Selecting or rewarding teachers?

International evidence from primary schools♣

Michela Braga[a], Daniele Checchi[b],
Christelle Garrouste[c] and Francesco Scervini[d]

[a]Bocconi University, Milan (Italy)
[b]University of Milan (Italy) and IZA
[c]Paris 12 Val de Marne University (France)
[d]University of Pavia (Italy)

This version: January 15th, 2020

Abstract: Using data from the existing four waves of PIRLS, this paper examines the effect of teacher quality on fourth-grade students' literacy test scores by exploiting variations induced by reforms in teachers' selection and/or reward schemes. We construct an original data set of reforms taking place at the national level after World War II and affecting the working conditions of primary school teachers. We match these indicators of government reforming activities with the year of entry into the profession of each teacher, our identifying assumption being that new entrants are mainly affected by the reform, with the impact of reforms dissipating with the time passing by. We show that more selective recruitment processes are positively correlated to pupil achievements, while better teacher working conditions (including to some extent retirement policies) exhibit the opposite correlation.

JEL code: *H52 (Government Expenditures and Education), I21 (Analysis of Education), I28 (Government Policy), J44 (Professional Labour Markets – Occupational Licensing)*

Keywords: *Student achievements, PIRLS, Teacher recruitment, Teacher reward.*

♣Paper presented at the II SISEC conference (Milan, January 2018), EPCS annual meeting (Rome, April 2018), XXX SIEP annual meeting (Padua, September 2018), IX Workshop on Economics of Education (Barcelona, September 2018) and Fondazione Agnelli (Turin, December 2018). We thank all participants for their helpful comments. We also thank two referees that have helped us in streamlining and focussing the message of the paper. All remaining errors are our own responsibility.

1. Introduction

There is a widening literature on cross-country comparison on the effectiveness of national educational systems, using data from international surveys like PIRLS, TIMMS or PISA. In their review of the literature, Hanushek and Woessman (2011) discuss institutional differences associated to student curricula (pre-primary education, secondary school tracking) and to school management (school autonomy and accountability, competition between public and private sector). Most of these studies exploit cross-country variations, though the existence of a sufficiently long series of surveys allows now the inclusion of temporal variations. In few cases of federal countries (like Belgium, Canada and UK), whenever samples were representative at state level, within-countries variation has also been exploited. While quantitative input measures (typically educational expenditure) explain little of cross country variation, several studies show that measures of institutional frameworks and of the quality of the teaching input (for example local autonomy in hiring and/or compensating teachers) account for significant portions of international differences in level and dispersion of student achievements. Since information on teachers and teaching practices is limited, the first wave of studies reviewed in their survey leaves teacher policies in the background. More specific on teacher compensation were Woessman (2011) and Dolton and Marcenaro-Gutierrez (2011), matching OECD aggregate information on teacher compensation to student achievements. The first contribution identifies a positive relationship between pay-to-performance and mean test scores in reading and numeracy; the second one correlates the actual and relative compensation of teachers to student test scores, exploiting cross-country and over time variations to identify the impact of the former onto the latter variable. In a more recent contribution, Woessman (2016) reviews studies exploiting within-country variations to identify the effect of teachers' competences or compensation onto students' achievements. However, within this literature, the contributions more similar to ours are Nagler et al. (2018) and Hanushek et al. (2019). In the first paper, Nagler and co-authors exploit the date of entry into the teaching profession to show that variation in the outside opportunities induced by business cycle fluctuations affect the quality of aspiring teachers, which then induces variation in pupil test scores. They use administrative data from Florida schools to obtain a precise match between student and teacher characteristics. In the second paper, Hanushek and co-authors investigate whether differences in cognitive skills of teachers can explain international differences in student performance across countries. They exploit between-subject variation in teachers' skills to identify their effect onto within-student differences. They also consider the PIAAC sample as pseudo-panel for the female labour supply, showing that international differences in women's opportunities to enter (other) high-skill occupations provide part of the explanation for the observed variation in teacher cognitive skills across countries.

In the present paper, we contribute to the literature on the effects of teachers' quality onto student achievements by exploiting the cohort of entry into the profession. Instead

of using variations induced by outside labour market conditions (as in Nagler et al., 2018) and/or by career perspectives (as in Dolton and Marcenaro-Gutierrez, 2011, or more recently in Britton and Propper, 2016), we build a novel dataset on national policy reforms targeted to selection and compensation of teachers. We match teachers to the institutional framework prevailing at the date of entry in the profession and we show that whenever governments were actively reforming the hiring procedures in the teacher labour market (i.e. when entry in the teaching profession was made more selective), the corresponding pupil achievements rises, while the opposite happens whenever more favourable working conditions for teachers are negotiated and implemented. We focus on primary school where essential competences for future life are formed, and we use standardized test score at fourth grades from the four available waves of PIRLS; in these surveys, classes and their prevailing teacher are jointly surveyed every five years from 2001 to 2016. The development of reading, writing and numerical abilities in primary school is fundamental for one's intellectual capabilities in future life. Their deficiency makes any investment in skill formation in the subsequent stages of a student's educational process costlier and less effective. Indeed, any significant human capital accumulation requires solid foundations built in the early stages of the learning process. The peculiar nature of the surveys allows exploiting the within-country between-school/classes variation of teacher seniority into the profession to identify the impact of government actions onto teachers' quality. For each survey year, IEA (the managing agency for PIRLS) also provides detailed information on cross-country differences in teacher recruitment procedures, and we briefly review this as motivating evidence. However, we ignore under which rules the teachers in our sample were selected, and similarly we ignore what are their effective compensations. For these reasons, we resort to a juridical database ('Database of National Labour, Social Security and Related Human Rights Legislation' (NATLEX) by the International Labor Office-ILO's International Labour Standards Department) to reconstruct the reforming activities of government in four areas of the teacher labour contracts (selection, working conditions, compensation, and retirement). We do not have detailed information on the content of these reforms, and therefore our constructed variables can be considered as proxies of government attention to these issues. We find that, other things constant, teachers hired in periods when local governments were active on recruiting policies are associated to better performance of their pupils. On the contrary, if hired when government were easing their working conditions, they are associated to lower performance. While reforms to the pay systems are not correlated with pupils' achievement, for a subset of OECD countries we are also able to control for actual compensation, finding positive effects. Our policy conclusions emphasize the role of selective policies as a tool for improving the quality of the educational systems. We cannot obviously exclude other competing explanations, like the fact that different cohorts of teachers were trained to different teaching styles, which then reflect into higher pupils' performances.[1]

[1] Contributions in this perspective can be found in Schwerdt and Wuppermann (2011) and in Britton and Popper (2016). Both papers investigates the association of teaching styles (traditional lecture style versus

However, cross-country differences in the timing of reforming activities leave us confident in our identification strategy. Our results are obtained when controlling for a large set of pupil, teacher and classes/school characteristics, as well as for country×year fixed effects, which minimize the possibility of confounding factors. Several robustness checks confirm our main findings.

The rest of the paper is organised as follows. Section 2 reviews some literature on personnel economics in the teaching profession. Section 3 presents some motivating evidence on cross-national institutional varieties in teacher recruitment systems, while Section 4 presents the data used in our analysis. Section 5 contains the empirical strategy and our main results, while in Section 6, we perform robustness checks and sensitivity analysis to corroborate our results. Finally, Section 7 concludes the paper.

