I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 12307

Does Evaluating Teachers Make a Difference?

Simon Briole
Éric Maurin

APRIL 2019

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 12307

Does Evaluating Teachers Make a Difference?

Simon Briole
Paris School of Economics

Éric Maurin
Paris School of Economics and IZA

APRIL 2019

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Does Evaluating Teachers Make a Difference?*

In France, secondary school teachers are evaluated every six or seven years by senior experts of the Ministry of education. These external evaluations mostly involve the supervision of one class session and a debriefing interview, but have nonetheless a direct impact on teachers' career advancement. In this paper, we show that these evaluations contribute to improving students' performance, especially in math. This effect is seen not only for students taught by teachers the year of their evaluations but also for students taught by the same teachers the subsequent years, suggesting that evaluations improve teachers' core pedagogical skills. These positive effects persist over time and are particularly salient in education priority schools, in contexts where teaching is often very challenging. Overall, a system of light touch evaluations appears to be much more cost-effective than more popular alternatives, such as class size reduction.

JEL Classification: I20, I28, J24

Keywords: teacher quality, evaluation, feedback, teaching practices, supervision, education

Corresponding author:
Éric Maurin
Paris School of Economics
48 Bd Jourdan
75014 Paris
France

E-mail: eric.maurin@ens.fr

* We would like to thank Marc Gurgand, Sandra McNally and Elise Huillery for their helpful comments on previous versions of this paper. We would also like to thank the French Ministry of Education for providing us with the administrative data exploited in this paper.

Introduction

There is a large body of research suggesting that teachers vary a lot in their ability to improve students' performance (Hanushek & Rivkin (2010)). It is also generally admitted that teacher evaluation can be a way to improve teachers' effectiveness, either by making it possible to provide them with useful feedbacks or by creating incentives to implement better practices (Isoré (2009); Taylor & Tyler (2012)). However, despite the recent evidence that existing evaluation systems produce accurate measures of teacher productivity (Jacob & Lefgren (2008); Kane et al. (2011); Bacher-Hicks et al. (2017)) there is still very little evidence on the actual impact of teacher evaluation on student performance. Teacher evaluations take many different forms across the world and vary a lot in terms of resources involved per teacher, but there is no consensus on what a good evaluation system should be and on how intensive it should be (Isoré (2009); OECD (2013a,b); Jackson et al. (2014)).

To shed light on this issue, this paper builds on administrative data with exhaustive information on the exact timing of secondary school teachers' evaluations in France, in a context where evaluations take place every six or seven years, involve very little resources per teacher and year, but have nonetheless a direct impact on teacher career advancement.

Evaluations are conducted by senior experts of the ministry of education, called *inspecteurs d'académie - inspecteurs pédagogiques régionaux* (hereafter *inspecteurs*), but each one of these *inspecteurs* is responsible for more than 350 teachers and has to perform on average about 40 evaluations per year, on top of many other managerial activities within the education system (IGEN (2011); IGEN/IGAENR (2016)). Evaluations mostly encompass the supervision of one class session and a debriefing interview with the teacher and we can estimate the cost to be about 600 euros per evaluation, namely about 100 euros per year and teacher. The results of these evaluations are used, however, to determine teachers' progression in the wage scale[1]. As a consequence, evaluations may not only help teachers improve their skills through the provision of evaluators' feedbacks, but they also give teachers strong incentives to provide effort to improve their teaching practices in order to be as good as possible on the day of evaluation.

Each year, each teacher is assigned to a given set of classes and, consequently, teaches the same group of students over the whole year. Our empirical strategy exploits data on the exact timing of evaluations to compare the average performance of students assigned to a teacher *before* and *after* his/her

[1] In most developed countries, teachers' evaluations are either conducted by internal evaluators only or not related to career advancement (OECD (2013a)). Only in a few countries (including Portugal, Switzerland and some regions of Germany) are teachers' evaluations conducted by external evaluators and have a direct impact on teachers' wage and promotion, as in the French system (OECD (2013b) ; Eurydice (2018)). Another important feature of the French system is that evaluations are conducted each year, in each subject, by the same group of highly qualified civil servants (*inspecteurs*) who likely develop a specific expertise in this task.

