μ_j.

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

and one needs to personally and originally elaborate on one's own understanding of the course material. Of course, there are costs and benefits to each type of exam and multiple-choice tests are often adopted because they can be marked quickly, easily and uncontroversially. For the sake of simplicity, however, we abstract from cost-benefit considerations.

Furthermore, equation 2.4 assumes that teaching-to-the-test does not require students to exert effort. All our results would be qualitatively unchanged under the weaker assumption that teaching-to-the-test requires less effort by the students. We also assume that μ_j is a fixed characteristic of teacher j, so that the model effectively describes the conditions for selecting teachers of different types, a key piece of information for hiring and promotion decisions. Alternatively, μ_j could be treated as an endogenous variable under the control of the individual teacher, in which case the model would feature a rather standard agency problem where the university tries to provide incentives to the teachers to choose a μ_j close to 1. Although, such a model would be considerably more complicated than what we present in this section, its qualitative results would be unchanged and the limited information on teachers in our data would make its additional empirical content redundant in our setting.

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

In all cases, a key assumption is that μ_j is unobservable by the university administra-

tors (the principal) and, although it might be observable to the students, it cannot be credibly communicated to third parties.

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

$$U_i = x_i - \frac{1}{2}\frac{e_i^2}{\eta_i} \tag{2.5}$$

where η_i is a measure of student's ability.

For simplicity, we assume that students are perfectly informed about the production function of grades, i.e. they know the type of their teacher, they know the return to their effort and there is no additional stochastic component to equation 2.4. This assumption can be easily relaxed by introducing either imperfect information about the teacher's type or about the exact specification of the production function and, consequently, by rewriting the utility function in equation 2.5 in expected terms. The main intuition of our results would be unchanged. Although the perfect information assumption is obviously a modeling device and does not correspond to reality, we do believe that students know a lot about their professors, either through conversations with older students or by observation through the duration of the course.

The utility function in equation 2.5 implicitly assumes that students are myopic, in the sense that they care only about grades and not about real learning. The main implications of the simple theory in this section would remain unchanged also with a different utility function that incorporates real learning, as long as students of different abilities care equally about it (just like they are equally myopic in the current specification).

The quasi-linearity of equation 2.5 simplifies the algebra of the model. Alternatively, we could have introduced some curvature in the utility function and assumed a linear production process without affecting the results. With non-linearities both in the produc-

tion and in the utility functions one would have to make explicit a number of additional assumptions to guarantee existence and uniqueness of the equilibrium.

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

$$\mu_j \frac{\partial h(e)}{\partial e_i}(e_i^*) = \frac{e_i^*}{\eta_i} \tag{2.6}$$

Using equation 2.6 it is easy to derive the following results:

$$\frac{de_i^*}{d\eta_i} > 0 \tag{2.7}$$

$$\frac{de_i^*}{d\mu_j} > 0 \tag{2.8}$$

$$\frac{de_i^*}{d\mu_j d\eta_i} > 0 \tag{2.9}$$

Equation 2.7 shows that more able students exert higher effort. Equation 2.8 shows that more real teaching induces higher effort from the students and equation 2.9 indicates that such an effect is larger for the more able students

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

$$\frac{\partial U_i(e_i^*)}{\partial \mu_j} = h(e_i^*) - \overline{x} \tag{2.10}$$

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

$$\frac{\partial U_i(e_i^*)}{\partial \mu_j} > 0 \quad \text{if} \quad \eta_i > \overline{\eta} \tag{2.11}$$

$$\frac{\partial U_i(e_i^*)}{\partial \mu_j} < 0 \quad \text{if} \quad \eta_i < \overline{\eta} \tag{2.12}$$

Equations 2.11 and 2.12 are particularly important under the assumption that,

especially when answering questions about the overall quality of a course, students give a better evaluation to teachers (or classes) that are associated with a higher level of utility. Equations 2.11 and 2.12 suggest that high ability students evaluate better teachers or classes that are more focused on real learning while low ability students prefer teachers that teach to the test. Hence, if the (benchmark) teacher effects estimated in Section 2.3 indeed measure the real learning value of a class (μ_j, in the terminology of our model), we expect to see a more positive (or less negative) correlation between such class effects and the students' evaluations in those classes where the concentration of high ability students is higher.

2.7 Further evidence

In this section we present three additional pieces of evidence that are consistent with the implications of the model of Section 2.6.

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

In order to provide support for our specification, in Table 2.17 we produce evidence that the students' evaluations respond to the weather conditions on the day when they were filled. There is ample evidence that people's utility (or welfare, happiness,

satisfaction) improves with good meteorological conditions (Barrington-Leigh, 2008; Denissen, Butalid, Penke, and van Aken, 2008; Keller, Fredrickson, Ybarra, Coté, Johnson, Mikels, Conway, and Wager, 2005; Pray, 2011; Schwarz and Clore, 1983) and finding that such conditions also affect the evaluations of professors suggests that they indeed reflect utility rather than (or together with) teaching quality.

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

Obviously, teachers might be affected by meteorological conditions as much as their students and one may wonder whether the estimated effects in the odd columns of Table 2.17 reflect the indirect effect of the weather on teaching effectiveness. We consider this interpretation to be very unlikely since the questionnaires are distributed and filled before the lecture so that students should not be able to incorporate in their answers the performance of the teacher in the day the evaluation forms are filled in. Moreover, students' are asked to evaluate teachers' performance over the entire duration of the course and not exclusively on the day of the test.

Nevertheless, in the even columns of Table 2.17, we also condition on our benchmark measure of teaching effectiveness and, as we expected, we find that the estimated effects of both the weather conditions and teacher effectiveness itself change only marginally.

Second, our specification of the production function for exam grades in equation 2.4 implies a positive relationship between grade dispersion and the professor's propensity to engage in real teaching (μ_j). In our empirical exercise our measures of teacher effectiveness can be interpreted as measures of the μ_j's in the terminology of the model. Hence, if grades were more dispersed in the classes of the worst teachers one would

Table 2.17: Students' evaluations and weather conditions

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

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

have to question our specification of equation 2.4.

In Figure 2.6 we plot the coefficient of variation of grades in each class (on the vertical axis) against our measure of teacher effectiveness (on the horizontal axis). To take proper account of differences across degree programs, the variables on both axes are the residuals of weighted OLS regressions that condition on degree program, term and subject area fixed effects, as in standard partitioned regressions (the weights are the squared roots of class sizes). Consistently with equation 2.4 in our model, the two variables are positively correlated and such a correlation is statistically significant at conventional levels: a simple univariate OLS regression of the variable on the vertical axis on the variable on the horizontal axis yields a coefficient of 0.011 with a standard error of 0.004.

Next, according to equations 2.11 and 2.12, we expect the correlation between our measures of teacher effectiveness and the average student evaluations to be less negative in classes where the share of high ability students is higher. This is the hypothesis that we investigate in Table 2.18. We define as high ability those students who score in the upper quartile of the distribution of the entry test score and, for each class in our data, we compute the share of such students. Then, we investigate the relationship between the students' evaluations and teacher effectiveness by restricting the sample to classes in which high-ability students are over-represented. Results seem to suggest the presence of non linearities or threshold effects, as the estimated coefficient remains relatively stable until the fraction of high ability students in the class goes above one quarter or, more precisely, 27% which corresponds to the top 25% of the distribution of the presence of high-ability students. At that point, the estimated effect of teacher effectiveness on students' evaluations is about a quarter of the one estimated on the entire sample. The results, thus, suggest that the negative correlations reported in Table 2.15 are mostly due to classes with a particularly low incidence of high

ability students.

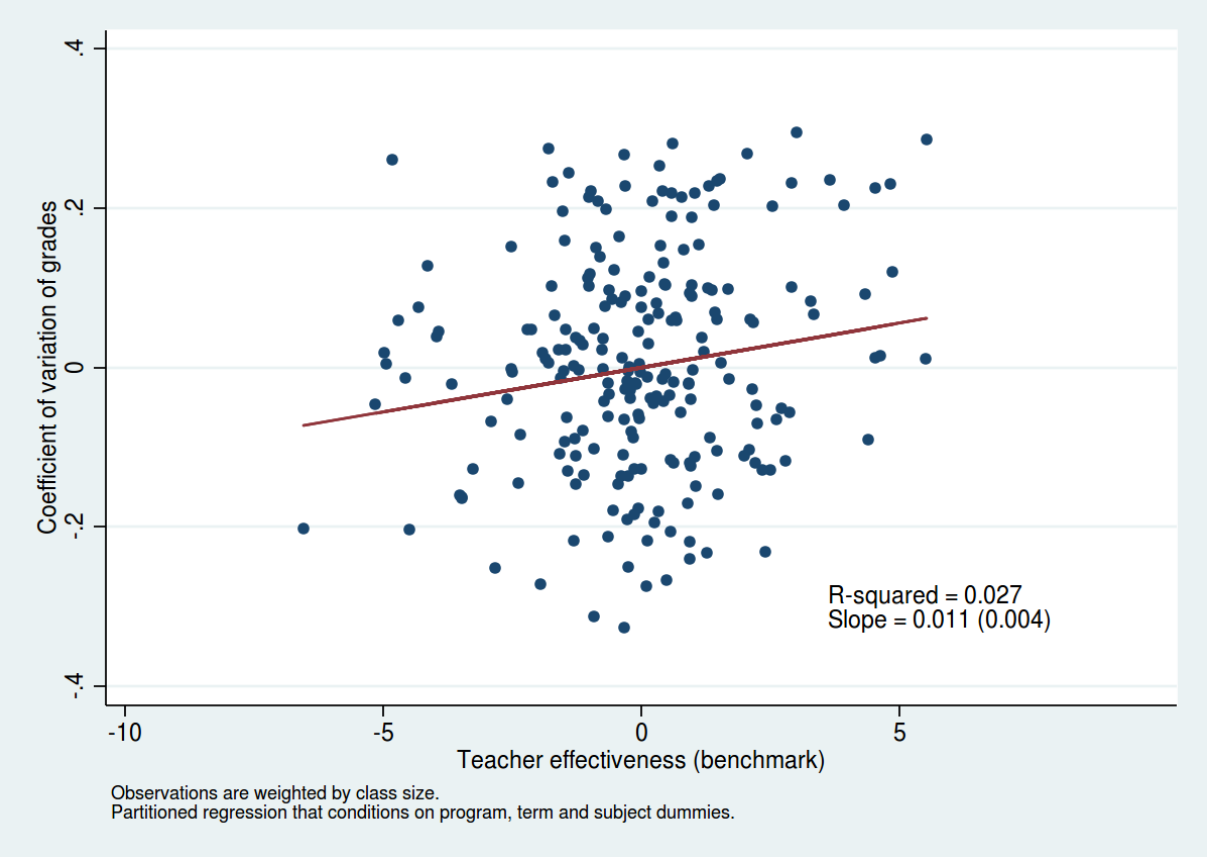


Figure 2.6: Teacher effectiveness and grade dispersion

Table 2.18: Teacher effectiveness and students evaluations by share of high ability students

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

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

2.8 Policies and conclusions

Using administrative archives from Bocconi University and exploiting random variation in students' allocation to teachers within courses we find that, on average, students evaluate positively classes that give high grades and negatively classes that are associated with high grades in subsequent courses. These empirical findings can be rationalized with a simple model featuring heterogeneity in the preferences (or ability) of teachers to engage in real teaching rather than teaching-to-the-test, with the former requiring higher effort from students than the latter. Furthermore, we also find that weather conditions on the day the questionnaires are filled are correlated with the

students' evaluations of teachers. This is consistent with the assumption of our model that the evaluations reflect students' perceived utility more than teachers' ability as such. Overall, our results cast serious doubts on the validity of students' evaluations of professors as measures of teaching quality or effort.