2. Policies for selecting and motivating best teachers

The literature on Human Resource management suggests that there are three main goals for any employer regarding manpower: recruiting (which implies attracting and selecting), motivating and then retaining the employees to achieve their highest level of productivity (Lazear and Gibbs, 2007). Because the ability to teach is a job-specific human capital, retaining teachers is usually not a main concern for school principals and policy makers at large although, in the literature, alternative opinions are considered (Moor Johnson, 2006). On the contrary, teacher selection and motivation are at the core of any attempt to improve the quality of educational systems. In their application of personnel economics to the public sector, Cameron et al. (2016) put emphasis on intrinsic motivation of public employees whenever performance is imperfectly observable and/or performance is not contractible. They argue that a key factor in the success of some public administration with respect to others is designing internal personnel policies (especially wage and promotion standards) that build cadres of highly motivated and capable civil servants. Teachers represent a perfect case study in this respect, since the goals assigned to them are imprecise and multidimensional. Children achievements (as measured by test scores, in levels or in terms of value added) constitute only one of the many dimensions of teacher engagement, since socialization (i.e. education to societal values) does represent as much an important task assigned to teachers. If we then add the issue of inducing curiosity to culture, we realize that one of the main concerns for policy makers is the selection of teachers who have an adequate intrinsic motivation.

When selecting on unobservable, probation periods, possibly associated to steeper wage schedules after tenure, do represent a suitable strategy: "*A second alternative* [to the impossibility of performance contracting] *is to attract and then differentially promote or*

innovative problem solving style) and student test scores, finding positive effects of traditional teaching onto factual knowledge abilities and routine problem solving.

retain intrinsically-motivated individuals ("zealots") who – in contrast with purely financially-motivated "slackers" – find employment as public sector managers inherently satisfying." (Cameron et al. 2016, pg.2). Thus public wage and promotion standards become crucial as selecting devices in order to attract effective teachers. This is consistent with the evidence proposed by Dolton and Marcenaro-Gutierrez (2011) on teachers' pay differentials across 39 OECD countries, who showed that recruiting more talented individuals into teaching and permitting quicker salary advancements have a positive effect on pupil outcomes.

Thus, selection on motivation and compensation policies are two available instruments in the hands of school principals, local educational authorities and/or central governments whose aim is to improve the quality of the schooling system. However, selection criteria are often based on educational credentials, though there is mixed evidence on their correlation with student achievements;[2] in addition, sorting into the teaching profession is significantly affected by outside options created by business cycles.[3]

However, high selectivity alone at the entry of the profession may be ineffective or even counterproductive, discouraging good potential candidates.[4] Thus selection at the entry should be combined with wage ladders and internal promotion rules that continue the

[2] In the literature review Hanushek and Rivkin (2006) do not report any consistent relationship between the level of credentials of teachers and corresponding student achievement. Various country specific studies find more mixed evidence: Santibañez (2006) on student achievement in Mexico finds a small positive relationship between teacher test scores and average student achievement scores; in a study on Sweden Andersson et al. (2011) show that the share of noncertified teachers decreases students' grade; Harris and Sass (2011) do not find any evidence that teacher preservice (undergraduate) training or college entrance exam scores being related to student achievements in US schools. Equivalent results are also obtained by Kane et al. (2008) regarding newly hired teachers in the New York City public schools (the initial certification status of a teacher has only small impacts on student test performance) and by Buddin and Zamarro (2009) in Los Angeles primary schools, where they show that neither the teacher licensure (*a regime where schools are forbidden from hiring teachers who have not completed a program of study in a teacher education program*) test scores nor the possession of an advanced degree are related to student achievement.

[3] Bacolod (2007) shows that the U.S. experienced a marked decline in the quality of young women entering teaching between 1960 and 1990, contrasting with a simultaneous increase in the quality of those who became professionals. Similarly, Falch et al. (2009) measure teacher shortages in Norway as the share of teachers without certified credentials, finding a negative relationship between teacher shortages and regional unemployment rates in the period 1981–2002. Nagler et al. (2018) obtain analogous results by exploiting business cycle condition at a teacher's start of career as a source of exogenous variation in the outside options for potential teachers. Similarly Carroll at al. (2018) show that the relative returns to education across occupations for men and women can explain vocational choices in the Australian context (and in particular gender segregation, with female teachers mostly concentrated in preprimary and primary schools, while male tend to specialize in secondary schooling and administrative roles). Thus, according to these authors, the teaching profession would remain a residual one because of the lack of career advancement, leading to a counter-cyclical selection into the teacher profession.

[4] "*High-performing countries use various mechanisms to select the best candidates to the teaching profession. In Finland, Hong-Kong (China), Macao (China) and Chinese Taipei, students who wish to enter teacher-training programmes must pass a competitive entry examination. In Japan, teaching graduates must pass a competitive examination to start teaching and in Singapore, they must complete a probation period. These requirements, however, are also found among some low-performing countries suggesting that early selection, while important, is not enough to ensure a highly qualified teaching force.*" (OECD 2017, pg. 2).

screening over the entire working life cycle. In this respect, professional development activities, particularly those that promote teacher collaboration, can reveal effective in forging a high quality teacher.[5]

If good teachers are to be retained in the teaching profession and supported in doing their work – and doing it well – they should have a workplace that promotes their efforts in a variety of ways (Moor Johnson, 2006). Since the 1980s, the United States and United Kingdom have passed measures to implement performance-based incentives, that is, monetary benefits to teachers and/or school principals who are considered the best according to the level of (or the variation in) their students' achievements. However, these policy measures have proven to have contradictory effects.[6] Indeed, because performance-based incentives are not easy to introduce in public schools, most countries have instead opted for reforms that unconditionally increase the level of teacher salaries. These measures have been found to be significantly correlated to student achievement in Dolton and Marcenaro-Gutierrez (2011), but it remains unclear through which channels monetary incentives play a role. As Dohmen and Falk (2010) have clearly shown, the presence or absence of monetary incentives in the teaching profession induces the self-selection of different individuals.[7]

Whether these two policy instruments, selection and reward, are substitutive or complementary in nature is hard to judge because self-selection occurs based on unobservable characteristics, which in turn can be correlated to (unobservable) teachers quality. Merit pay wage policies should attract people who are expecting to benefit the most from such a scheme, but whether they are better able and/or greedier than average is difficult to gauge: as a consequence, it is almost impossible to predict what the overall effect on student achievement will be because the 'selection' and 'incentive' effects may work in opposite directions. If, therefore, it is impossible to derive uncontroversial predictions about what the most effective teacher policies are to improve school quality,[8] we do not have other alternatives than taking these questions to the data. In the

[5] *"High-performing countries try to attract the most promising candidates to the teaching profession early on, but they also understand that talent can and must be nurtured through high-quality training and continuous learning. These countries strive to boost teachers' knowledge base, enhance the professional qualifications of teachers and involve them in professional development activities, particularly teacher collaboration*." (OECD 2017, p.6).

[6] Atkinson et al. (2009) find that a performance-related pay scheme implemented in the UK did improve test scores and the value added increased on average by about 40% of grade per pupil. In an earlier study on US, Ballou (2001) showed that efforts to implement merit pay in public education have generally been unsuccessful, mainly because of the opposition from teachers and teachers' unions. In Israel, Lavy (2015) reports persistent gains in labour market achievements of students whose teachers have been exposed to pay-to-performance schemes.