evaluations, the basic question being whether evaluations coincide with specific improvement in students' average performance. Identification relies on the assumption that external evaluations do not coincide with teachers being assigned to better classes. Empirically, we checked that there is no specific change in students' characteristics before and after evaluations and, in particular, no changes in the proportion of students who have been held back a grade or in the proportion who take prestigious non-compulsory courses (such as Latin or ancient Greek courses). We also checked that external evaluations are not followed by specific changes in the level of teaching experience of colleagues who teach other subjects in the same class. If teachers were systematically assigned to better classes after external evaluations, we would observe a different pattern, namely a mechanical increase in colleagues' level of experience after external evaluations. Eventually, we provide evidence that the timing of teacher evaluations is unrelated to teacher mobility and that, more specifically, teachers don't move to better performing schools after an external evaluation. By contrast, as regards performance, we provide clear evidence that the visit of a math teacher by an external evaluator is followed by a significant increase (of about 4.5% of a SD) in students' scores in math at the end of middle school (9th grade). The effect of math teachers' evaluations is observed on performance in math, not in other subjects, consistent with the assumption that increased performance in math are driven by improved teaching practices of math teachers, not by an increase in students' overall academic ability or in math workload (which would likely be detrimental to performance in other subjects). Furthermore, math teachers' increased effectiveness is observed not only at the end of the evaluation year, but also at the end of the following years. Such persistent effects on teachers' effectiveness are consistent with the assumption that the visit of an evaluator is associated with an improvement in teachers' pedagogical skills, not just a temporary increase in teachers' effort. In the same spirit, the influence of math teachers' evaluations on their students can still be seen several years later, in high-school, as a larger proportion of their former students keep on studying math and succeed in graduating in fields of study which involve taking math exams. These longer term effects on students' outcomes are further suggestive that external evaluations do not simply help math teachers to "teach to the test", but make them able to improve students' core skills as well as students' perception of the discipline. These improvements can be seen for less experienced teachers as well as for more experienced ones. They are even more significant for math teachers assigned to education priority schools, in context where students' academic level is often very weak and teaching more challenging.

Building on the same identification strategy, we show that external evaluations have smaller effects on French language teachers than on math teachers, even though evaluations of French language teachers are followed by significant improvement in French language test scores in education priority schools.

The smaller effects of external evaluations on French language teachers are consistent with the existing literature on teacher effectiveness, which typically finds that teacher effects are much weaker on language exams than on math exams, maybe because students learn language in many settings outside schools, so that the influence of teachers is diluted and distorted by that of many other factors (Lavy (2009); Hanushek & Rivkin (2010); Harris & Sass (2011); Taylor & Tyler (2012); Wiswall (2013); Jackson et al. (2014); Papay & Kraft (2015)).

Eventually, when we consider the joint sample of math and French language teachers, we find an average effect of teacher evaluation of about 3% of a SD on test scores. Such an average effect is about the same order of magnitude as the average effect of a 5-student reduction in class size, as estimated by Piketty & Valdenaire (2006) for French middle schools. Our program of teacher evaluations involves, however, much smaller cost per teacher and year.

Our paper contributes to the growing literature on the causal impact of policies aimed at improving teachers' effectiveness. These policies include program of peer mentoring for new teachers (Rockoff (2008); Glazerman et al. (2008, 2010)) as well as programs of formal training and professional development (Angrist & Lavy (2001); Harris & Sass (2011)) and policies designed to evaluate and provide feedbacks to teachers (Weisberg et al. (2009); Allen et al. (2011); Taylor & Tyler (2012); Murphy et al. (2018)). Generally speaking, most existing papers focus on US programs and are suggestive that teacher-related programs can make a difference only insofar as they are high intensity. For example, the evaluation program in Cincinnati public schools involve the observation of four classroom sessions during the year of the evaluation, three by an external expert and one by an internal one (Taylor & Tyler (2012)). Both external and internal evaluators have to complete an intensive evaluation training program, so as to be able to measure several dozens of specific skills and practices. Overall, the Cincinnati program has a significant effect on math teacher effectiveness (about +10% of a SD on student' scores), but involves a total budget of about 7,500 dollars per evaluation, namely a cost per evaluation that we estimate to be about 10 times more important than the budget involved by the program analyzed in our paper.

The remainder of the paper is organized as follows. Section 1 describes the teacher evaluation system as well as the organization of secondary schooling and national exams in France. Section 2 presents the databases exploited in this paper and the construction of our working samples. Section 3 develops our empirical approach and shows the effects of external evaluations on student outcomes through a graphical analysis. Section 4 implements a regression analysis to show the robustness of our main results and to explore the potential heterogeneity in the effects of evaluations. The final section concludes with a brief discussion on the implications of our results.