At the same time, the strong effects of teaching quality on students' outcomes, as documented in Section 2.3, suggest that improving the quantity or the quality of professors' inputs in the education production function can lead to large gains. Under the interpretation offered by our model in Section 2.6, this could be achieved through various types of interventions.

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

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

increased substantially.

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

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

Obviously, these, as well as other potential alternative measurement methods, are costly but they should be compared with the costs of the current systems of collecting students' opinions about teachers, which are often non trivial.

Chapter 3

Performance-Related Pay and Firms' Productivity

3.1 Introduction

Wage bargaining institutions are of paramount importance in various fields of economics. On the macroeconomic side, mechanisms that link wages to firms' performances should make labor costs more flexible, and real wages more responsive to the business cycle, with important macroeconomic implications in terms of employment levels, risk smoothing, inflation and business cycle dynamics (Ichino, 1994; Nucci and Riggi, 2011; Nuti, 1987; Weitzman, 1985a,b). On the microeconomic side, incentive pay is widely believed to be a useful tool to correct market imperfections and to better align principals' and agents' objectives, thus solving the agency problem (Lazear, 2000; Prendergast, 1999). Incentive schemes are therefore often introduced with the goal of motivating workers, making them more efficient and, in the end, making firms more profitable. The increase in productivity may occur through two main channels (Lazear, 1986): by extracting the efficient amount of effort from workers (incentive effect) or by

inducing the most able workers to join the firm and the least able to leave it (selection effect). The introduction and the actual design of incentive schemes (which can be individual- or group-based) can have many other consequences, in terms of sorting and selection of workers and efficiency-insurance tradeoffs (see Bryson, Freeman, Lucifora, Pellizzari, and Pérotin, 2010 for an exhaustive report on incentive pay schemes).

The purpose of this work is to contribute to the empirical literature that investigates the effects of the introduction of collective performance-related pay (PRP) schemes on firms' productivity. I use firm-level panel data from the Bank of Italy's Survey of Manufacturing Firms, a longitudinal survey conducted since the early 1980's on a nationally representative sample of Italian medium- and large-sized manufacturing firms (with more than 50 employees). Given that firms can freely choose wheter or not to introduce PRP schemes, in order to carefully construct an appropriate counterfactual outcome and reduce the bias associated with firms' self-selection I exploit the richness of the dataset and use propensity-score matching (PSM) estimators (Rosenbaum and Rubin, 1983); furthermore, I exploit the longitudinal structure of the data and combine PSM with a difference-in-differences (DD) strategy (Abadie, 2005; Heckman, Ichimura, and Todd, 1998, 1997). In the short-run (one year after the introduction of PRP scheme) I do not find statistically significant effects on firms' productivity; coefficient estimates range from 2 to 9% (depending on the choice of the matching algorithm) and are very close to zero when using DD. In the medium run (three years after the introduction of PRP) estimates are always in the 3-5% range in all specifications, but continue to be not statistically significant.

The previous literature has generally found a strong positive correlation between the adoption of PRP schemes and firms' productivity. In particular, there is substantial evidence that the introduction of individual incentives (such as piece rate schemes) increases productivity through higher effort and positive sorting of the most productive

workers (Bandiera, Barankay, and Rasul, 2005, 2007, 2009; Lazear, 2000; Paarsch and Shearer, 2000; Shearer, 2004); however, such schemes may be costly to maintain and can have negative consequences on other aspects of workers' behaviour, leading to small or even negative effects on firms' profitability. This might explain why group-based incentive schemes are becoming increasingly common (Pendleton, Withfield, and Bryson, 2009). Some studies have used detailed data from a single firm in order to estimate the productivity effects of PRP. This "case-study" approach trades off internal validity (results are more convincing since it is possible to compare the *same* firm and the *same* workers under different regimes) with external validity (results obtained for a specific firm are not necessarily valid for other firms). Knetz and Simester (2001), for instance, find positive effects from the introduction of a firm-wide incentive scheme in Continental Airlines, while Burgess, Propper, Ratto, von Hinke Kessler Scholder, and Tominey (2010) also find positive effects of group incentive schemes in the UK tax collection authority. Other studies have used representative firm surveys. Early work by Kruse (1992) found that the adoption of profit sharing in US manufacturing firms is associated with productivity increases between 2.8 and 3.5%. Cable and Wilson (1989) estimate effects between 3 and 8% for UK engineering firms, while Cahuc and Dormont (1997) find a positive effect of 2% for French manufacturing firms. More recently, Gielen, Kerkhofs, and van Ours (2010), using a panel of Dutch firms, estimated that PRP schemes increased productivity by 9%.

For the case of Italy, Amisano and Del Boca (2004) studied the determinants of the introduction of profit-sharing schemes, finding higher propensity to adopt them in larger firms, and that the associated increase in wages is used to boost an initially low level of productivity; Cristini and Leoni (2007) estimate the elasticity of wages to profits per employee to be as low as 2%. An early study by Biagioli and Curatolo (1999) argue that productivity gains of PRP are around 10%; similarly, Damiani and Ricci (2008) use

balance sheet data for a panel of firms merged with cross-sectional information from a survey on the diffusion of PRP schemes and find that the introduction of PRP schemes increased productivity by 13%. Origo (2009) is the most recent study, and also the most similar to this paper. She uses panel data from a sample of metalworking firms and, using propensity score matching methods, estimates positive effects on productivity between 7 an 11%. She also finds that productivity effects are much smaller in highly unionized firms. This can help reconciling the findings of the two papers. In the Bank of Italy's Survey that I use, only firms with more than 50 employees are surveyed, while the sample of metalworking firms used by Origo (2009) includes also smaller firms. Given the positive association between firm size and unionization rates, it is likely that the results of this paper are more comparable with what Origo (2009) finds for her subsample of highly unionized firms.

The rest of the paper is organized as follows. In section 3.2 I present the institutional context that has led to the diffusion of firm-level contract and to the introduction of PRP schemes in Italy during the 1990s. In section 3.3 I present the dataset used in the empirical analysis and the econometric strategy adopted to estimate the effect of PRP on firms' productivity. In section 3.4 I present the results of the empirical analysis, and in section 3.5 I conclude.

3.2 The institutional setting

The first major reform of the wage bargaining system was introduced in Italy in 1993, with an agreement signed by the Italian government and the major national trade unions and employers' association. The main purpose of the agreement was to curb the inflation rate (in the light of the EU Maastricht targets) by breaking the wages-prices spiral that had characterized the Italian economy in the 1980s (for this purpose, in 1992 the system

that automatically indexed wages to price level - *Scala mobile* - was abolished). The agreement introduced a two-level wage bargaining system: national-level bargaining (by sector) was meant to preserve wages' purchasing power, while local bargaining (at the firm or at a lower territorial level) was meant to allow for sharing of productivity gains through performance-related pay mechanisms. The variable pay schemes could only add to minimum wage levels set by national agreements, and the amount of the bonuses was usually the same for all workers involved (but for differences linked to the individual position in the firm's hierarchy or to indicators of individual absenteeism).

Following the 1993 agreement, firm-level contracts had a rapid diffusion. In 1996, 40% of workers in firms with at least 10 employees were estimated to be covered by such contracts; however, such workers were employed by only 10% of the firms surveyed: the propensity to adopt a firm-level contract was therefore sharply increasing in firms' size and unionization rates (Istat, 2000). Wage issues were the single most important bargaining topic, and PRP schemes were bargained in 40% of bargaining firms, employing 60% of the workers covered by local contracts (Istat, 1999; Origo, 2009). The majority of workers covered by firm-level contracts were employed in the manufacturing sector. In 2001 the Bank of Italy Survey was extended to cover firms with 20-49 employees: roughly a third of these smaller firms were covered by local schemes; between 1997 and 2002 coverage was significantly increased in the credit sector and among large retailers (Casadio, 2003, 2010).

The actual incidence of PRP over the total wage bill has been quite limited: in the last ten years, PRP bonuses have increased total wages by roughly 1% (Brandolini, Casadio, Cipollone, Magnani, Rosolia, and Torrini, 2007; Casadio, 2010). Since the mid nineties, collective PRP has also been subject to some forms of tax exemption (Bryson, Freeman, Lucifora, Pellizzari, and Pérotin, 2010). From 1997 to 2007 firms were exempted from paying social security contributions on part of the actual amount

of the collective PRP, and between 2008 and 2010 firms could enjoy a tax relief equal to 25 percentage points of the total social security contribution rate due on the actual amount of collective PRP (available funds were allocated to eligible firms according to a first-come-first-served rule).

3.3 Data and empirical strategy

3.3.1 The dataset

The data used in this paper come from the Bank of Italy Survey of Manufacturing Firms. Informations concerning the presence and the characteristics of company-level agreement are present in the 1994, 1999 and 2001 waves. Other informations were collected by some regional branches of the Bank[1]. Up to 2001, the Survey was restricted to manufacturing firms with more than 50 employees. Smaller manufacturing firms (20-49 employees) were surveyed since 2001, and firms in the service sector since 2002. The need to have as much firm-level informations as possible for the longest possible time span is the main reason why in this paper I restrict attention to large manufacturing firms (more than 50 employees), for which I have consistent longitudinal data starting in 1984. As already discussed in section 3.2, among these larger firms the adoption of firm-level contracts and PRP schemes was quite common. Table 3.1 reports, for each year between 1984 and 2001, the share of firms covered by a local agreement, the share of firms with a PRP scheme in place, and the firms that, in each year, introduce a PRP scheme (as a percentage of firms without PRP). The majority of firms had already adopted a firm-level contract in the mid 80's; however, PRP schemes started to become more common following the 1993 national agreement. Between 1993 and

[1] I am especially grateful to Piero Casadio and Leandro D'Aurizio for having put together and made available all the relevant informations.