[7] '*…teachers are more risk averse than employees in other professions, indicating that relatively risk adverse individuals sort into teaching occupations under the current system. Using survey measures on trust and reciprocity, we find that teachers trust more and are less negatively reciprocal than other employees'* (Dohmen and Falk, 2010, p. F256).

[8] Similar wide-ranging lessons are to be learnt in comparative cross-country analysis: *"A variety of approaches to selecting and evaluating teachers, and a wide range of career and compensation structures for teachers, can be found across the best-performing countries in PISA. But at least three elements tend to be common to high-performing countries' professional development policies for teachers: a mandatory*

next section, we provide some descriptive evidence on the varieties of institutional models adopted by different countries in teacher recruiting and rewarding policies, while, in the next sections, we provide more precise identification of the effect of reforms onto teachers' effectiveness in raising student achievement.

3. Descriptive evidence on selection and reward of teachers

In the present paper we study primary school achievements relying on four PIRLS surveys, conducted by the International Association for the Evaluation of Educational Achievement (IEA), aiming to test the reading literacy competences of fourth graders.[9] Though not fully representative of the dynamics of all levels of schooling, the use of primary school data is appropriate for our purpose, since the existence of a prevailing teacher in each class allows for a more precise identification of the teacher's effect onto pupils' achievements. This has obviously drawbacks, since it prevents the analysis of more complex cognitive abilities of students (like numeracy and scientific reasoning) as well as studying the contribution of teaching practices, since 4th graders are typically unable to report on them. Thus, our results do not necessarily extend to the whole educational process in each country; nevertheless, it remains rather indicative of the overall attitude of educational authorities with respect to teachers' policies. Nevertheless, in primary school pupils acquire those abilities and skills that are essential for further stages of their education.

In addition to microdata on student achievements and teacher characteristics discussed in the next Section, the PIRLS survey also collects some information about the institutional framework prevailing for primary and lower secondary school level in each country/wave, information that is provided by national country experts. With the goal of maximising country coverage to emphasise institutional diversities, we have selected five dimensions of a country/region recruitment system whose presence/absence could make the teaching profession more or less selective, hence potentially affecting the future quality of aspiring teachers:

i) having a compulsory training period before/during the teacher educational programmes required for teaching (dummy variable TRAIN);

ii) having an official process to license or certify teachers by one institution (variable EXAM);

---

and extended period of clinical practice as part of pre-service teacher education or of the induction period; the presence of a variety of bespoke opportunities for in-service teachers' professional development, such as workshops organised by the school; and teacher-appraisal mechanisms with a strong focus on teachers' continuous development. […] Countries with higher teachers' salaries (relative to GDP) had, on average, larger shares of students who expected to work as teachers. And while in all countries girls were more likely to expect a career in teaching than boys, students' expectations of a teaching career were more gender-balanced in countries with higher teachers' salaries. However, there is no evidence that higher salaries attract high-achieving students into the teaching profession more than low-achieving students." (OECD 2018a, p.11).

[9] More details on the PIRLS assessment are provided in the Section 4.

iii) having a compulsory probation period (dummy variable PROBATION);
iv) the length of the potential probation period at the early stages of a teacher's career (continuous variable PROB_LENGTH);
v) having a mentoring programme for teachers (dummy variable MENTOR).
There are two additional variables, but only available in a subset of waves:
vi) passing a standardised test or an official examination as a basic requirement for teaching (dummy variable TEST);
vii) receiving a specific preparation on teaching techniques for reading (dummy variable TECHNIQUES).
The descriptive statistics for these variables are reported in Table 1, while their correlation matrix is in Table A.1 of the online appendix. In order to summarise the information contained in these institutional indicators, we have applied a factor analysis, which suggests the existence of at least two latent variables that account for 60% of the observed variance and are orthogonal by construction (see Table A.2 of the online appendix):[10,11]

a) the first one gathers the contribution of the variables PROBATION, PROB_LENGTH and MENTORING, since the highest factor loadings are associated to these variables. Considering that both institutional features are referred to a selection of teachers after having observed them on the job, we call this first factor SELECTION EX-POST.
b) the second one collects the contributions of the variables TRAIN and EXAM. In such a case, the selection tends to occur before experimenting an aspiring teacher on the job. For this reason we call this second factor SELECTION EX-ANTE.
If we observe the distribution of countries along these two latent dimensions (see Figure A.1 of the online appendix), we notice that a group aligns along a sort of frontier of the "maximal selectivity" combining, in different ways, the selections based on certification/training and on probation/mentoring (notably France, the Netherlands, Canada, but also US, UK-England, Singapore and Japan). How do these institutional features correlate with pupils' achievements? In Table 2 we report the unconditional correlations between the two latent variables and the mean test scores in reading. It is apparent that achievements are higher in countries where teachers are ex-ante selected through training and/or examinations (see also Figures A.2 and A.3 of the online appendix).

[10] If we apply the factor analysis to the seven variables indicated in the text, we lose 43 observations (mostly in the last two surveys) and obtain indication on the existence of three factors, accounting for 67% of total variance. After rotation, the first factor coincides with the first one reported in the text, while the second and the third gather respectively EXAM-TEST and TRAIN-TECHNIQUES. Given the descriptive nature of this section, we have preferred to stick to the simplest version of two factors. This additional factor analysis is available from the authors.
[11] Given the large majority of dummies variables, one could resort to the polychoric variance-covariance matrix to obtain the eigenvectors. However, given the high correlation of the two methods (the first latent variables extracted with the two methods are correlated at 0.99, while the secondo components are correlated at 0.94).

In the previous section we have also argued that selection *per se* may be insufficient, because of self-sorting of aspiring teacher may be driven by expected wage for teachers. For this reason, we complemented the data on the institutional setting for primary schools with information on the corresponding average pay in each country. The average pay earned by primary school teachers is an indicator of the relative attractiveness of the profession compared with other professions that require similar qualifications in terms of education. Higher relative pay for teachers should attract better candidates (in terms of both observable and unobservable credentials) and/or enhance their quality in terms of skills and motivation. Hence, from various issues of the OECD's *Education at a Glance* (the most recent one being OECD, 2018b), we collect the ratio of primary teacher salaries to GDP per capita for each wave of PIRLS. Unfortunately, information on wages is available only for a subsample of 37 OECD countries. Going back to Table 2, we observe two facts. More generous pay policies are associated to selection procedures that involve lengthy probation and/or mentoring by senior colleagues; however, irrespective of the relative generosity of pay conditions, average pupils' achievements seem not correlated to teachers' wage (see Figure A.4 of the online appendix – the apparent positive correlation is entirely attributable to the Indonesian observation). The first fact is consistent with theoretical expectations, since a lengthy induction period can attract good candidates if and only if the expected future wage compensates for current reduced opportunities (Garibaldi, 2006). The second fact seems in contradiction with part of the literature (e.g. Dolton and Marcenaro-Gutierrez, 2011), but should be qualified on at least two details: the wage variable consists of an average across different educational levels and different levels of tenure, and the correlations are unconditional (i.e., they do not control for compositional differences across countries and surveys).