1 Institutional context

In France, secondary school teachers are recruited through national competitive exams organized each year, in each field of study, by the ministry of education[2]. Once recruited, teachers' progression through the wage scale depends not only on internal evaluations conducted each year by school heads, but also on external evaluations conducted every 6 or 7 years by senior experts of the ministry of education[3]. Internal evaluations focus on teachers' behavior at school (punctuality, absenteeism, participation in cross-class collaboration projects) whereas external evaluations focus on teaching practices and pedagogical skills.

Teacher external evaluations

Teacher external evaluations are under the responsibility of a group of senior civil servants of the ministry of education, called *inspecteurs d'académie - inspecteurs pédagogiques régionaux* (hereafter *inspecteurs*). The vast majority of evaluations are conducted by *inspecteurs* themselves. A small fraction is conducted by senior teachers temporarily appointed to help *inspecteurs*[4].

Inspecteurs are recruited through national competitive exams restricted to experienced civil servants. There is one such competitive examination per field of study each year. Most candidates are experienced teachers who look for a career change. According to the staff directory of the ministry of education, *inspecteurs* are on average about 52 years old and have about 6 years of experience as *inspecteur* (see Table A1 in the online appendix). Once recruited, each *inspecteur* is assigned to a specific education region by a centralized assignment system. There are 31 education regions in France and the average number of *inspecteurs* per region and field of study is typically very small compared to the number of teachers. For instance, according to the staff directory of the ministry, there are on average only about 5 math *inspecteurs* per region and they have to evaluate about 1,700 math teachers (Table A1)[5]. According to the same data source, about 250 math teachers are evaluated each year, in each region. Assuming that

[2]The vast majority (93%) are granted the basic degree required to teach secondary school students, namely the *Certificat d'Aptitude au Professorat de l'Enseignement Secondaire* (hereafter CAPES). A small minority (about 7%) are recruited through an even more selective examination and hold an advanced degree, called the *Agrégation*. Most *Agrégation* recipients teach in high school or in higher education. In the remainder, given our focus on students' performance at end-of-middle school exams, we will focus on CAPES recipients.

[3]Teachers' basic promotion rate on the wage scale is based on their number of years of experience. But teachers who get good evaluations can be promoted at a faster rate. Going from the first to the last level of the wage scale takes about 30 years with the basic promotion rate versus only 20 years for the 30% teachers with the best evaluations. Teachers' access to the faster promotion track is determined by the weighted sum of the administrative grade that they get from school heads (/40) and the pedagogical grade that they get from external evaluators (/60).

[4]According to IGEN (2011), the proportion of external evaluations who are not conducted by *inspecteurs* vary across regions, but is never above 15%. Senior teachers appointed each year to help *inspecteurs* typically belong to the category who intend to take the exam to become *inspecteurs*.

[5]Overall, there were 142 math *inspecteurs* and 165 French language *inspecteurs* in France in 2008.

85% of these evaluations are conducted by *inspecteurs*, it means that each *inspecteur* conducts on average about 40 evaluations per year.

Each evaluation involves the supervision of one class session. It also involves a debriefing interview with the evaluated teacher, during which the *inspecteur* provides feedbacks and advices. *Inspecteurs* can also provide teachers with suggestions about the specific training sessions that they could attend to improve their teaching practices or class management practices. On the day of the evaluation, *inspecteurs* also examine students' notebooks as well as the class book, namely the book where teachers have to report class sessions' contents, the exams that they give, etc. Eventually, *inspecteurs* have to produce a written report (so called, *rapport d'inspection*) where they provide an analysis of the class session that they supervised and provide explanations for the overall grade that they give to the evaluated teacher. In general, teachers are notified well in advance of the visit of the *inspecteur*, if only because the date of the visit has to coincide with a day when they teach. However, there is no legal constraint on notification delays.

Symbolically, the evaluation of teachers represents the most important task assigned to *inspecteurs*. But, in practice, *inspecteurs* are in charge of many other aspects of the education policy, so that the evaluation of teachers represents only a small part of their activities. As a matter of fact, *inspecteurs* are also in charge of the conception of the many national exams organized each year in France[6]. In each education region, *inspecteurs* also have to contribute to the conception and organization of teacher training and professional development programs. As regards human resources management, they are also expected to play a consulting role with teachers, namely they are expected to answer queries about both career advancement and teaching practices. More generally, *inspecteurs* are expected to supervise the actual enforcement of education policies in each education region and each school. Overall, according to surveys conducted by the ministry of education on the working condition of *inspecteurs*, the evaluation of teachers represents on average only between 20% and 30% of *inspecteurs*' activities (IGEN (2011); IGEN/IGAENR (2016)). Given that the total wage cost of an *inspecteur* is about 100,000 euros per year and assuming that about 20-30% of this cost compensates for evaluation tasks, we can estimate that 20,000-30,000 euros compensate for about 40 evaluations, meaning about 500-700 euros per evaluation[7]. Given that there is only one evaluation every six or seven year, the cost per teacher and year is about 100 euros.