2001, the share of firms covered by such schemes increased from 17 to 60%. After the introduction of PRP schemes, no firms switched back to other contractual arrangements (the most common of which is the provision of premia of a fixed amount, bargained with local unions).

Table 3.1: Coverage of Performance-related pay schemes

year	No. firms[a]	% firms with firm-level contract	% firms with a PRP scheme	% firms introducing PRP [b]
1984	135	28.148	0.000	0.000
1985	219	55.708	0.000	0.000
1986	297	67.340	0.000	0.000
1987	346	71.965	0.578	0.290
1988	482	80.083	7.469	6.303
1989	581	83.821	10.499	3.704
1990	705	88.085	11.489	1.730
1991	880	91.591	12.159	1.899
1992	916	92.031	14.410	2.726
1993	961	92.508	16.961	3.277
1994	1,019	93.229	26.791	10.577
1995	1,131	94.872	41.910	16.000
1996	1,311	96.339	51.259	13.260
1997	1,374	96.652	54.731	6.028
1998	1,418	96.756	57.334	5.920
1999	1,469	97.005	59.156	4.801
2000	1,471	97.008	59.279	0.347
2001	1,471	97.008	59.279	0.000

[a] Number of firms for which information on firm-level contract is available.
[b] As a share of firms without PRP.

The introduction of PRP schemes is clearly a free choice of the firm. Table 3.2 reports some basic descriptive statistics for the sample used in the subsequent empirical analysis, covering the 1984-2001 period. Most of the variables presented in the table will then be used to implement the propensity-score matching estimation strategy.

Firms with a PRP schems are on average larger, more productive (when productivity is measured by real sales per worker), export a higher share of total sales, make a more intense use of installed capacity and have a lower share of blue collars. Firms

Table 3.2: Descriptive statistics

	All firms		With an active PRP scheme		Never introducing PRP	
	Mean	Std.Dev.	Mean	Std.Dev.	Mean	Std.Dev.
Real sales per worker[a]	248.5	778.7	296.3	985.0	194.4	168.12
Export share	0.33	0.27	0.36	0.27	0.30	0.27
Investment per worker[a]	11.4	22.0	12.1	17.2	10.8	28.5
No. employees	665.2	2,412.7	790.0	2,862.4	444.9	1,072.0
Turnover rate[b]	0.14	0.2	0.13	0.1	0.16	0.3
Employment growth[c]	-0.01	0.07	-0.004	0.06	-0.03	0.09
% Capacity utilization	80.3	10.37	81.5	9.64	78.9	12.0
Share blue collars	0.7	0.19	0.6	0.19	0.7	0.18
Temp. lay-off hours per worker	93.4	217.3	56.1	171.9	125.0	291.7
Hours worked per worker	1,634.1	178.6	1,652.9	154.6	1,617.8	217.4
Hours of strike[d]	0.005	0.006	0.003	0.005	0.004	0.007
Firm age	38.8	27.5	41.9	29.3	34.7	24.8
No. yrs with PRP	3.2	3.4	6.9	2.7	0.0	0.0
No. yrs with firm contract	7.3	4.4	10.6	4.3	5.7	4.2
			Distribution by Sector			
Food and beverage	0.09		0.088		0.087	
Textile	0.18		0.133		0.251	
Chemicals	0.14		0.153		0.120	
Minerals	0.07		0.067		0.087	
Metalworking	0.42		0.461		0.353	
Other manufacturers	0.10		0.096		0.101	
Mining and energy	0.002		0.002		0.000	
			Distribution by geographical areas			
North-West	0.46		0.489		0.424	
North-East	0.24		0.271		0.188	
Centre	0.18		0.153		0.224	
South and Islands	0.12		0.086		0.164	
No. firms [e]	1,365		839		526	

[a] Thousand Euro 2009

[b] Number of employees leaving over total average number of employees.

[c] Difference between number of employees joining the firm and number of employees leaving the firm, over the total average number of employees.

[d] Over total hours worked.

[e] Referred to Real Sales per Worker, No. Employees, Investment per Worker. For other variables we tipically have fewer observations.

with a PRP scheme are also older (41 versus 34 years), are on average covered by a firm level contract for 10 years out of 17 (less than 6 years for firms without PRP) and stay with a PRP scheme in place for more than 6 years on average. The incidence of PRP schemes is more pronounced among firms located in the North of the country and operating in the metalworking sector.

3.3.2 Empirical strategy

The main purpose of this paper is to estimate the effect of the introduction of a PRP scheme on firm's productivity; this task can be expressed as a standard treatment evaluation problem. Let D_i be a dummy variable denoting whether firm i switches to a PRP scheme; $Y_i(1)$ is the outcome of interest (a measure of firm productivity) when $D_i = 1$, and $Y_i(0)$ is the outcome of interest when $D_i = 0$; the Average Treatment Effect on the Treated (ATT) can be expressed as

$$ATT = E\left\{Y_i(1) - Y_i(0)|D_i = 1\right\} \tag{3.1}$$

The "fundamental problem of treatment evaluation" lies in the fact that the counterfactual outcome $Y_i(0)|D_i = 1$ is (by definition) not observable. In the absence of random variation in treatment status, the average outcome of non-treated firms is not a consistent estimate of the counterfactual outcome for treated firms, since different observable and unobservable firms' characteristics may determine both selection into the treatment and the outcome.

If one is willing to assume that the effect of the treatment is linear and common across participants and that selection into the treatment is only driven by a set of variables X_i that are observable to the econometrician, then a consistent estimate of the ATT can be obtained by simply running OLS on the following equation:

$$Y_i = \alpha + \beta X_i + \gamma D_i + \epsilon_i \tag{3.2}$$

A more convincing strategy can be adopted by exploiting the longitudinal structure of the dataset. Firm fixed-effects allow to control for all observable and unobservable time-invariant characteristics that could induce firms to adopt a PRP scheme. This is the route taken by Gielen, Kerkhofs, and van Ours (2010) and Bryson, Freeman, Lucifora, Pellizzari, and Pérotin (2010), which estimate the following equation:

$$Y_{it} = \alpha + \delta_i + \lambda_t + \beta X_{it} + \gamma D_{it} + \epsilon_{it} \tag{3.3}$$

This approach has some drawbacks as well. The estimated ATT will be inconsistent if selection into the treatment is driven by unobservable time-varying factors, or by past realizations of the outcome variable. Furthermore, this specification still assumes a linear, common and constant effect of the treatment. Non-parametric methods can improve upon OLS and FE estimation by relaxing the linearity assumption, by making more transparent the construction of the comparison group and by allowing more flexibility in estimating heterogeneus effects (for instance, at different points in time). Widely used in the program evaluation literature is the Propensity Score Matching estimator (PSM). Letting $P(X)$ be the probability of treatment conditional on X, Rosenbaum and Rubin (1983) showed that $D \perp X|P(X)$ (balancing property). The propensity score can thus helps in solving the so-called "curse of dimensionality": if the counterfactual outcome $Y(0)$ is unrelated to treatment status after conditioning on observables X, (Conditional independence assumption - CIA), it will be so also conditioning on the scalar measure $P(X)$. Treated and untreated units can then be compared on the basis of their propensity score:

$$\widehat{ATT} = \sum_{i \in T} \left\{ Y_i - \sum_{j \in C} w_{ij} Y_j \right\} w_i \tag{3.4}$$

where T and C denote treated and non-treated units, w_{ij} is the weight associated to unit j when matched to unit i and w_i is a system of weights to control for the distribution of the treated sample. By matching individual units on the basis of their observable characteristics, this approach allows to effectively compare only comparable individuals, avoiding the extrapolation necessarily performed with parametric methods when the distribution of observable characteristics is very different between treated and non-treated units. Several matching procedures have been developed in the literature, mainly differing in the system of weights given to the matched controls[2].

It is important to stress that, as with OLS estimation, identification of treatment effects still relies on the assumption that selection into treatment is driven by observable firm's characteristics (CIA). The conditional independence assumption can be partly relaxed by exploiting the longitudinal structure of the data (Abadie, 2005; Heckman, Ichimura, Smith, and Todd, 1998; Heckman, Ichimura, and Todd, 1997). Even if the CIA does not hold, one can assume that the associated bias is constant over time. The so-called Conditional Bias Stability Assumption can be expressed as

$$\Delta Y(0) \perp D | P(X) \tag{3.5}$$

The difference-in-differences propensity score matching estimator can thus be written as

$$\widehat{ATT} = \sum_{i \in T} \left\{ (Y_{i,t+\sigma} - Y_{i,t-\tau}) - \sum_{j \in C} w_{ij} (Y_{j,t+\sigma} - Y_{j,t-\tau}) \right\} w_i \tag{3.6}$$

[2]Useful discussions about the pros and cons of different matching procedures can be found, among others, in Becker and Ichino (2002); Caliendo and Kopeinig (2008); Heinrich, Maffioli, and Vázquez (2010).

In the remaining of the paper, t will always denote the year od treatment (i.e. the year in which the firm adopted a PRP scheme), while τ will be always set to 1. σ will be set equal to 1 (to evaluate short-run effects) or 3 (to evaluate medium-run effects).

3.4 Econometric results

A first set of result is presented in table 3.3, using the parametric specifications of equation 3.2 and 3.3. Panel A reports results from a pooled OLS estimation, while panel B reports panel FE estimates. The dependent variable is the log of real sales per worker. Column 1 of panel A shows that the unconditional difference in productivity between firms with and without a PRP scheme is as large as 26%. Adding a set of controls like time, macro-region[3] and sector dummies, the estimated effect is reduced to 10%. Further controlling for firm-level characteristics, the estimated effect is 4.5% and is not statistically different from zero. A similar result is obtained in column 4, where the sample is restricted to firms having a firm-level contract. In column 5 I follow the approach taken by Bryson, Freeman, Lucifora, Pellizzari, and Pérotin (2010), restricting the sample to firms introducing PRP since 1995 (i.e., two years after the institutional reform of 1993); the estimated effect is around 7%.

Pooled OLS estimates are likely to be upward biased, since they are unable to control for possible selection into the treatment driven by unobservables firm characteristics. The bias due to time-invariant unobservables is eliminated by adding firm fixed-effects (in panel B). The estimated effect of PRP schemes on productivity is 3.6%, much smaller than what previously found in the literature. Using an equivalent specification, Origo (2009) estimates an effect of almost 7%, while Bryson, Freeman, Lucifora, Pellizzari, and Pérotin (2010) estimate effects between 5 and 6%; estimates for Dutch firms presented

[3]We use dummies for the four Italian macro-regions: North-West, North-East, Centre, South and Islands.

by Gielen, Kerkhofs, and van Ours (2010) are as large as 9%.