However, even if these results are simply suggestive of the importance of the cross-country institutional differences in teacher policies, they do not allow a clear identification of their impact onto teachers' quality. Indeed, these measures of recruitment attitudes are effective at the time of the survey and therefore affect the quality of teachers entered since their last change and all future teachers.[12] On the contrary, the wage policy affects the working conditions of the incumbent teaching staff, as well as the attractiveness of the profession for future candidates. Thus, they are not fully comparable since they affect different populations; in addition, it is difficult to identify who are the teachers treated by these measures.

For these reasons, we prefer an alternative strategy that exploits within-country/wave variations induced by teachers' seniority that allows for the identification of the year of entry into the teaching profession. In this way, even controlling for confounding effects

[12] One could argue that variations in these institutional (latent) variables could be exploited to identify potential effects on teacher qualities (see Figure A.5 in the online appendix). However, we cannot exclude that these variations do represent different coding patterns of different national experts (thus nothing but a measurement error), especially when we try to check their consistency overtime using alternative sources.

by country×wave fixed effects, there is still variability at class level induced by the seniority of teachers, which we correlate with our measure of reforming activities of governments. This has some cost in terms of institutional detail, since we could not find detailed information about teacher selection at hiring going back 70 years. At best, we have been able to reconstruct information on reforming activities of governments, that capture the government attention to the recruitment and working conditions of teachers.

The identification strategy and the description of the data and the results are the content of the next sections.

4. Data and Descriptive Statistics

To test the effect of teachers on pupils' performance, we use microdata on students' achievements drawn from the four available waves of the PIRLS assessment together with information about the country-level reforming activity relevant for teachers collected from secondary data sources. Let us briefly describe the content of the data and present the basic descriptive statistics.

The PIRLS assessment tests the reading literacy of fourth graders. Starting from 2001, PIRLS has been administrated every five years, covering seventy country/state/region entities with legal autonomy in educational policy making. The study defines reading literacy as the ability to understand and use the written language forms required by society and/or valued by the individual. Three dimensions are assessed: the processes of comprehension, the purposes of reading and reading behaviours and attitudes. Test scores measure student performance in reading literacy and they are standardised to an international mean of 500 with a standard deviation of 100.

The test scores are intended to be nationally representative. National samples are drawn through a two-stage stratified sampling design. First, the participating schools are randomly selected. Then, within each school, a random sample of classes from the targeted grade is drawn and, within each class, all the students participate in the assessment. Together with students' reading achievement scores, the survey collects detailed background information on students, parents, teachers, schools and curricular activities. The questionnaires are administered to the tested students, to their parents, to their reading teachers and to their school principals. The teachers' information refers to the main or unique reading teacher of the class.[13]

[13] The structure of the data set is nested, with four levels of information aggregation: pupil – class and teacher – school – country. To have a perfectly nested sample, we dropped the very few (less than 1%) classes with more than one teacher of 'reading', while the inclusion of different classes with the same teacher is less harmful, unless one argues that teacher quality declines with the number of classes taught (this happens only in 611 classes over the 43,367 analysed). Results are robust to the exclusion of these classes.

Table 3 provides a summary of the statistics of the core variables used to perform the empirical analysis. Besides the reading test scores, we include individual socio-demographic features, school and teacher characteristics. Among the individual features potentially responsible for the differences in performance, we consider gender, age at the date of the survey and language spoken at home. When available, socio-economic background is proxied by parental education as the highest level of education of either parent. We also include the index of home educational resources based on the number of books at home, having a computer, a tablet, a study desk/table for own use, newspapers and internet connection. The school features refer to the geographical location, the presence of a library for students and the share of disadvantaged pupils. Among teacher characteristics, we focus on gender, age, tenure, having a graduate degree and number of students in the typical reading class. We restricted the analysis to those students with a complete record of data related to these dimensions. Missing values on individual, school or teacher characteristics account for less than 3 per cent of the initial sample: this share is reasonable in any survey and it does not raise particular concerns. On the contrary, data on parental background are missing for a higher fraction of the sample, with huge variability both across countries and over time, and in the US the parental questionnaire was not even administrated. For this reason, to account for potential selection problems in Section 6 we deal with missing values in parental background. Finally, it is important to note that the sample of countries is not balanced since some countries are missing in some waves. Table 4 reports the list of countries included in our analysis according to the set of available variables required in the empirical analysis. Namely, we consider thirty-four countries/state/regions in 2001, forty-one in 2006, fifty-one in 2011 and 2016.

To address our research question regarding the role of selection and reward policies on teachers' quality, starting from the paper by Garrouste (2010), we collected detailed information about the reforming process of the teacher profession that took place over the last 70 years and we assembled a final dataset containing yearly information for the period 1947–2016 covering all countries surveyed by PIRLS. Information is drawn from the 'Database of National Labour, Social Security and Related Human Rights Legislation' (NATLEX) produced by the International Labor Office-ILO's International Labour Standards Department. The database lists and classifies all the legislative actions in fields broadly related to the labour market and working conditions from the mid-1800s. Among them we focused on the following categories 'Constitutional laws', 'Labour codes, general labour and employment acts', 'Economic and social policy', 'Education, vocational guidance and training', 'Conditions of employment', 'Conditions of work', 'Social security', 'Employment security, termination of employment' and 'Specific categories of workers ⇨ teachers'. Furthermore, we focused only on those acts targeted to primary school teachers since our search strategy included these two keywords thus leaving out any potential reform affecting all the other public employees.

Since the selection and rewarding mechanisms potentially affect the quality of teachers, among all the legislative actions recorded, we selected those relevant to our scope identifying four reforming areas: *i*) reforms affecting teacher recruitment processes; *ii*) reforms on teacher working conditions; *iii*) reforms on the pay scheme; and *iv*) reforms affecting retirement possibilities. The first group of reforms refers to the *ex-ante* selection process to become a primary school teacher, while the next three reform areas involve different dimensions of rewarding. Among the reforms of the recruitment process, we included those that are changing the prerequisite criteria, through changes in the minimum marks to enter teacher colleges, in the level of educational attainment or in the prerequisites for teacher certification or licensing. Reforms of the working conditions refer to changes in the working hours, in the legal rights for special leaves or in continuous training. Reforms of pay include changes in the wage policy towards teachers, either as a part of a global civil servant reform or as a teacher-specific measure, which often comes from pressure by teacher unions. Finally, reforms of the retirement rules include legislative changes in the retirement entitlements that are specific to teachers, since we are interested in the incentive mechanisms of teacher attractiveness versus other careers prospects (e.g prerequisites for early retirement and/or the level of pension benefits).

For each legislative change we identified the year of implementation, as well as the direction of the change marked by the policy-makers (i.e., whether it was favourable or unfavourable for teachers). Whenever, in a given year we recorded a change in a specific dimension, we assigned a value of one from then onwards, while if no changes occur, we assigned a value of zero. When legislators have repeatedly reformed a specific dimension over the sample period, our created step dummies were summed over the years. The time plot of these variables is reported in Figures A.6 in the Appendix, while the original timing of the reforms is reported in Table A.3 of the Appendix.