[6]Most notably, they are in charge of the different types of end-of high school *Baccalauréat*, as well as the different types of end-of-middle school *Brevet*, the different *Certificat d'Aptitudes Professionnelles*, etc.

[7]More information on the duties and compensations of *inspecteurs* can be found at the following address: http://www.education.gouv.fr/cid1138/inspecteur-de-l-education-nationale.html.

School context and exams

In France, middle school runs from 6th to 9th grade and high school runs from 10th to 12th grade. Students complete 9th grade the year they turn 15. The curriculum is defined by the central government. It is the same in all middle schools and there is no streaming by ability[8]. The 20% most underprivileged middle-schools benefit from education priority programs which provide them with additional resources[9].

An important feature of the French system is that students stay in the same class, in all subjects, (with the same teacher in each subject), throughout the school year. Classes are groups of about 25 students which represent, each year, very distinct entities. School principals assign students and teachers to classes before the beginning of the school year. In the remainder of this paper, we will mostly focus on teachers who teach 9th grade classes and our most basic measure of their effectiveness will be defined by the average performance of their students at the (externally set and marked) national exam taken at the end of 9th grade, which is also the end of middle school. This exam involves three written tests (in math, French language, history-geography) and our first question will be whether external evaluations of 9th grade teachers improve their ability to prepare their students for these tests. Specifically, we will mostly focus on math teachers and ask whether their external evaluations are followed by an improvement in the math scores of their students[10].

After 9th grade, students enter into high school, which runs from grade 10th to 12th grade. At the end of their first year of high school (10th grade), French students can either pursue general education or enter a technical or a vocational education program. Furthermore, those who pursue general education have to specialize in a specific field of study. There are three main fields: science (field "S"), economics and social sciences (field "ES") or languages and literature (field "L"). This is a key choice: each field of study corresponds to a specific curriculum, specific high school examinations, and specific opportunities after high school. Another important research question will be whether the effect of 9th grade teachers' evaluation on their students can still be seen one year later, at the end of 10th grade, on students' probability to choose S as field of specialization. The first year of high school (10th grade) is dedicated

[8] 9th grade students get about 25 hours of compulsory courses per week: 4 hours of French language, 3.5 hours of mathematics, 3.5 hours of History and Geography, 3 hours of Science, 1.5 hours of Technology, 5.5 hours of foreign languages, 3 hours of sport, 1 hour of art course. They also have the possibility to take additional (non compulsory) courses, such as Latin or ancient Greek. Principals can decide to assign students taking these additional courses to the same classes. Given that these students are typically good students, we may observe some segregation by ability across classes within schools.

[9] As shown in table A2 in online appendix A, the proportion of students from low-income families is twice bigger in education priority schools than in non-priority schools. Education priority schools also exhibit higher proportions of repeaters and students in this type of schools get lower scores at the end-of-middle school national examination on average.

[10] In the last section of this paper, we also present an analysis of the effects of external evaluations on French language teachers' effectiveness, as measured by their students' French language score. Generally speaking, we find much weaker effects on French language teachers than on math teachers, except in priority education schools.

to exploring the different subjects and to choosing a field of specialization. The two last years of high school (11th and 12th grade) are dedicated to the preparation of the national high school exit exam, the *Baccalauréat*, which is a prerequisite for entry into post-secondary education. Students have to take one exam per subject, and they obtain their diploma if their weighted average mark across subjects is 10/20 or more, where subjects taken and weights depend strongly on their field of specialization. Given our focus on math teachers, a last research question will be whether the effect of 9th grade teachers' evaluation on their students can still be seen three years later, at the end of 12th grade, on students' ability to graduate in science (S).

2 Data and samples

In this paper, we use administrative data with detailed information on secondary school teachers for the period between t_0=2008-2009 and t_1=2011-2012. For each teacher j, this dataset gives information on whether (and when) j underwent an external evaluation between t_0 and t_1. It also gives information on whether (and when) teacher j taught 9th grade students and on the average performance of these students at exams taken at the end of 9th grade as well as at exams taken subsequently at the end of high school. Online appendix B provides further information on how we build this database.