Table 3.3: Parametric estimates

	Panel A: Pooled OLS				
	[1]	[2]	[3]	[4[a]]	[5[b]]
PRP scheme	0.259***	0.098***	0.045	0.041	0.073**
	(0.013)	(0.009)	(0.021)	(0.019)	(0.018)
Year dummies	NO	YES	YES	YES	YES
Macro-region dummies	NO	YES	YES	YES	YES
Sector dummies	NO	YES	YES	YES	YES
Firm-level controls [c]	NO	NO	YES	YES	YES
R^2	0.035	0.220	0.404	0.405	0.395
No. observations	10,642	10,642	9,643	9,143	7,566
	Panel B: Fixed Effects				
	[1]	[2]	[3]	[4[a]]	[5[b]]
PRP scheme	0.230***	0.040***	0.036***	0.028**	0.053***
	(0.014)	(0.015)	(0.013)	(0.013)	(0.016)
Year dummies	NO	YES	YES	YES	YES
Macro-region dummies	NO	YES	YES	YES	YES
Sector dummies	NO	YES	YES	YES	YES
Firm-level controls [c]	NO	YES	YES	YES	YES
R^2	0.035	0.052	0.046	0.046	0.092
No. observations	10,642	10,642	9,643	9,143	7,566
No. firms	1,365	1,365	1,343	1,298	1,131

Robust standard errors in parentheses. * p<0.1, ** p<0.05,***p<0.01

[a] In column 4 we restrict the sample to firm with a firm-level contract.

[b] In column 5 of Panel B we restrict the sample to treated firms introducing PRP since 1995.

[c] Employment level and its squared, change in employment level, workers turnover, capacity utilization, share blue collars, export share, investment per worker, firm's age, M&A, hours worked, presence of a firm-level contract.

In order to partly address the drawbacks of parametric estimation, and to check the robustness of the results to a different methodology, I then perform non-parametric estimation based on propensity-score matching. These methods easily allow to evaluate the treatment effect at different points in time. Following Origo (2009), I will consider outcome in the short-run (one year after the treatment) and in the medium-run (three

years after the treatment). Firms are considered treated if they introduced a PRP scheme at time t; the control group is formed by all firms that never introduced PRP in the period considered (1984-2001). The horizon at which the effect is evaluated imposes different restrictions on the sample; for this reason, the propensity score has been estimated separately for the subsequent estimation of short- and medium-run effects. To estimate the propensity score, I include all available variables that could have an effect both on the choice of adopting PRP and on the outcome of interest. I also include lagged values of some variables; in particular, the inclusion of lagged values of the outcome variable is of great importance, given that I don't have a very accurate measure of productivity (I use real sales per worker). Nickell, Wadhwani, and Wall (1992), however, showed that sales are a good proxy of gross output in dynamic terms, and the introduction of lagged values of sales should reduce the impact of heterogeneus labor costs between firms[4].

Results of propensity score estimation are presented in tables 3.4 (short-run) and 3.5 (medium-run). In both specifications, the balancing property was satisfied, as reported in tables 3.6 and 3.7.

[4]The same measure of productivity is used by Origo (2009) and Bryson, Freeman, Lucifora, Pellizzari, and Pérotin (2010).

Table 3.4: Propensity score estimation - Short-run estimates

Dependent variable: adoption of PRP scheme		
	Coefficient	Standard error
Contemporaneous variables		
Export share	-0.313	0.578
Investment per worker	-0.003	0.004
No. employees	0.002	0.000
No. employees ^{2}	-0.000	0.000
Turnover rate	-0.502	0.264
Employment growth	1.121	0.528
Share blue collars	-0.858	0.812
Hours worked per worker	0.000	0.000
% Capacity utilization	-0.004	0.006
Firm age	0.006	0.002
M&A [a]	-0.063	0.156
Food and beverage	0.090	0.187
Textile	-0.692	0.126
Chemicals	0.247	0.146
Minerals	-0.042	0.173
Other manufacturing	-0.191	0.151
Nort-West	1.287	0.216
Nort-East	1.420	0.221
Center	1.106	0.225
Lagged variables		
Real sales per worker	0.764	0.249
Real sales per worker $t-2$	-0.555	0.244
Export share	0.510	0.586
Investment per worker	0.001	0.004
No. employees	-0.001	0.001
No. employees ^{2}	0.000	0.000
Share blue collars	1.180	0.848
Temporary lay-off hours per worker	-0.000	0.000
Hours worked per worker	-0.001	0.000
% Capacity utilization	0.008	0.006
Firm-level contract	0.773	0.186
No. observations	1,607	

The regression also includes year dummies. See notes to table 3.2 for a more detailed description of the variables used.

[a] Merger and acquisitions at t or at $t-1$, where t is the year of adoption of PRP schemes.

Table 3.5: Propensity score estimation - Medium-run estimates

Dependent variable: adoption of PRP scheme		
	Coefficient	Standard error
Contemporaneous variables		
Export share	-0.179	0.622
Investment per worker	-0.001	0.005
No. employees	0.001	0.001
No. employees ^{2}	-0.000	0.000
Turnover rate	-0.622	0.228
Employment growth	1.015	0.593
Share blue collars	-0.672	0.888
Hours worked per worker	0.000	0.000
% Capacity utilization	-0.002	0.006
Firm age	0.005	0.002
M&A [a]	-0.065	0.176
Food and beverage	0.308	0.209
Textile	-0.770	0.141
Chemicals	0.374	0.164
Minerals	-0.085	0.187
Other manufacturing	-0.167	0.163
Nort-West	-0.164	0.116
Center	-0.458	0.143
South	-1.606	0.243
Lagged variables		
Real sales per worker	0.895	0.285
Real sales per worker $t-2$	-0.721	0.280
Export share	0.271	0.630
Investment per worker	0.001	0.004
No. employees	-0.001	0.001
No. employees ^{2}	0.000	0.000
Share blue collars	1.195	0.924
Temporary lay-off hours per worker	-0.000	0.000
Hours worked per worker	-0.000	0.000
% Capacity utilization	0.006	0.006
Firm-level contract	0.683	0.204
No. observations	1,350	

The regression also includes year dummies. See notes to table 3.2 for a more detailed description of the variables used.

[a] Merger and acquisitions at t or at $t-1$, where t is the year of adoption of PRP schemes.

Table 3.6: Blocks of estimated propensity score - Short-run estimates

Inferior of block of pscore	No. Controls	No. Treated	Total
0.006	536	20	556
0.1	236	38	274
0.2	226	104	330
0.4	67	47	114
0.5	31	48	79
0.6	22	42	64
0.8	2	8	10

The final number of block is 7. This number of blocks ensures that the mean propensity score is not different for treated and controls in each block. The balancing property is satisfied. The common support option has been selected.

Table 3.7: Blocks of estimated propensity score - Medium-run estimates

Inferior of block of pscore	No. Controls	No. Treated	Total
0.005	448	20	468
0.1	183	23	206
0.2	177	86	263
0.4	92	71	163
0.6	21	64	85
0.8	2	13	15

The final number of block is 6. This number of blocks ensures that the mean propensity score is not different for treated and controls in each block. The balancing property is satisfied. The common support option has been selected.

In figure 3.1 I plot the distribution of the estimated propensity score by treatment status. The distribution is clearly quite different between treated and non-treated firms: it will thus be important, in the rest of the analysis, to restrict estimation to observations belonging to the common support.

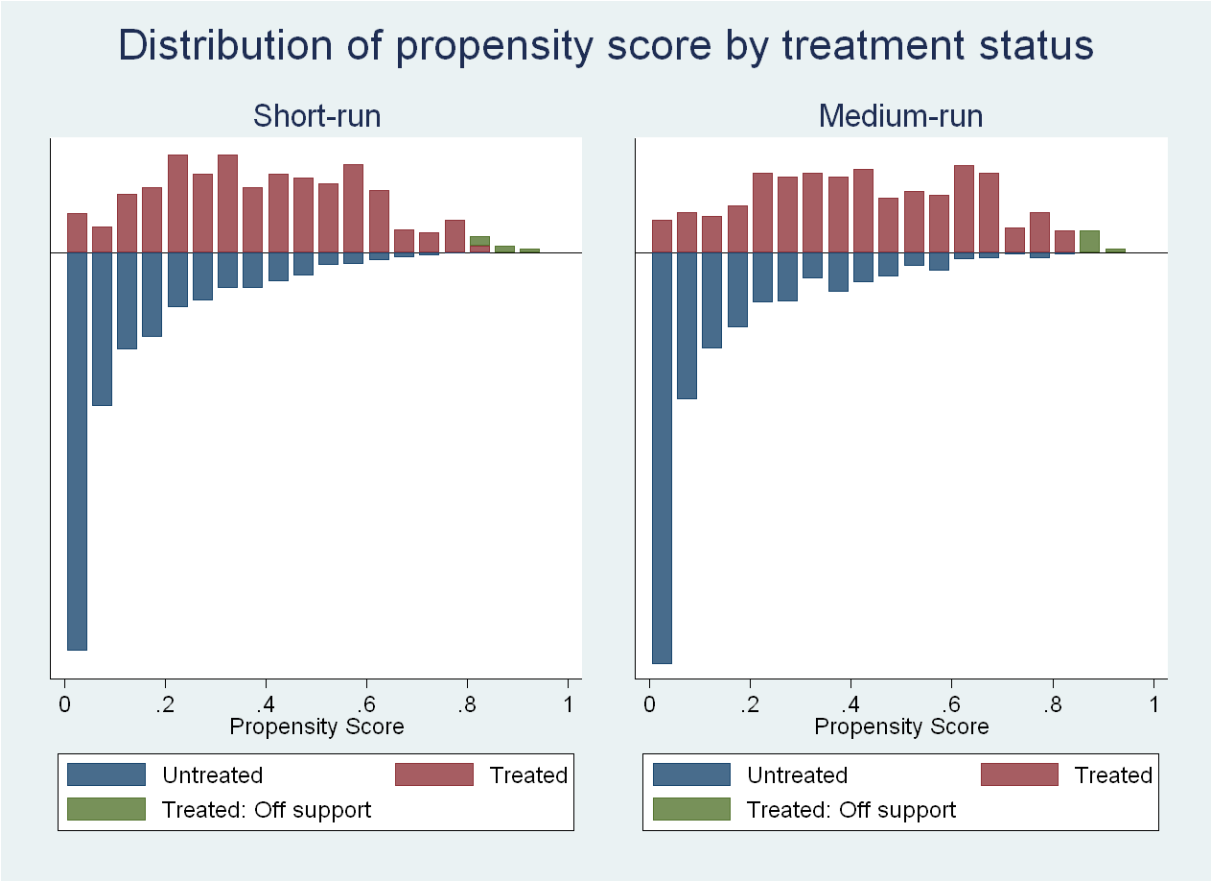


Figure 3.1: Propensity score distribution by treatment status

In table 3.8 I present a first set of estimates of short-run effects, both in levels (panel A) and using differences-in-differences (panel B). For robustness reason, I tried different matching algorithms: simple nearest-neighbour matching, caliper matching (which matches the nearest neighbour within a radius of 1% from the propensity score of the treated unit), matching to the 5 nearest-neighbours (within a 1% radius) and kernel matching (which uses all available controls, giving them different weights according

to the distance in the propensity score). Given that Abadie and Imbens (2008) have recently shown that bootstrapped standard errors are not valid for nearest-neighbour matching, in the table I always report analytical standard error (except for the case of kernel matching)[5]. The choice of different matching algorithms implies a bias-precision tradeoff. With nearest-neighbour matching, bias should be reduced by the fact that treated units are compared to the most similar control units; using a higher number of matches increases precision of the estimate, but also the bias, due to the fact that matches of average lower quality are selected. The table also reports a number of checks on matching quality[6], like the average standardised bias before and after matching[7], the pseudo-R^2 from a Probit regression of treatment status on covariates before and after matching, and the value of a likelihood ratio test of joint significance of covariates before and after matching. Overall, these statistics indicate that all matching procedures have been quite successful in balancing the distribution of observable variables across the two samples. The estimated effects in the short run range from 2% (using 5 nearest neighbours) to almost 10% (using caliper matching), but are never statistically significant. Estimated effects using a diff-in-diff strategies are much lower, ranging from 0.8 to 1.5%.