We then normalized our step dummy variables in a unitary range of variation and we ended up with four final indicators of the reforming activity at country/year level. The first index refers to changes in the teacher recruitment process, with an increase corresponding to more restrictive selection criteria. The second one is related to changes in working conditions, an increase referring to more favourable working conditions (workload, holidays, standard requirements and the like). The third indicator is defined according to changes in the wage policy and salary conditions; also in this case, an increase means more generous wage allowances for primary school teachers. Finally, the fourth one captures variations in the stringency of retirement conditions including the retirement allowance, severance pay and retirement age: the indicator increases whenever retirement conditions become more favourable. Table 5 reports the summary statistics of our normalized reform variables while in Table 6 we report the pairwise correlation matrix among them. Notice that the decomposition of the standard deviation into the 'between' and the 'within' components indicates that there is a significant variation within countries.[14] We then match these indicators to the teachers surveyed by

[14] In a previous version of the paper, with fewer countries and waves we used as data source for the European countries also *Eurybase*, the *Eurydice* database providing detailed information on European

PIRLS according to the year when they entered into the labour market. As an example, consider the case of our 'teacher recruitment reform indicator' in a country where the data source reports more stringent reforms in 1987, 1988 and 1990. Therefore, our index is constructed as a variable that is coded zero before 1987, one-third in 1987, two-thirds in 1988 and 1989, and one afterward. Every teacher entering the profession in that country before 1987 gets a zero value for this reform variable, those entering in 1987 get one-third, those in 1988–1989 receive two-thirds and teachers hired more recently obtain a value of one.

Regarding to our new dataset on reforming activity three aspects must be discussed. First, the constructed indicators for the reforming activity is based only on the direction of the legislative change. An increase or decrease in the variable refers to a legislative change that is favourable or unfavourable to teachers, but we are not able to quantify the size of its potential impact with respect to the preceding situation (either in terms of coverage among teachers or in terms of individual change induced by the reform) thus being unable to distinguish between "major" and "minor" reforms. Second, although comprehensive and constantly updated, the NATLEX database could unintentionally misreport or omit some legislative act or regulation. Symmetrically, specific collective agreements regulating contracts in the private sector are not recorded. However, it collects legislation actions with *erga omnes* effects, which are more interesting for our purpose. Furthermore, NATLEX may not report wage adjustments not requiring an explicit normative act, such as price indexing. Third, the database classifies legislative intervention recording year of adoption and entry into force of the law with no more details on when specific regions or federal states could have actually implemented it. For all these reasons, our estimates are likely to be downward biased because of measurement errors in the reforming variable. However, given the impossibility to appropriately measure the true size of the impact, it is also impossible to assess the size of the bias.[15]

5. Empirical Strategy and Main Results

The aim of the empirical analysis is to identify whether some policies intended to *attract, select, reward* and/or *motivate* good teachers, who can improve student performance in primary schools, are actually effective. Whether teachers matter for student performance should be tested by correlating student achievement with direct measures of teacher quality. However, measuring teacher quality is somehow

---

education systems and policies since the end of World War II. For that subset of countries we also used country-specific descriptions of national education systems, thematic studies on specific institutional features and we double-checked our data by directly contacting national experts in the field. In the current version of the paper, due to the inclusion of a significant number of non-EU countries, we preferred to rely on a single data source. However, when correlating the reforms identified through the NATLEX database only with the ones previously identified using also Eurybase, we find a correlation of the reform indicators ranging between 0.92 and 0.96.

[15] In a paper adopting a similar strategy for building reform variables (see Braga et al., 2013) it was tested the exogeneity of the reform variables by instrumenting it with the political orientation of the governments undertaking the reform. In the present case it is rather difficult to attribute areas of reform to specific political orientation, and therefore we do not pursue it further.

problematic: the observable characteristics of teachers are weakly correlated with student achievement, and the reverse strategy of inferring teacher quality from observed student achievements is only valid when either the students are randomly allocated to teachers (inapplicable for countries where there is explicit or implicit streaming) or one possesses longitudinal samples where repeated observations of different student cohorts are exposed to the same teacher (as in Rivkin et al., 2005). This strategy is even more complicated when we consider that students are often exposed to more than one teacher (a sort of *group production*) and that teacher mobility is often driven by perceived student teachability (thus inducing a *self-sorting* of teachers to schools/classes). Thanks to the repeated cross-sectional structure of the data available in PIRLS, we exploit both cross-sectional and temporal variations to identify legislative changes that may be effective either because they attract or select better teachers or because they solicit a higher level of effort. In addition, we focus only on the effect of the main instructor in charge of teaching reading to fourth-grade students. Unfortunately, the survey does not collect information about teachers to whom students have been exposed to in previous grades, if different from the current one, making it impossible to distinguish between the relative contribution of each effect.

To this goal, our identifying strategy consists of comparing the average results of pupils exposed to teachers with the same characteristics who entered the profession in different years and were exposed to different labour market rules and features. We are able to implement this strategy by exploiting the four waves of PIRLS, making it possible to disentangle the effect of teachers' age, tenure and year of entry in the labour market. More in details, for each pupil i associated to class/teacher j in school s of country c surveyed at time t, we estimated a standard educational production function for the student's reading achievement a_i through the following equation:

$$a_{ijsct} = \alpha + \beta X_{it} + \gamma Y_{ijt} + \eta Z_{st} + \vartheta R_{c\tau t} + \delta_t + \delta_c + \delta_t \delta_c + \varepsilon_{ijsct} \qquad (1)$$

where the vector X_{it} associated to each student contains information about gender, age in years, language spoken at home, and, in some specifications, also parental education and home educational resources. The vector $Y_{ijt} = [X_{-ijt}, T_{j\tau t}]$, also associated to each student, includes class and teacher features and can be decomposed into two sub-vectors: the first one, X_{-ijt} controls for class and peer effects by considering the average features of the class computed with the exclusion of the considered pupil (like the share of females in the class, average age, share of students speaking a different language at home and – when available - an index for household educational resources and the average educational attainments among the parents in the class). The second sub-vector $T_{j\tau t}$ is class-invariant and contains information regarding the main or unique reading teacher of class j, who entered the labour market in year τ and was surveyed at time t: gender, age (in ten-year intervals), tenure (in years) and educational attainment (being graduated), in addition to the class size. The third vector Z_{st} accounts for school characteristics such as location (urban/rural), average teachers' tenure in the school, availability of a library and the share of disadvantaged students in the school. Last, but most important, we estimate the effects of teacher quality on pupils' by exploiting exogenous variations in the labour market setting relevant for teachers prevailing in the

year τ of their entry in the labour, as measured by the vector $R_{c\tau t}$ of the implemented reforms[16] in country c at time τ, for teachers surveyed at time t. In addition, to control for time invariant national differences in the educational systems, in the labour markets or in the institutional setting affecting teachers and their teaching practice, we include the country-fixed effects δ_c. Instead, the wave-fixed effects δ_t account for exogenous unobservable shocks affecting pupils performance that change over time but not across countries. Furthermore, we include the set of country×wave dummies $\delta_t\delta_c$ that control for possibly divergent trends across countries over the 15 years time-span considered due to non-observable country specific characteristics. Finally, the usual idiosyncratic error component, clustered at the country and year of hiring level, is ε_{ijsct}.