To construct our working sample of math teachers, we first extract from our main database the sample of math teachers who have less than 25 years of teaching experience, who taught 9th grade students in t_0, but who were not evaluated in t_0[11]. The size of this sample is about 40,000, which represents about 85% of the total number of 9th grade math teachers. About 57% of teachers in our sample are externally evaluated during the period under consideration and our objective is to evaluate the effect of these external evaluations on their students' math performance[12].

To explore this issue, we have to further focus on the subsample who teach 9th grade students at least one additional time after t_0, so as to be able to look at the evolution of students' performance at the end of 9th grade. The size of the corresponding working sample is about 30,000, which represents about 80% of the main sample. Most of our empirical analysis will be conducted on this working sample. One potential issue with this working sample, however, is that external evaluations may have an impact on

[11] We drop the small fraction of 9th grade teachers who are evaluated on year t_0=2008-2009 because the vast majority (about 96%) are not (re)evaluated before t_1 and cannot contribute to the identification of the effect of external evaluations. We also drop teachers with more than 25 years of teaching experience (on t_0) so as to minimize attrition rate. As it happens, many teachers with more than 25 years of experience are near the end of their working career and about 31% leave the education system between t_0 and t_1 (against only 4% for teachers with less than 25 years of experience). We checked, however, that results remain similar when we keep teachers with more than 25 years of teaching experience in our working sample (see online appendix C1 and C2).

[12] The sample of French language teachers used in the last section of the paper will be constructed in a similar way.

teachers' probability to teach 9th grade students after t_0, meaning the selection into the working sample may be endogenous to the "treatment" under consideration. To test for such an endogenous selection, we considered the main sample of 40,000 observations and we tested whether the probability to teach 9th grade students on a year t after t_0 is different for teachers who are evaluated between t_0 and t and for those who are not evaluated in this time interval. As shown in online Appendix Table A3, we find no significant difference between the two groups of teachers. The probability to teach 9th grade student on a given year after t_0 is on average about 78% for non-evaluated teachers and about 0.8 percentage point higher for evaluated teachers, the difference between the two groups being non-significant at standard level. The same diagnosis holds true when we replicate this sample selection analysis on subsamples defined by type of schools, teachers' experience or teachers' gender. Generally speaking, these results are consistent with the assumption that attrition is negligible.

Overall, our working sample includes 9,451 math teachers who teach 9th grade students at least two times between t_0 and t_1, which represents 30,414 observations in total. We provide some descriptive statistics in online Appendix A (see column (1) of Table A4)).

3 The effect of evaluations: conceptual framework and graphical evidence

In the remainder of the paper, we ask whether teachers' external evaluations are followed by an improvement in their effectiveness, as measured by their ability to prepare 9th grade students for national exams or for high school. We first focus on math teachers and the last section provides results for French language teachers. The underlying educational production function is straightforward: (a) students' achievement is assumed to depend not only on their individual characteristics, but also on the effectiveness of their teachers and (b) the effectiveness of teachers is assumed to depend not simply on their level of experience, but also on the number of external evaluations they underwent since the beginning of their career. In this framework, assuming that teachers are assigned to the same type of classes on the years before and after the visit of an *inspecteur*, the comparison of the effectiveness of evaluated and non-evaluated teachers before and after an additional evaluation provides a means to identify the impact of such an additional evaluation on effectiveness. Before moving on to our econometric investigations, we start by providing simple graphical evidence on this issue.

The impact of external evaluations: graphical evidence

For each group of evaluated math teachers defined by the year t_e of their evaluation (with $t_0 < t_e \leq t_1$), let us consider Y_{ed} the average performance in math of their 9th grade students at national exams taken at the end of year $t_e + d$ and Y_{-ed} the average performance of the students of non-evaluated teachers at the end of the same year $t_e + d$. Denoting Y_d and Y_{-d} the average of Y_{ed} and Y_{-ed} across all possible evaluation year t_e, Figure 1(a) shows the evolution of Y_d and Y_{-d} when d increases from d=-3 to d=+2 (i.e., the range of variation of d in our sample). The Figure reveals a marked increase in the average performance of students of evaluated teachers just after evaluations (i.e, for $d \geq 0$). The average performance of the evaluated and non-evaluated groups follows a similar pattern for exams taken before evaluations, but the gap widens for exams taken after evaluations.