More consistent results (not changing a lot across different specifications) are obtained by loooking at outcome in the medium run, and are presented in table 3.9. The ATTs continue to be estimated quite imprecisely, and results are never statistically significant at conventional confidence levels. Estimated effects range from 3 to 6%, once again less than what found by Origo (2009) (between 5 and 12%).

[5]Bootstrapping standard errors for other algorithms as well does not significantly change the results.
[6]As suggested, for instance, by Sianesi (2004).
[7]The standardised bias is the difference of the sample means in the treated and non-treated sub-samples as a percentage of the square root of the average of the sample variance in the treated and non-treated subgroups (Rosenbaum and Rubin, 1985).

Table 3.8: Short-run PSM estimates

	Nearest Neighbour	Caliper[a]	5 Nearest neighbours[b]	Kernel matching[c]
	Panel A: Level estimates			
ATT	0.087	0.096	0.020	0.028
	(0.068)	(0.068)	(0.100)	(0.048)
No. treated	301	294	294	301
No. matched controls	189	187	1,361	1,300
Mean std. bias before matching [d]	25.287	25.287	25.287	25.287
Mean std. bias after matching	4.284	3.511	2.507	2.141
Pseudo R^2 before matching [e]	0.258	0.258	0.258	0.258
Pseudo R^2 after matching	0.030	0.014	0.014	0.005
LR χ^2 before matching [f]	403.950	403.950	403.950	403.950
LR χ^2 after matching	24.520	11.040	11.040	4.340
	Panel B: Diff-in-diff estimates			
ATT	0.007	0.008	0.013	0.015
	(0.028)	(0.028)	(0.018)	(0.018)
No. treated	301	294	294	301
No. matched controls	189	187	1,361	1,300
Mean std. bias before matching [d]	25.287	25.287	25.287	25.287
Mean std. bias after matching	4.451	4.284	3.511	1.408
Pseudo R^2 before matching [e]	0.258	0.258	0.258	0.258
Pseudo R^2 after matching	0.029	0.030	0.014	0.005
LR χ^2 before matching [f]	403.950	403.950	403.950	403.950
LR χ^2 after matching	23.850	24.520	11.040	4.340

Analytical standard errors in parentheses. Bootstrapped standard errors with 300 replications for Kernel estimates.

[a] Nearest-neighbour matching with a 1% caliper.

[b] 5 neighbours within a 1% caliper.

[c] Epanechnikov kernel with 0.06 bandwidth.

[d] The standardised bias is the difference of the sample means in the treated and non-treated sub-samples as a percentage of the square root of the average of the sample variance in the treated and non-treated subgroups, as in Rosenbaum and Rubin (1985)

[e] Pseudo R^2 from probit estimation of the conditional treatment probability.

[f] Likelihood-ratio test of the joint insignificance of all the regressors in the probit estimation of the conditional treatment probability.

Table 3.9: Medium-run PSM estimates

	Nearest Neighbour	Caliper[a]	5 Nearest neighbours[b]	Kernel matching[c]
	Panel A: Level estimates			
ATT	0.039	0.045	0.058	0.057
	(0.078)	(0.078)	(0.095)	(0.044)
No. treated	270	266	266	270
No. matched controls	170	170	1,154	1,073
Mean std. bias before matching [d]	26.378	26.378	26.378	26.378
Mean std. bias after matching	5.673	5.246	3.408	2.785
Pseudo R^2 before matching [e]	0.288	0.288	0.288	0.288
Pseudo R^2 after matching	0.052	0.049	0.014	0.010
LR χ^2 before matching [f]	394.640	394.640	394.640	394.640
LR χ^2 after matching	38.970	36.200	10.210	7.750
	Panel B: Diff-in-diff estimates			
ATT	0.052	0.048	0.046	0.032
	(0.034)	(0.033)	(0.031)	(0.029)
No. treated	270	266	266	270
No. matched controls	170	170	1,073	1,073
Mean std. bias before matching [d]	26.378	26.378	26.378	26.378
Mean std. bias after matching	5.673	5.246	3.408	2.785
Pseudo R^2 before matching [e]	0.288	0.288	0.288	0.288
Pseudo R^2 after matching	0.052	0.049	0.014	0.010
LR χ^2 before matching [f]	394.640	394.640	394.640	394.640
LR χ^2 after matching	38.970	36.200	10.210	7.750

Analytical standard errors in parentheses. Bootstrapped standard errors with 300 replications for Kernel estimates.

[a] Nearest-neighbour matching with a 1% caliper.

[b] 5 neighbours within a 1% caliper.

[c] Epanechnikov kernel with 0.06 bandwidth.

[d] The standardised bias is the difference of the sample means in the treated and non-treated sub-samples as a percentage of the square root of the average of the sample variance in the treated and non-treated subgroups, as in Rosenbaum and Rubin (1985)

[e] Pseudo R^2 from probit estimation of the conditional treatment probability.

[f] Likelihood-ratio test of the joint insignificance of all the regressors in the probit estimation of the conditional treatment probability.

It could be interesting to see if the timing of adoption of PRP schemes had some influence on its effectiveness. To do so, I ran exact matching based on treatment year (performing different estimates of the propensity score by each treament year) and then estimated year-specific treatment effects, that are plotted in figures 3.2 and 3.3. For short-run effects, no clear pattern emerges. The estimated effects are negative for firms introducing PRP in 1996, 1998 an 1999. As far as medium-run effects are concerned, PRP is apparently more beneficial to firms that have introduced it during years of economic slowdown (1992, 1993, 1996, 1997). Estimated effects, however, are never statistically significant.