The match of the reforms to teachers according to their year of entry into the labour market allows identifying the effect of policies by comparing students' achievement in classes taught by 'treated' teachers against classes taught by 'non-treated' teachers acting as control cases. In fact, by matching the reforms to teachers based on their tenure, conditional on the year of survey, we can distinguish those who were affected by the reforms from those who were not. By so doing we compare the effects of having two teachers that, other things equal, were hired just before or just after a legislative change relevant for the teacher profession. For example, suppose a reform introducing, in a given country, the requirement of a university degree (BA level) to become teacher was approved in a country in 1990. As a consequence, candidates leaving teaching schools in the same year were forced to undertake three additional years of college to obtain the degree. Thus, all other things constant, we can test whether the students in classes with teachers hired before 1990 exhibited worse performance compared with those taught by teachers hired after 1990 (presumably with a BA degree – because of a lack of information, we are forced to assume perfect compliance). The identification assumption is that the timing of the reforming activity is uncorrelated with the error term once observables and unobservables are accounted for. In addition, in the present case, the effect of the reforms is more precisely identified because the age and tenure effects are distinct from the timing of the reforms, thanks to repeated survey effect. On the one side, we observe individuals in the same labour market with the same age and tenure but matched with different sets of teacher policies because they are surveyed in different periods. Similarly, we observe teachers with different age but identical tenure and year of entry in the labour market, as well as teachers of identical age who entered the profession in the same year, but reporting different seniority in surveys conducted in different years.

Our prior is that the two reforming areas potentially affecting student performance through teacher quality are the introduction of a more selective recruitment process and harder working conditions, though compensated by more generous compensation policies. A more selective or targeted recruitment process allows selecting well-qualified candidates who have specific skills and hence should translate into a more effective teaching practices improving learning. Symmetrically, reducing the generosity

[16] Since we are unable to claim that each reform perfectly identifies a single specific change, we include the four reform indicators together in order to reduce the risk of spurious correlation among reforms. However, results are robust to the inclusion of separate single reforms.

of the reward scheme *lato sensu* could improve the quality of the teaching workforce by attracting more motivated candidates (i.e. individual who are available to work in the profession despite the harder working conditions), reducing turnover, increasing retention and, hence, enhancing students' achievements. Instead, relaxing the workload in terms of teaching times or providing more flexibility could attract less motivated teachers, consequently reducing student achievements (adverse selection). Ex ante, it is not clear whether changes in the retirement possibilities could exert an effect on newly hired teachers. One the one hand, we could expect that early retirement or more generous allowances may lower teachers' motivation, possibly because of adverse selection. On the other hand, it could be also possible that when entering in the labour market individuals do not consider the retirement scheme that will become relevant later on only after a given (unpredictable) tenure.

In our estimates we assume that reforms in teacher policies (especially recruitment ones) mostly affect beginner teachers, leaving already tenured teachers almost unaffected (i.e., any imitative behaviour can be considered negligible). For the other reforming areas, through our identification strategy, we are capturing the effect on the attractiveness of the profession compared to others. Since we actually want to test whether the considered legislative changes may affect the quality of new entrants into the profession, in our setting the treated teachers are therefore the ones for which the reform may have affected the choice of entry into the profession and the non-treated teachers, the ones who were already in the profession at the time of the reform. However, the effect of the reform could vary throughout the teacher career. Presumably, reforms regarding working condition, pay and retirement rules could affect also already hired teachers but at a different degree of intensity, here being stronger the fresher the teacher is (i.e., the smaller the time period between entry into the profession and the reform is).

Furthermore, the lack of detailed information on each reform makes it impossible to construct a quantitative measure of the effect, allowing for a comparison of the magnitude of their impact across countries and over time. As such, our variables capture the frequency and intensity of the reforming activity of subsequent governments *vis-à-vis* teachers within each country.[17] What we are actually estimating is the effect of being taught by teachers hired under alternative institutional framework where the regulator has different preferences and hence gives different level of attention to the quality and the working conditions of primary school teachers.

We now turn to discussing the estimates of the model presented in equation (1) together with some variants to check its robustness. In particular, for the subset of OECD countries where information is available, we include the average remuneration of teachers to control for the effect exerted by the pay level. As previously mentioned, the primary teacher pay level is provided by the OECD only for a broader subset of 37 countries entering into our analysis.

[17] A similar strategy has been pursued by Braga et al. (2013) while studying the impact of educational policies on educational inequalities.

Table 7 reports the baseline model under different specifications and sample sizes. The first column presents the coefficients of interest for the full sample including the country-specific time trends. In order to consider the pay index, that is perfectly collinear with these trends, we have to remove them (Column 2) and reduce the sample size to countries for which the salary index is available (Column 3) and actually include the pay index (Column 4). Size and significance of the coefficients are comparable across specifications, with the exception of the coefficient relative to salary reforms, which is negative and significant only in Column 2, possibly because the salary reform variables captures specific trends for some countries: in facts it reverts sign and loses significance when reducing the set of countries in column 3. All the models are estimated controlling for student, class, teacher and school characteristics (the full set of coefficients for these baseline models can be found in Table A.4 in online appendix). In order to retain the maximum sample size, information on parental background is not included in this baseline specification.[18]

As expected, the selectivity of the recruitment process is positive and significantly correlated to the students' achievement in all specifications, suggesting that more selective recruiting polices could have a positive impact on the 'quality' of the teachers. Conversely, our estimates confirm the potential adverse selection hypothesis on teachers: teachers entering the labour market just after a relaxation of working conditions are associated, on average, with lower performance of their students. Results signal that such policies tend to attract individuals exerting less effort on their job because less-motivated or because having conciliation problems between work and family life (e.g., teachers who find this profession easier to combine with caregiving in the family, especially in countries with poor family leave policies). Finally, no effect is found when considering changes in the retirement schemes, indicating that individuals are predictably not forward looking when entering the labour market. Instead, the compensation level over the per capita GDP included in the model in Column (4) suggests that, independently of the reforms, the wage plays a role in attracting better quality teachers, leaving unaffected the role of the recruitment policies. Overall, the results confirm that selection at entry is as good as improving pay conditions when it comes to raising student performance. Unfortunately, as we have already discussed in the previous section, our reform variables are scale-free, making it impossible to assess the size of the existing trade-off between the two alternatives.

Some teacher characteristics are also associated to better student performance: female and graduate teachers generally obtain a positive and significant coefficient in most of the specifications, while tenure is always insignificant.[19] When moving to the other covariates (reported in table A4 in the appendix), regardless of the specification girls outperform their boy counterparts by an average of 13 points. The point estimates also show a small negative effect of age that probably captures the lower skills of students

[18] As discussed in Section 4, the missing information on parental background cuts the sample size by a sizeable amount. We cope with this issue in the next section.

[19] This is important from our perspective because we match teachers and reforms based on this variable. The absence of significance in this regression reduces the risk of a spurious correlation with the reform variables.

repeating the year. Moreover, students speaking a different language at home are at a disadvantage.[20]

A non clear-cut issue regards the clustering of the standard errors. Different strategies can be implemented in our case, according to different studies. A recent paper by Abadie et al. (2017) points to the fact that clustering should follow the sampling strategy, therefore – in our case – suggesting to cluster at class level, given the random selection of classes within randomly selected schools in regionally stratified samples, or at country level. Other studies (for instance Bertrand et al., 2004) would suggest clustering according to the level of variation of the treatment variable of interest, that is, in our case, country and year of hiring. We decided to follow the latter strategy and cluster the standard errors at country and year of hiring level. However, we replicate the baseline models clustering errors both at country and at class level and no significant differences emerge.[21]

6. Robustness Checks and Sensitivity Analysis

Starting from the baseline model, we consider a number of alternative specifications in order to assess the consistency and robustness of our results.