To take one step further, Figure 1(b) plots the difference between evaluated and non-evaluated groups, with the last pre-evaluation year (i.e, t_e-1) being taken as a reference. It confirms that the evaluation year coincides with an improvement in the relative performance of evaluated teachers' students. The difference between the two groups of teachers is not statistically different from zero before the evaluation, but becomes statistically different from zero just after the evaluation.

Overall, Figures 1(a) and 1(b) are suggestive that evaluations have an impact on math teachers' effectiveness, as measured by the math scores of their 9th grade students. The basic identifying assumption is that evaluations do not coincide with teachers being assigned to better classes.

To further explore the credibility of our identifying assumption, Figures 2(a) and 2(b) replicate Figures 1(a) and 1(b) using average standardized scores in humanities as dependent variable, where scores in humanities are defined as the average of French language and history-geography scores[13]. Comfortingly, Figures 2(a) and 2(b) do not reveal any improvement in students' performance in humanities after external evaluations of math teachers. These Figures are in line with the assumption that external evaluations do not coincide with any overall improvement in the ability of students assigned to teachers. They are also consistent with the assumption that increased performance in math are driven by improved teaching practices of math teachers, not by an increase in math workload, since an increase in math workload would likely be detrimental to performance in other subjects.

A symmetrical falsification exercise consists in testing whether students math performances are affected by the evaluation of non-math teachers. Figures 3(a) and 3(b) shows that this is not the case, namely

[13] As mentioned above, students take three written tests at the end of 9th grade, namely a test in math, a test in French language and a test in history-geography. For each student, the score in humanities correspond to the average of the French language score and the history-geography score. Results are similar when we use separately the French language score and the history-geography score.

the Figures do not show any improvement in student math performance after the evaluation of French language teachers, which further suggest that teachers are not assigned to intrinsically better classes after external evaluations.

In online appendix A, Figures A1 (a) to A1 (c) provide additional evidence that external evaluations are not associated with teacher mobility (as captured by variation in teachers' seniority level) and do not coincide with teachers moving to better schools. In particular, these figures show that external evaluations do not coincide with any change in teachers' probability to teach in education priority schools. More generally, we do not see any variation in the academic level of the schools where they teach (as measured by the math average performance of 9th grade students at national exams taken in 2008, pre-treatment).

4 The effect of teachers' evaluations: regression analysis

The previous subsection provides us with simple graphical evidence on the effects of external evaluations on math teachers' effectiveness, as measured by the performance of their students at externally set and marked examinations. In this section, we explore the robustness of this finding - as well as the potential heterogeneity of effects across teachers and schools - using more parsimonious regression models. Specifically, we keep on focusing on the same working sample of math teachers as Figure 1(a) and we consider the following basic event-analysis model:

$$Y_{jt} = \beta T_{jt} + \theta X_{jt} + u_j + \gamma_t + \epsilon_{jt} \tag{1}$$

where Y_{jt} still represents the average standardized math score of teacher j's students at exams taken at the end of year t, while T_{jt} is a dummy indicating that an evaluation took place between t_0 and t. Variable X_{jt} represents a set of controls describing the average characteristics of the students taught by teacher j on year t (proportion of girls, average age, proportion studying ancient languages, etc.). X_{jt} also includes dummies controlling for teachers' number of year of teaching experience and for teachers' seniority level as well as a dummy indicating whether the teacher works in an education priority school and dummies indicating the education region. Eventually, the u_j and γ_t parameters represent a comprehensive set of teacher and year fixed effects while ϵ_{jt} represent unobserved determinants of students' performance.

In this set-up, parameter β can be interpreted as the effect of one additional external evaluation between t_0 and t on students' performance at the end of t. It should be emphasized that this basic parameter encompasses the effect of evaluations which took place on t (the very year of the exam) and the effect of evaluations which took place between t_0 and $t-1$. To separate these two effects, we will also

consider models with two basic independent variables, namely a dummy (denoted T_{1jt}) indicating that the evaluation took place on t and a dummy (T_{2jt}) indicating that the evaluation took place between t_0 and $t-1$.