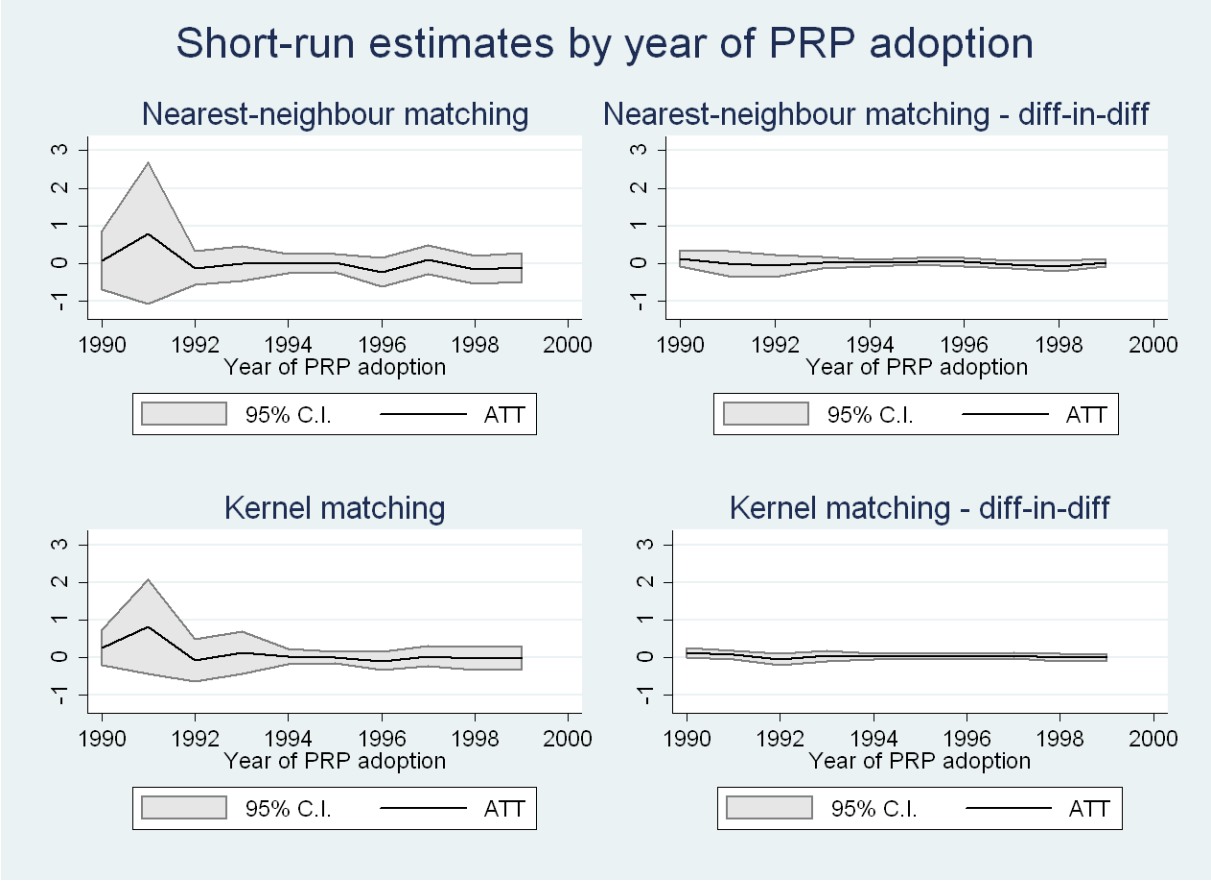


Figure 3.2: Short-run PSM estimates by year of PRP adoption

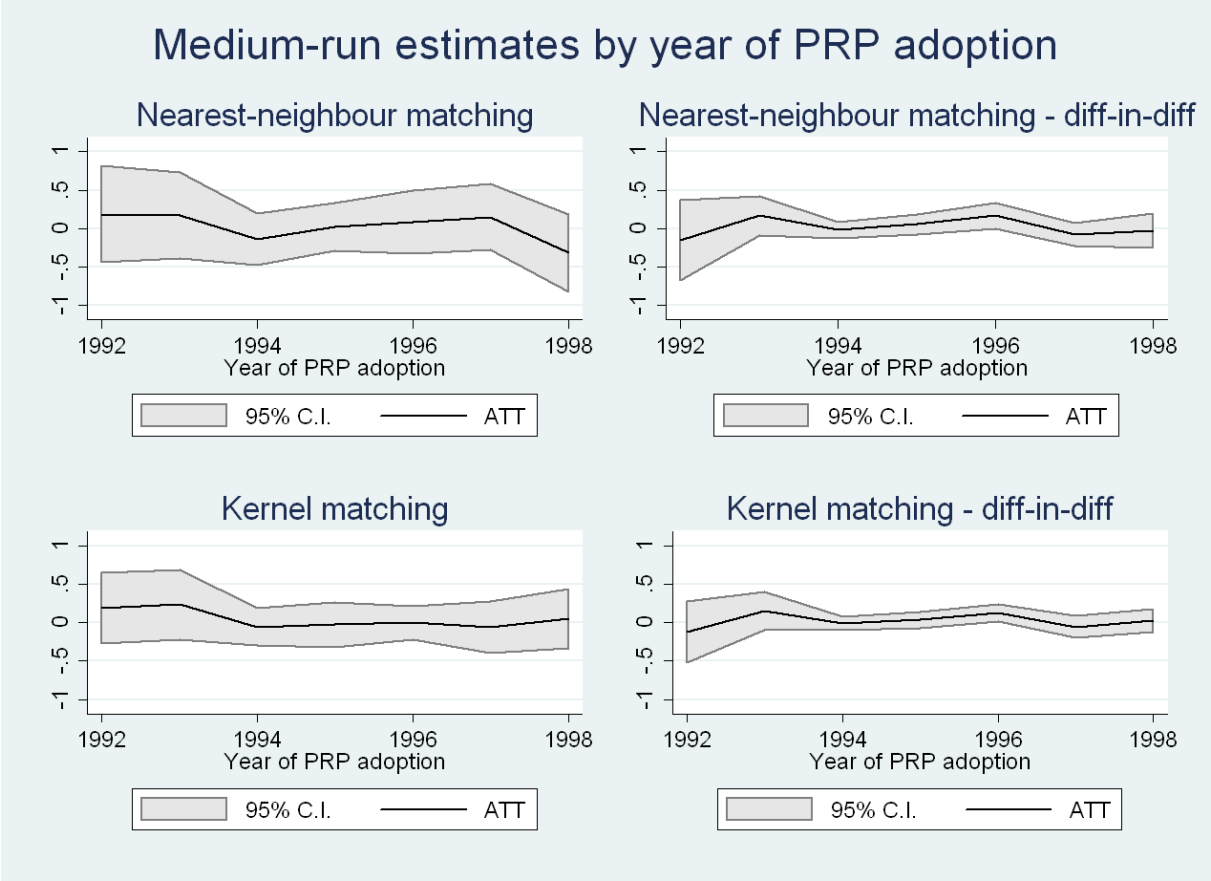


Figure 3.3: Medium-run PSM estimates by year of PRP adoption

Finally, I also tried to estimate treatment effects using an extended definition of the control group. In order to improve matching quality I used as controls also firms that would have introduced PRP in the future. In fact, for the estimation of short-run effects, a firm introducing PRP in year t could be a valid control for a firm that introduced PRP in year $t - j$, with $j > 1$. Similarly, for the estimation of medium-run effects, a firm introducing PRP in year t could be used as a valid control for a firm that introduced PRP in year $t - j$, with $j > 3$. The controls used in this "'larger sample"' should in principle be more similar to the matched treated units, both in terms of observables and unobservables characteristics, given that some of them would introduce PRP schemes

themselves in later years. Estimation results for this "larger sample" are reported in the appendix in figure 3.4 and in tables 3.10 and 3.11. Estimated effects of PRP adoption are even smaller than the ones obtained with the benchmark sample, and in some cases are even negative.

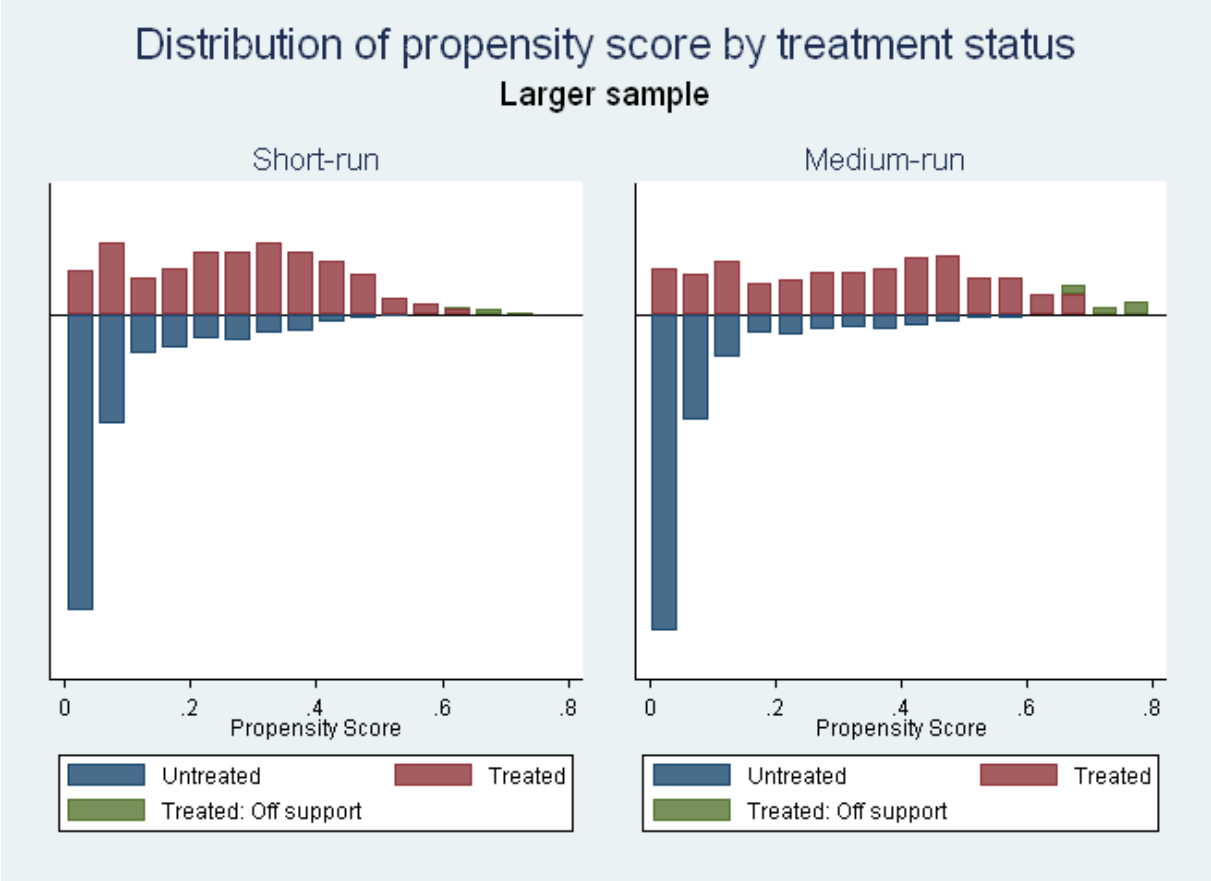


Figure 3.4: Propensity score distribution by treatment status - larger sample

Table 3.10: Short-run PSM estimates for a larger sample

	Nearest Neighbour	Caliper[a]	5 Nearest neighbours[b]	Kernel matching[c]
	Panel A: Level estimates			
ATT	0.006	0.002	0.019	0.034
	(0.058)	(0.058)	(0.049)	(0.023)
No. treated	303	300	300	303
No. matched controls	230	228	1,478	2,273
Mean std. bias before matching [d]	20.810	20.810	20.810	20.810
Mean std. bias after matching	4.929	5.145	2.208	2.201
Pseudo R^2 before matching [e]	0.217	0.217	0.217	0.217
Pseudo R^2 after matching	0.026	0.027	0.010	0.004
LR χ^2 before matching [f]	408.590	408.590	408.590	408.590
LR χ^2 after matching	21.500	22.400	8.000	3.710
	Panel B: Diff-in-diff estimates			
ATT	0.002	0.004	0.004	0.001
	(0.024)	(0.024)	(0.016)	(0.014)
No. treated	303	300	300	303
No. matched controls	230	228	1,361	2,273
Mean std. bias before matching [d]	25.287	25.287	25.287	25.287
Mean std. bias after matching	4.451	5.145	3.511	2.201
Pseudo R^2 before matching [e]	0.217	0.217	0.217	0.217
Pseudo R^2 after matching	0.026	0.027	0.014	0.004
LR χ^2 before matching [f]	403.950	403.950	403.950	403.950
LR χ^2 after matching	21.500	22.400	11.040	3.710

Analytical standard errors in parentheses. Bootstrapped standard errors with 300 replications for Kernel estimates.

[a] Nearest-neighbour matching with a 1% caliper.

[b] 5 neighbours within a 1% caliper.

[c] Epanechnikov kernel with 0.06 bandwidth.

[d] The standardised bias is the difference of the sample means in the treated and non-treated sub-samples as a percentage of the square root of the average of the sample variance in the treated and non-treated subgroups, as in Rosenbaum and Rubin (1985)

[e] Pseudo R^2 from probit estimation of the conditional treatment probability.

[f] Likelihood-ratio test of the joint insignificance of all the regressors in the probit estimation of the conditional treatment probability.

Table 3.11: Medium-run PSM estimates for a larger sample

	Nearest Neighbour	Caliper[a]	5 Nearest neighbours[b]	Kernel matching[c]
	Panel A: Level estimates			
ATT	-0.004	-0.041	0.027	0.055
	(0.087)	(0.088)	(0.066)	(0.041)
No. treated	199	197	197	199
No. matched controls	142	142	929	1,406
Mean std. bias before matching [d]	21.942	21.942	21.942	21.942
Mean std. bias after matching	5.958	5.954	4.000	3.404
Pseudo R^2 before matching [e]	0.282	0.282	0.282	0.282
Pseudo R^2 after matching	0.044	0.046	0.016	0.012
LR χ^2 before matching [f]	350.290	350.290	350.290	350.290
LR χ^2 after matching	24.350	25.080	8.840	6.460
	Panel B: Diff-in-diff estimates			
ATT	0.010	0.011	-0.005	-0.003
	(0.037)	(0.037)	(0.028)	(0.027)
No. treated	199	197	197	199
No. matched controls	142	142	929	1,406
Mean std. bias before matching [d]	21.942	21.942	21.942	21.942
Mean std. bias after matching	4.589	5.954	4.000	3.404
Pseudo R^2 before matching [e]	0.282	0.282	0.282	0.282
Pseudo R^2 after matching	0.044	0.046	0.016	0.012
LR χ^2 before matching [f]	350.290	350.290	350.290	350.290
LR χ^2 after matching	24.350	25.080	8.840	6.460

Analytical standard errors in parentheses. Bootstrapped standard errors with 300 replications for Kernel estimates.