As mentioned in Section 4, we start by considering, that about one fifth of the sample is missing information on parental background, the reason why we do not include parental background in the main specification. However, since parental background is a key determinant of pupils' achievement, we run the same model considering also parental education and the index of home educational resources. In addition, omitting pupils with missing values in these variables in countries that collected this information introduces potential selection bias in our estimates. For this reason, we impute missing values by country and wave to include these observations in our estimates.[22] Table 8 reports the results of the model controlling for parental background, while Table 9 shows point estimates with imputation of parental background missing information, including the set of dummies variable to control for the imputation. All results are confirmed, with the addition of the retirement reforms, that turn out to be negatively and significantly correlated to pupils' attainment in some specifications. As previously discussed, this negative association could suggest an adverse selection similar to the effect of working conditions.

[20] It is important to note that the PIRLS survey directly tests linguistic competencies that are extremely correlated with the language usually spoken in everyday life. Interestingly, the same characteristics averaged by class play a similar the same role in determining pupil performance and strengthening the effect of the corresponding individual feature, indicating a significant peer effect, in particular for the share of students speaking a different language at home (Table A.4 in the appendix).

[21] The only exception is the coefficient on recruitment reforms on the restricted sample that loses significance, when clustering at country level. However, this does not weaken the main results of our study, since the recruitment coefficient always tends to be less significant in the subset of OECD countries.

[22] Since our specification relies on country×year fixed effects, we replace missing values with the mean of the non-missing observations by country and wave. In addition, we create dummy variables (taking value 1 for imputed values and 0 otherwise) to control for the imputation when estimating the model.

Second, there could be an issue of sample distortion. The country-wave panel is highly unbalanced, with only few countries participating to all the four waves, and this could bias the estimates, even after controlling for country-specific time trends. In order to check the robustness of the results, we run the baseline model with and without parental background including only those countries that are present in all waves (Table A.5, Col.1-2 indicated as "balanced sample") or excluding one wave at turn (Table A.5, Col.3-10). Again, results are fully confirmed for the working conditions reforms, while for some samples they weaken the correlation with the recruitment reform variable.

A third possibility we have considered is spurious correlation between the reforms and other trends in the labour market that may affect the selection into the teaching profession. To control for this, we run a placebo test by matching the year of entry in the profession with the reforms indicators in place 5, 10 and 15 years before and after the actual year of entry. Table 10 shows that the main results of the baseline model do not survive when placebo 'lagged' reforms are introduced (col.1-3), while when placebo 'forwarded' reforms are considered (col. 4-6) some correlation is still in place but point estimates are lower both in terms of magnitude and significance, suggesting some kind of dissipative effect over time. It should be noticed, however, that this is in line with our expectations. On the one side, our identification strategy relies on the fact that reforms affect only teachers hired *after* their implementation. However, one may think that some features of the reforms can be anticipated by the potential teachers, therefore showing their effects even before their implementation. Results in Columns 1 to 3 show that this is *not* the case, confirming our identification strategy. On the other side, reforms on the teachers' labour market are relatively rare events, therefore manifesting their effect over a relatively long time span. Indeed, Columns 4 to 6 present results that corroborate this interpretation, with effects decreasing over time. On top of that, our reform variables identify the year of adoption and entry into force of the law and hence we assume that from that date onwards the change would be fully adopted. Instead, compliance could be less than perfect because the ratification process takes times or because the implementing decrees must be completed.[23]

A further possible concern about our results is that countries differ both in terms of education system and in terms of labour markets by a great extent. Different education systems may push to different reforms in the institutional setting for teachers, while different labour market conditions may modify the outside options for a potential teacher. To account for these factors, we follow two different strategies. On the one side, we control for GDP growth and employment rate in the year of hiring (only for a subset of years/countries, due to the lack of data from the Penn World Tables before 1950). In both cases, we do not find evidence of distortions in our main estimates attributable to these effects (see Tables A.6 and A.7 in the appendix): point estimates are unchanged in terms of magnitude and level of statistical significance. On the other hand, we run a model with year-of-hiring fixed effects that, together with country fixed

[23] Compared to the baseline model presented in Table 7, in this specification we lose some observations because the matching with the placebo vectors of reforms cannot be performed for the first/last cohorts of teachers.

effects and year fixed effects, should capture any unobservable macroeconomic shock and any unobservable specific cohort effect.[24]

Furthermore, since we are unable to claim that each reform perfectly identifies a single specific change in the institutional setting, our preferred specification considers the four reform indicators together in order to reduce the risk of spurious correlation in the estimates. However, results are robust to the inclusion of each single reform indicator separately. The size and the significance level in the two specifications are fully comparable, both considering or excluding parental background. In addition, to take into account the fact that a teacher could be hired when more dimensions of the working setting were reformed at the same time, we control for the number of reforms implemented in the hiring year or for the implementation of a full package of reforms. All our previous results are confirmed, and these additional controls do not exert any statistical effect on the outcome variable.[25]

Finally, to check whether the results are driven by the behaviour of a single country, we re-estimate our baseline regressions excluding from the sample one country at a time or considering each country separately and by specific geographic areas (Europe vs other). Although not reported due to space limitations, the results continue to hold and are not driven by the behaviour of a single country.[26,27]

In all previous specifications, we have focused on legislative changes that occurred just before each teacher entered the labour market since these reforms could affect the quality of the applicants and, hence, the subsequent performance of their students. However, having been exposed to reforms also throughout the career could influence the incentives to be effective in teaching. Therefore, in Table A.8, we study whether the intensity of the reform process throughout one's career has an effect on teacher quality and translates into different levels of student achievement. In particular, for each of the four reforming areas, we identify the number of legislative changes affecting a given teacher *after* his or her entry into the school system up to the date of the survey when the students' competencies are tested. Results on recruiting process and working conditions are confirmed. In addition, there seems to be some positive effect of salary reforms implemented during the career, suggesting that might be, on average, some incentive effect on teachers already in the profession. All the effects disappear when looking at the restricted sample. In this case, only a weak effect of the reforms of the retirement requirements is present, likely due to the fact that considered countries have an older labour force.

[24] The results are available upon request.
[25] Due to space constrains, all the results commented in this paragraph are available upon request.
[26] We also distinguish between formerly planned economies, North America, Latin America, East Asia, the Middle East and North Africa, Oceania and sub-Saharan Africa and results still hold.
[27] The results are available upon request.

7. Discussion and Conclusions

The current paper provides new evidence on the effect of teacher quality on student performance in primary school. Based on international standardised tests for literacy conducted with fourth-grade students and using variations in the institutional setting relevant for teachers when entering the profession, our analysis shows that teacher quality matters. In particular, we identify two channels to improve learning achievements and to ensure high-quality standards in compulsory school. On the one hand, teachers hired in periods when local governments were active on recruiting policies are associated to better performance of their pupils. On the other hand, if hired when government were easing their working conditions, including retirement policies to some extent, they are associated to lower performance. While reforms to the pay systems are not correlated with pupils' achievement, for a subset of countries we are also able to control for a measure of actual compensation, finding positive effects. Our policy conclusions emphasize the role of selective policies as a tool for improving the quality of the educational systems. However, to improve student performance, policy makers must adequately balance the selectivity level of the recruitment process of potential applicants with the generosity of the reward scheme in terms of flexibility and working conditions. In addition, also the costs of such policies should be considered, both directly and indirectly. Although a cost-benefit analysis is beyond the scope of this paper, our results suggest that the two reforming areas associated to changes in pupils' performance would potentially imply interventions that are cheaper than those related to compensation or retirement that, on average, turn out not being quite effective.