To identify the parameters of interest in Equation (1), we assume that the timing of evaluations (as captured by changes in T_{jt}) is unrelated to changes in unobserved determinants of students' performance in math (as captured by changes in ϵ_{jt}), namely the same identifying assumption as in the previous graphical analysis. It amounts assuming that the evolution of the effectiveness of evaluated and non-evaluated teachers would have been the same across the period under consideration, had evaluated teachers not been evaluated. Table A5 in the online appendix shows the results of regressing students' observed characteristics (gender, age, family background as well as the study of ancient languages or the study of German language) on T_{jt} using model (1). Consistent with our identifying assumption, the Table shows that the timing of external evaluation does not coincide with any significant variation in students' characteristics. We also checked that when we regress T_{jt} on all student observed characteristics, a F-test does not reject the joint nullity of the estimated coefficients[14]. These results hold true regardless of whether we use the full sample of math teachers or subsamples defined by level of experience, gender or type of schools. Eventually, Table A6 in the online appendix confirms that the timing of evaluation does not coincide with teacher mobility (as captured by changes in teachers' seniority level) or with changes in the academic level of the schools where teachers work (as measured by school pre-treatment average scores or by priority education). The Table also reveals that the timing of evaluation does not coincide with changes in the level of experience or in the level of seniority of colleagues teaching other subjects to the same class. This finding is consistent with our assumption that evaluations are not followed by assignment to specific classes. If that were the case, evaluations would also mechanically coincide with assignment to classes with more senior and experienced colleagues.

4.1 Main effects on math scores

The first column of Table 1 shows the basic effect of external evaluations on math teachers' effectiveness, as measured by their students' performance in math at end-of-middle school national exams. Consistent with our graphical analysis, it confirms that external evaluations are followed by a significant improvement in math score of about 4.5% of a SD. The second column shows the impact of external eval-

[14]Specifically, we have F(5, 20857) = 0.49 ; p-value = 0.78

uations of math teachers on students' performance in humanities and, comfortingly, it shows no effect[15]. Column 3 shows the results of re-estimating the effect of math teachers' evaluations on math scores when we consider separately the effect on exams taken at the end of the evaluation year (T_{1jt}) and the effect on exams taken at the end of the following years (T_{2jt}). Both effects appear to be significant. The effect on exams taken at the end of the following years tend to be stronger (5.3% of a SD), but the difference between the two effects is non-significant at standard level. Eventually, column 4 confirms that math teachers' evaluations have no effect on performance in humanities, be they measured at the end of the evaluation year or later.

4.2 Heterogeneous effects

Table 2 shows the results of replicating our basic analysis separately on subsamples of math teachers defined by their gender, number of years of teaching experience (less than 11 years vs 11 years of more, where 11 is the median number of years of experience in our sample) or type of school (education priority schools vs regular schools). The Table shows that the impact of external evaluations on math scores is similar for men and women as well as for teachers with higher and lower level of work experience. By contrast, the impact appears to be significantly stronger for teachers in education priority schools (9.4% of a SD) than for teachers in non-priority schools (+3.1% of a SD). This finding is suggestive that external evaluations tend to be even more effective in school contexts where the average academic level of students is weaker and where teaching is more challenging[16].

Consistent with our identifying assumption, Table 2 also confirms that external evaluations of math teachers have no significant effect on students' performance in humanities, regardless of the subsample. As mentioned above, Tables A5 and A6 in the online appendix provide balancing tests for the different subsamples which further confirm that external evaluations are not followed by any systematic variations in class composition, teacher mobility or colleagues' characteristics.

4.3 Longer term effects

Previous sections suggest that external evaluations improve the effectiveness of math teachers, as measured by their ability to prepare their 9th grade students for exams taken at the end of 9th grade.

[15] As mentioned above, the score in humanities correspond to the average of the score in French language and the score in history-geography. We have checked that math teachers' evaluation have no effect on any of the two scores when we consider them separately.

[16] A survey conducted in 2006 provides an analysis of the specific challenges faced by teachers in education priority schools, due to students' social environment (poor working conditions at home, fatigue, diet) as well as to students' disruptive behaviors and low academic ability. The survey report emphasizes that most teachers lack the pedagogical skills that are necessary to adapt teaching to this specific context (IGEN/IGAENR (2006)).

Table 3 shows that the influence of math teachers on their 9th grade students can still be seen one year later at the end of 10th grade (when students have to choose their major field of study) or even three years later, at the end of 12th grade, when they have to take their high school exit exams. Specifically, the Table focuses on the same sample of 9th grade math teachers as Tables 1 or 2 and looks at the probability that their students subsequently choose science as major field of study as well as at the probability that they subsequently succeed in graduating in science. The first column of the Table shows an increase in both probabilities. Specifically, it suggests an increase of about 0.5 percentage points in the probability to choose science at the end of 10th grade and to graduate in science at the end of 12th grade, which represent an increase of about 3% in this probability. Consistent with Table 2, the following columns shows that this increase is particularly significant for teachers in education priority schools (+10%). These longer term effects on students' choices and performance are suggestive that external evaluations do not simply help teachers to "teach to the test", but make them able to improve students' core skills as well as students' perception of the discipline.