[a] Nearest-neighbour matching with a 1% caliper.

[b] 5 neighbours within a 1% caliper.

[c] Epanechnikov kernel with 0.06 bandwidth.

[d] The standardised bias is the difference of the sample means in the treated and non-treated sub-samples as a percentage of the square root of the average of the sample variance in the treated and non-treated subgroups, as in Rosenbaum and Rubin (1985)

[e] Pseudo R^2 from probit estimation of the conditional treatment probability.

[f] Likelihood-ratio test of the joint insignificance of all the regressors in the probit estimation of the conditional treatment probability.

3.5 Conclusions

This paper has presented new estimates of the productivity effects of collective performance-related pay schemes, using a panel dataset from the Bank of Italy Survey of Manufacturing firms.

The richness of the dataset has allowed to use different econometric methodologies to address the fundamental problem of non-random selection into the adoption of such schemes and to evaluate the robustness of the results. Overall, the estimated effects of PRP are lower than what found by previous studies, and are often non statistically significant. In the short-run (1 year), PRP schemes do not seem to have a significant impact on productivity, while in the medium-run (3 years), estimated productivity gains are around 5%. Such results can be partly reconciled with recent findings by Origo (2009) in the light of the fact that the dataset used in this paper only includes medium and large sized companies (above 50 employees). In such firms, it is maybe not very surprising that collective schemes provide low incentives. This argument is supported by parallel evidence that the amount of bonuses paid to workers has so far been quite limited (Casadio, 2003, 2010), probably due to the prominent role played by unions, that traditionally favour an egualitarian treatment for all workers over schemes that, by rewarding different workers differently, also tend to increase wage inequalities.

In the absence of natural or experimental variation in contractual arrangements, the identification of such effects necessarily relies on assumptions that might be sometimes questionable. Experiments or pilot projects would provide more robust evidence, needed to better evaluate current policies aimed at promoting and incentivizing the diffusion of performance-related pay schemes.

Even though the empirical evidence provided so far by the economic literature generally indicates positive (or non-negative) productivity effects of PRP schemes, the

case for *active* government intervention in this area (through subsidies or other forms of fiscal incentives) is not obvious. Such policies are likely to primarily benefit firms that would have adopted these schemes in any case, or may induce firms to strategically change pay schemes with the only purpose of gaining tax advantages.

Bibliography

ABADIE, A. (2005): "Semiparametric Difference-in-differences Estimators," Review of Economic Studies, 72(1), 1–19.

ABADIE, A., AND G. W. IMBENS (2008): "On the Failure of the Bootstrap for Matching Estimators," Econometrica, 76(6), 1537–1557.

AMISANO, G., AND A. DEL BOCA (2004): "Profit Related Pay in Italy. A Microeconometric Analysis," International Journal of Manpower, 25(5), 463–478.

ATTANASIO, O., AND K. KAUFMANN (2009): "Educational Choices, Subjective Expectations, and Credit Constraints," NBER Working Papers 15087, National Bureau of Economic Research, Inc.

BAKER, G., R. GIBBONS, AND K. J. MURPHY (1994): "Subjective Performance Measures in Optimal Incentive Contracts," Quarterly Journal of Economics, 109(4), 1125–1156.

BANDIERA, O., I. BARANKAY, AND I. RASUL (2005): "Social Preferences and the Response to Incentives: Evidence from Personnel Data," The Quarterly Journal of Economics, 120(3), 917–962.

——— (2007): "Incentives for Managers and Inequality among Workers: Evidence from a Firm-level Experiment," The Quarterly Journal of Economics, 122(2), 729–773.

——— (2009): “Social Connections and Incentives in the Workplace: Evidence from Personnel Data,” Econometrica, 77(4), 1047–1094.

BARRINGTON-LEIGH, C. (2008): “Weather as a Transient Influence on Survey-reported Satisfaction with Life,” Draft research paper, University of British Columbia.

BECKER, S. O., AND A. ICHINO (2002): “Estimation of Average Treatment Effects Based on Propensity Score,” The Stata Journal, 2(4), 358–377.

BECKER, W. E., AND M. WATTS (1999): “How Departments of Economics Should Evaluate Teaching,” American Economic Review (Papers and Proceedings), 89(2), 344–349.

BELZIL, C. (2007): “Subjective Beliefs and Schooling Decisions,” Discussion Paper Series 2820, IZA.

BELZIL, C., AND M. LEONARDI (2007): “Can Risk Aversion Explain Schooling Attainments? Evidence from Italy,” Labour Economics, 14(6), 957–970.

BIAGIOLI, M., AND S. CURATOLO (1999): “Microeconomic Determinants and Effects of Financial Participation Agreements: an Empirical Analysis of the Large Italian Firms of the Engineering Sector in the Eighties and Early Nineties,” Economic Analysis, 2(2), 99–130.

BÉNABOU, R., AND J. TIROLE (2002): “Self-Confidence And Personal Motivation,” The Quarterly Journal of Economics, 117(3), 871–915.

BOWLES, S., H. GINTIS, AND M. OSBORNE (2001): “The Determinants of Earnings: a Behavioral Approach,” Journal of Economic Literature, XXXIX(4), 1137–1176.

BRANDOLINI, A., P. CASADIO, P. CIPOLLONE, M. MAGNANI, A. ROSOLIA, AND R. TORRINI (2007): “Employment Growth in Italy in the 1990s: Institutional Arrangements

and Market Forces," in Social Pacts, Employment and Growth. A Reappraisal of Ezio Tarantelli's Thought, ed. by N. Acocella, and R. Leoni, AIEL Series in Labour Economics, pp. 31–68. Physica-Verlag, Heidelberg.

BROWN, B. W., AND D. H. SAKS (1987): "The Microeconomics of the Allocation of Teachers' Time and Student Learning," Economics of Education Review, 6(4), 319–332.

BRUNELLO, G., AND M. SCHLOTTER (2011): "Non Cognitive Skills and Personality Traits: Labour Market Relevance and their Development in Education & Training Systems," IZA Discussion Papers 5743, Institute for the Study of Labor (IZA).

BRYSON, A., R. FREEMAN, C. LUCIFORA, M. PELLIZZARI, AND V. PÉROTIN (2010): "Paying for Performance. Incentive Pay Schemes and Employees' Financial Participation," European Conference XII, Fondazione Rodolfo Debenedetti.

BURGESS, S., C. PROPPER, M. RATTO, S. VON HINKE KESSLER SCHOLDER, AND E. TOMINEY (2010): "Smarter Task Assignment or Greater Effort: the Impact of Incentives on Team Performance," The Economic Journal, 120(547), 968–989.

CABLE, J., AND N. WILSON (1989): "Profit-sharing and Productivity: An Analysis of UK Engineering Firms," The Economic Journal, 99(396), 366–75.

CAHUC, P., AND B. DORMONT (1997): "Profit-sharing: Does it Increase Productivity and Employment? A Theoretical Model and Empirical Evidence on French Micro Data," Labour Economics, 4(3), 293–319.

CALIENDO, M., AND S. KOPEINIG (2008): "Some Practical Guidance for the Implementation of Propensity Score Matching," Journal of Economic Surveys, 22(1), 31–72.

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

CASADIO, P. (2003): "Wage Formation in the Italian Private Sector After the 1992-93 Income Policy Agreements," in Institutions and Wage Formation in the New Europe, ed. by G. Fagan, F. Mogelli, and J. Morgan, pp. 112–133. Edward Elgar, Cheltenham.

——— (2010): "Contrattazione Aziendale Integrativa e Differenziali Salariali Territoriali: Informazioni dall'Indagine sulle Imprese della Banca d'Italia," Politica Economica, 2, 241–292.

CESARINI, D., M. JOHANNESSON, P. LICHTENSTEIN, AND B. WALLACE (2009): "Heritability of Overconfidence," Journal of the European Economic Association, 7(2-3), 617–627.

CHECCHI, D., AND L. FLABBI (2007): "Intergenerational Mobility and Schooling Decisions in Germany and Italy: The Impact of Secondary School Tracks," IZA Discussion Papers 2876, Institute for the Study of Labor (IZA).

CHETTY, R., J. N. FRIEDMAN, N. HILGER, E. SAEZ, D. W. SCHANZENBACH, AND D. YAGAN (2011): "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence From Project STAR," Quarterly Journal of Economics, 126(4), 1593–1660.

CHEVALIER, A., S. GIBBONS, A. THORPE, M. SNELL, AND S. HOSKINS (2009): "Students' Academic Self-Perception," Economics of Education Review, 28(6), 716–727.

CHOWDRY, H., C. CRAWFORD, AND A. GOODMAN (2011): "The Role of Attitudes and Behaviours in Explaining Socio-economic Differences in Attainment at Age 16," Longitudinal and Life Course Studies, 2(1), 5–76.

CRISTINI, A., AND R. LEONI (2007): "The 1993 July Agreement in Italy: Bargaining Power, Efficiency Wage or Both?," in Social Pacts, Employment and Growth. A Reappraisal of Ezio Tarantelli's Thought, ed. by N. Acocella, and R. Leoni, AIEL Series in Labour Economics, pp. 97–119. Physica Verlag, Heidelberg.

CUNHA, F., AND J. J. HECKMAN (2007): "The Technology of Skill Formation," American Economic Review, 97(2), 31–47.

——— (2008): "Formulating, Identifying and Estimating the Technology of Cognitive and Noncognitive Skill Formation," Journal of Human Resources, 43(4), 738–782.

CUNHA, F., J. J. HECKMAN, AND S. M. SCHENNACH (2010): "Estimating the Technology of Cognitive and Noncognitive Skill Formation," Econometrica, 78(3), 883–931.

DAMIANI, M., AND A. RICCI (2008): "Flexible Wage Contracts and Firm Productivity: Evidence from Italy," mimeo.

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

DE GIORGI, G., M. PELLIZZARI, AND W. G. WOOLSTON (2011): "Class Size and Class Heterogeneity," Journal of the European Economic Association, forthcoming.

DELANEY, L., C. HARMON, AND C. REDMOND (2011): "Parental Education, Grade Attainment and Earnings Expectations among University Students," Economics of Education Review, 30(6), 1136–1152.

DENISSEN, J. J. A., L. BUTALID, L. PENKE, AND M. A. VAN AKEN (2008): "The Effects of Weather on Daily Mood: A Multilevel Approach," Emotion, 8, 662–667.