It is important to recall that recruitment policies may be directly effective in selecting better quality teachers, identified either ex-ante (via examinations) or ex-post (via probation and/or mentoring), while policies affecting the working and retirement conditions work indirectly, by making more or less attractive the teaching profession to the aspiring candidates. A similar role can be played by the compensation policies, especially when considered relatively to other job opportunities. Aggregate data do not allow us disentangling these effects from the incentive effects than can be exerted by pay-to-perform schemes, but as long as they lead to more generous pay policies, they are absorbed by the positively correlated pay index we find in our estimates. Thus, our overall conclusion points to the fact that policies can effectively enhance school quality, especially when targeted at primary school teachers, because these reforms are also effective in enhancing the overall quality of the educational system thanks to their cumulative effect on subsequent school grades.

Obviously, our previous estimates do not capture all aspects of a country setting that might crucially affect teachers' incentives. For example, the quality of infrastructures and equipment may affect the desirability of entering the profession. But also job amenities, financial and non-financial, influence the *type* of who enters and remains in the teaching profession. These dimensions, whether proxied or not by our contextual controls, could be responsible for heterogeneity of the reforming activities. In addition, the data at hand do not allow considering all dimensions related to teaching techniques, which other papers have found important in affecting student learning. However, our analysis has shed new light on the relevance of personnel policies in school

management and could help policy makers as well academics in increasing their understanding of the effect of intended teacher policies. Identifying the most effective way to recruit and motivate the best teachers is a non-trivial question for policy makers, also considering that the costs of teaching staff represent almost all total schooling expenses, in almost all countries.

References

Abadie, Alberto, Susan Athey, Guido W. Imbens and Jeffrey Wooldridge. 2017. When Should You Adjust Standard Errors for Clustering? NBER Working Papers 24003,

Andersson, C., Johansson, P. and Waldenström, N. (2011). Do you want your child to have a certified teacher? *Economics of Education Review*, 30(1): 65-78.

Atkinson, A., Burgess, S., Croxson, B., Gregg, P., Propper, C., Slater, H. and Wilson, D. (2009) Evaluating the impact of performance-related pay for teachers in England. *Labour Economics*, 16 (3): 251-261.

Bacolod, M.P. (2007). Do Alternative Opportunities Matter? The Role of Female Labor Markets in the Decline of Teacher Quality. *The Review of Economics and Statistics*, 89(4): 737-751.

Ballou, D. (2001). Pay for performance in public and private schools. *Economics of Education Review*, 20: 51–61.

Bertrand, Marianne, Esther Duflo and Sendhil Mullainathan. 2004. How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics*. 249-275

Braga, M., Checchi D and Meschi E. 2013. Institutional Reforms and Educational Attainment in Europe: A long run perspective, *Economic Policy* 73/2013: 45-100.

Britton, Jack and Carol Propper. 2016. Teacher pay and school productivity: Exploiting wage regulation. *Journal of Public Economics* 133: 75–89

Buddin, R. and Zamarro, G. (2009). Teacher qualifications and student achievement in urban elementary schools. *Journal of Urban Economics*, 66(2): 103–115.

Cameron, Charles, John M. de Figueiredo and David E. Lewis. 2016. Public sector personnel economics: wages, promotions, and the competence-control trade-off. NBER Working Paper 22966

Carroll, David, Jaai Parasnis and Massimiliano Tani. 2018. Teaching, Gender and Labour Market Incentives. IZA DP 12027.

Dohmen, Thomas and Armin Falk. 2010. You get what you pay for: incentives and selection in the education system. *Economic Journal* 120: F256-271

Dolton, P. and Marcenaro-Gutierrez, O.D. (2011). If You Pay Peanuts do You Get Monkeys? A Cross Country Comparison of Teacher Pay and Pupil Performance. *Economic Policy*, 1: 5–55.

Falch, T., Johansen, K. and Strøm, B. (2009). Teacher shortages and the business cycle. *Labour Economics*, 16(6): 648-658.

Garibaldi, Pietro. 2006. *Personnel economics and imperfect labour markets*. Oxford University Press 2006

Garrouste, C. (2010): *100 years of educational reforms in Europe: a contextual database.* EUR – Scientific and Technical Research series, Vol. 24487, (2010): pp. 1-349.

Hanushek, E.A. and Rivkin, S.G. (2006). Teacher Quality, in E.A. Hanushek and F. Welch (eds.), *Handbook of the Economics of Education Volume 2*. Amsterdam: Elsevier, Chapter 18, pp. 1051–1078.

Hanushek, Eric and Ludger Woessman. 2011. The Economics of International Differences in Educational Achievement. in Eric A. Hanushek, Stephen Machin, and Ludger Woessmann, editor: *Economics of Education*, Vol. 3, The Netherlands: North-Holland, pp. 89-200.

Hanushek, Eric, Marc Piopiunik and SimonWiederhold. 2019. The Value of Smarter Teachers: International Evidence on Teacher Cognitive Skills and Student Performance. *Journal of Human Resources* 54: 857-899.

Harris, D.N. and Sass, T.R. (2011). Teacher training, teacher quality and student achievement. *Journal of Public Economics*, 95: 798–812.

Kane, T.J., Rockoff, J.E. and Staiger, D.O. (2008). What Does Certification Tell us About Teacher Effectiveness? Evidence from New York City. *Economics of Education Review*, 27(6): 615-631.

Lavy, V. 2015. Teachers' pay for performance in the long-run: effects on students' educational and labor market outcomes in adulthood. NBER wp n.20983

Lazear, E. and M. Gibbs. 2007. *Personnel Economics for Managers* (2nd edition). Wiley

Moor Johnson, S. (2006). The Workplace Matters: Teacher Quality, Retention, and Effectiveness. Best Practices-NEA Research Working Paper (July 2006).

Nagler, Markus, Marc Piopiunik and Martin R. West. 2018. Weak markets, strong teachers: recession at career start and teacher effectiveness. forthcoming in *Journal of Labor Economics*

OECD 2017. *PISA in focus #70: what do we know about teachers' selection and professional development in high performing countries ?* Paris

OECD 2018a. *Effective teacher policies. Evidence from PISA*. Paris

OECD 2018b, *Education at a Glance* 2018: OECD Indicators, OECD Publishing, Paris

Rivkin, S., E.Hanushek and J.Kain. (2005). Teachers, Schools, and Academic Achievement, *Econometrica*, vol. 73(2): 417-458

Santibañez, L. (2006). Why We Should Care if Teachers Get A's: Teacher Test Scores and Student Achievement in Mexico. *Economics of Education Review* 25 (15): 510-520.

Schwerdt, Guido and Amelie C. Wuppermann. 2011. Is traditional teaching really all that bad? A within-student between-subject approach. *Economics of Education Review* 30: 365–379

Woessmann, Ludger 2011. Cross-country evidence on teacher performance pay. *Economics of Education Review* 30(3): 404-418

Woessmann, Ludger. 2016. The Importance of School Systems: Evidence from International Differences in Student Achievement. *Journal of Economic Perspectives* 30(3): 3-32