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

DUNNING, D., C. HEATH, AND J. SULS (2004): "Flawed Self-Assessment," Psychological Science in the Public Interest, 5, 69–106.

FALK, A., D. HUFFMAN, AND U. SUNDE (2006): "Self-Confidence and Search," Discussion Paper Series 2525, IZA.

FERLA, J., M. VALCKE, AND Y. CAI (2009): "Academic Self-Efficacy and Academic Self-Concept: Reconsidering Structural Relationships," Learning and Individual Differences, 19(4), 499–505.

FIGLIO, D. N., AND L. KENNY (2007): "Individual Teacher Incentives and Student Performance," Journal of Public Economics, 91, 901–914.

FILIPPIN, A., AND A. ICHINO (2005): "Gender Wage Gap in Expectations and Realizations," Labour Economics, 12, 125–145.

GIELEN, A. C., M. J. M. KERKHOFS, AND J. C. VAN OURS (2010): "How Performance Related Pay Affects Productivity and Employment," Journal of Population Economics, 23(1), 291–301.

GIULIANO, P. (2008): "Culture and the Family: An Application to Educational Choices in Italy," Rivista di Politica Economica, 98(4), 3–38.

GOLDHABER, D., AND M. HANSEN (2010): "Using Performance on the Job to Inform Teacher Tenure Decisions," American Economic Review (Papers and Proceedings), 100(2), 250–255.

GOODMAN, A., AND P. GREGG (eds.) (2010): Poorer Children's Educational Attainment: How Important are Attitudes and Behaviour?, Joseph Rowntree Foundation Report. London. Joseph Rowntree Foundation.

GREGG, P., AND E. WASHBROOK (2011): "The Role of Attitudes and Behaviours in Explaining Socio-economic Differences in Attainment at Age 11," Longitudinal and Life Course Studies, 2(1), 41–58.

HANUSHEK, E. A. (1979): "Conceptual and Empirical Issues in the Estimation of Educational Production Functions," Journal of Human Resources, 14, 351–388.

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

——— (2010): "Generalizations About Using Value-added Measures of Teacher Quality," American Economic Review (Papers and Proceedings), 100(2), 267–271.

HECKMAN, J., H. ICHIMURA, J. SMITH, AND P. TODD (1998): "Characterizing Selection Bias Using Experimental Data," Econometrica, 66(5), 1017–1098.

HECKMAN, J. J., H. ICHIMURA, AND P. TODD (1998): "Matching as an Econometric Evaluation Estimator," Review of Economic Studies, 65(2), pp. 261–294.

HECKMAN, J. J., H. ICHIMURA, AND P. E. TODD (1997): "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme," Review of Economic Studies, 64(4), 605–54.

HECKMAN, J. J., J. STIXRUD, AND S. URZUA (2006): "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior," Journal of Labor Economics, 24(3), 411–482.

HEINECK, G., AND S. ANGER (2010): "The Returns to Cognitive Abilities and Personality Traits in Germany," Labour Economics, 17(3), 535–546.

HEINRICH, C., A. MAFFIOLI, AND G. VÁZQUEZ (2010): "A Primer for Applying Propensity-Score Matching," IDB Publications 8292, Inter-American Development Bank.

HIRSCH, J. E. (2005): "An Index to Quantify an Individual's Scientific Research Output," Proceedings of the National Academy of Sciences of the United States of America, 102(46), 16569–16572.

HOFFMAN, F., AND P. OREOPOULOS (2009): "Professor Qualities and Student Achievement," The Review of Economics and Statistics, 91(1), 83–92.

HOGAN, T. D. (1981): "Faculty Research Activity and the Quality of Graduate Training," Journal of Human Resources, 16(3), 400–415.

HOLMSTROM, B., AND P. MILGROM (1994): "The Firm as an Incentive System," American Economic Review, 84(4), 972–991.

HVIDE, H. K. (2002): "Pragmatic Beliefs and Overconfidence," Journal of Economic Behavior & Organization, 48, 15–28.

ICHINO, A. (1994): "Flexible Labor Compensation, Risk Sharing and Company Leverage," European Economic Review, 38(7), 1411–1421.

ISTAT (1999): I Principali Risultati della Rilevazione sulla Flessibilità nel Mercato del Lavoro. Istat, Rome.

——— (2000): "La Flessibilità del Mercato del Lavoro nel Periodo 1995-96," Collana Informazioni 34, Istat, Rome.

JACOB, B. A., AND L. LEFGREN (2008): "Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education," Journal of Labor Economics, 26, 101–136.

JENSEN, R. (2010): "The (Perceived) Returns to Education and the Demand for Schooling," Quarterly Journal of Economics, 125(2), 515–548.

KANE, T. J., AND D. O. STAIGER (2008): "Estimating Teacher Impacts on Student Achievement: an Experimental Evaluation," Discussion Paper 14607, NBER Working Paper Series.

KAUFMANN, K. M. (2010): "Understanding the Income Gradient in College Attendance in Mexico: The Role of Heterogeneity in Expected Returns," Working Paper Series 362, IGIER, Bocconi University.

KELLER, M. C., B. L. FREDRICKSON, O. YBARRA, S. COTÉ, K. JOHNSON, J. MIKELS, A. CONWAY, AND T. WAGER (2005): "A Warm Heart and a Clear Head. The Contingent Effects of Weather on Mood and Cognition," Psychological Science, 16(9), 724–731.

KNETZ, M., AND D. SIMESTER (2001): "Firm-wide Incentives and Mutual Monitoring at Continental Airlines," Journal of Labor Economics, 19(4), 743–772.

KÖSZEGI, B. (2006): "Ego Utiliy, Overconfidence and Task Choice," Journal of the European Economic Association, 4(4), 673–707.

KRAUTMANN, A. C., AND W. SANDER (1999): "Grades and Student Evaluations of Teachers," Economics of Education Review, 18, 59–63.

KRUEGER, A. B. (1999): "Experimental Estimates of Education Production Functions," Quarterly Journal of Economics, 114, 497–532.

KRUSE, D. L. (1992): "Profit Sharing and Productivity: Microeconomic Evidence from the United States," The Economic Journal, 102(410), 24–36.

LAVY, V. (2009): "Performance Pay and Teachers' Effort, Productivity and Grading Ethics," The American Economic Review, 95(5), 1979–2011.

LAZEAR, E. P. (1986): "Salaries and Piece Rates," The Journal of Business, 59(3), 405–31.

——— (2000): "Performance Pay and Productivity," American Economic Review, 90(5), 1346–1361.

LINDQVIST, E., AND R. VESTMAN (2011): "The Labor Market Returns to Cognitive and Noncognitive Ability: Evidence from the Swedish Enlistment," American Economic Journal: Applied Economics, 3(1), 101–28.

MULLIS, I. V., M. O. MARTIN, D. F. ROBITAILLE, AND P. FOY (2009): TIMSS Advanced 2008 International Report. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

NGUYEN, T. (2008): "Information, Role Models and Perceived Returns to Education: Experimental Evidence from Madagascar," Working paper, MIT.

NICKELL, S., S. WADHWANI, AND M. WALL (1992): "Productivity Growth in U.K. Companies, 1975-1986," European Economic Review, 36(5), 1055–1085.

NUCCI, F., AND M. RIGGI (2011): "Performance Pay and Shifts in Macroeconomic Correlations," Temi di discussione (Economic working papers) 800, Bank of Italy, Economic Research Department.

NUTI, M. D. (1987): "Profit Sharing and Employment. Claims and Overclaims.," Industrial relations, 26(1), 18–40.

NYHUS, E., AND E. PONS (2005): "The Effect of Personality on Earnings," Journal of Economic Psychology, 26, 363–384.

OECD (2008): Education at a Glance. OECD Publishing, Paris.

——— (2009): Top of the Class. High Performers in Science in PISA 2006 Programme. OECD Publishing, Paris.

——— (2010): PISA 2009 at a Glance. OECD Publishing., Paris.

ORIGO, F. (2009): "Flexible Pay, Firm Performance and the Role of Unions. New Evidence from Italy," Labour Economics, 16(1), 64–78.

PAARSCH, H. J., AND B. SHEARER (2000): "Piece Rates, Fixed Wages, and Incentive Effects: Statistical Evidence from Payroll Records," International Economic Review, 41(1), 59–92.

PENDLETON, A., K. WITHFIELD, AND A. BRYSON (2009): "The Changing Use of Contingent Pay in the Modern British Workplace," in The evolution of the modern workplace, ed. by W. Brown, A. Bryson, J. Forth, and K. Whitfield, chap. 11, pp. 256–284. Cambridge University Press.

PRAY, M. C. (2011): "Some Like It Mild and Not Too Wet: the Influence of Weather on Subjective Well-Being," Cahiers de recherche 1116, CIRPEE.

PRENDERGAST, C. (1999): "The Provision of Incentives in Firms," Journal of Economic Literature, XXXVII (37), 7–63.

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

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

ROCKOFF, J. E. (2004): "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data," American Economic Review (Papers and Proceedings), 94(2), 247–252.

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

ROSENBAUM, P. R., AND D. B. RUBIN (1983): "The Central Role of Propensity Score in Observational Studies for Causal Effects," Biometrika, 70(1), 41–55.

——— (1985): "Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score," The American Statistician, 39(1), 33–38.

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

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

SCHWARZ, N., AND G. L. CLORE (1983): "Mood, Misattribution, and Judgments of Well-being: Informative and Directive Functions of Affective States," Journal of Personality and Social Psychology, 45(3), 513–523.

SHEARER, B. (2004): "Piece Rates, Fixed Wages and Incentives: Evidence from a Field Experiment," The Review of Economic Studies, 71(2), 513–534.

SIANESI, B. (2004): "An Evaluation of the Swedish System of Active Labor Market Programs in the 1990s," The Review of Economics and Statistics, 86(1), 133–155.

SJÖGREN, A., AND S. SÄLLSTRÖM (2004): "Trapped, Delayed and Handicapped," Working Paper Series 613, Research Institute of Industrial Economics.

SWAMY, P. A. V. B., AND S. S. ARORA (1972): "The Exact Finite Sample Properties of the Estimators of Coefficients in the Error Components Regression Models," Econometrica, 40(2), pp. 261–275.

TYLER, J. H., E. S. TAYLOR, T. J. KANE, AND A. L. WOOTEN (2010): "Using Student Performance Data to Identify Effective Classroom Practices," American Economic Review (Papers and Proceedings), 100(2), 256–260.

WEINBERG, B. A. (2009): "A Model Of Overconfidence," Pacific Economic Review, 14(4), 502–515.

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

WEITZMAN, M. (1985a): "The Simple Macroeconomics of Profit Sharing," American Economic Review, 75, 937–953.

WEITZMAN, M. L. (1985b): "Profit Sharing as Macroeconomic Policy," American Economic Review, 75, 41–45.