4.4 Effects of external evaluations on French language teachers

Until now, we have focused on math teachers. In this section, we extend our analysis to French language teachers. The corresponding working sample is constructed along the same line as the working sample of math teachers, meaning we focus on those who teach 9th grade students on t_0, who are not evaluated on t_0 and who have less than 25 years of teaching experience on t_0. Figures 4(a) and 4(b) replicate Figures 1(a) and 1(b) using this working sample of French language teachers. In contrast with what we find for math teachers, these Figures do not show any significant variation in performance at French language exams after French language teachers' evaluations. Tables A7 and A8 in online appendix A replicate Tables 1 and 2 using the sample of French Language teachers and confirm that external evaluations have only a small and marginally significant effect on their effectiveness, except when we focus on priority education schools (where the effect is about 7.6% of a SD). To further explore this issue, we looked at the effect of French language teachers' evaluations separately on reading test scores and writing test scores[17]. This analysis shows that the effects of external evaluations tend to be slightly stronger on writing test scores, but the difference across writing and reading tests is not significant at standard level (see Table A9 in online appendix A).

[17]The French language end-of-middle-school exam consists of a set of reading and a set of writing exercises. During the exam, students are given the same amount of time to complete each one of the two sets of exercises.

Generally speaking, the smaller effects of evaluations observed on French language teachers are in line with the literature on teachers' effects which recurrently finds that these effects are much weaker on language exams than on math exams (see e.g. Lavy (2009); Hanushek & Rivkin (2010); Harris & Sass (2011); Taylor & Tyler (2012); Wiswall (2013); Jackson et al. (2014); Papay & Kraft (2015)). One possible reason is that students learn language in many other settings outside schools, so that the influence of teachers is diluted and distorted by that of many other factors.

Eventually, Table A10 in the online Appendix shows the results of replicating our main regression analysis on the joint sample of math and French language teachers, so as to estimate the average effect of teacher evaluations on end-of-middle-school exams. The Table shows a significant effect of about 3% of SD (8% of a SD in priority education). Not surprisingly, this effect is close to the average of the effect for math teachers and the effect for French language teachers estimated in previous sections. Building on the same type of database as those used in this paper, Piketty and Valdenaire (2006) found that a 5-student reduction in class size improves 9th grade students' average score in math and French language by about 4% of SD. Hence, our estimated effect of teacher evaluation is about the same order of magnitude as the effect of a 5-student reduction in class size. The corresponding cost, however, is much smaller[18].

5 Conclusion

Despite the general consensus that teachers represent an important determinant of student achievement, there is still little evidence on successful policies aimed at improving teacher effectiveness. In this paper, we study the impact of teacher evaluation on students' performance, in a context where evaluations are conducted every six or seven years by senior experts of the Ministry of Education and represent a key determinant of teacher career advancement. We show that math teachers' evaluations increase their students' performance in math at end-of-middle school national exams. This effect is seen not only for students taught by the teacher the year of the evaluation but also for students taught by the same teacher the subsequent years, suggesting that evaluations improve teachers' core pedagogical skills. Math teachers' evaluations also generate persistent benefits for their students, who not only perform better at the end-of-middle school exam, but also graduate more often in science at the end of high school, three years later. The impact of evaluation appears to be much smaller for French language teacher, except

[18]Given that class size is about 25 students on average, a 5-student reduction corresponds to a class size reduction of about 20%. Hence, the corresponding cost per teacher and year can be estimated to be about 0.20 x 50,000 euros where 50,000 euros is a proxy for the total labor cost of a secondary school teacher. We end up with a cost per teacher and year of about 10,000 euros whereas the cost per teacher and year of the evaluation system is only about 100 euros (as discussed in section 2).

in education priority schools. For both math and French language teachers, the positive effects of evaluations are actually particularly salient in education priority schools, in contexts where teaching is often very challenging.

In terms of policy implications, our results suggest that a low-intensity low-cost evaluation program can be highly cost effective provided that it is conducted by external authorities and has a significant impact on teachers' career advancement. Our results also show that evaluations can generate significant benefits even after ten years of work experience. In most countries, evaluations tend to be concentrated on beginning teachers, whereas our findings suggest that it can be efficient to evaluate teachers all along their career, not simply at the start. Finally, our findings show that evaluations are particularly worthwhile in contexts where teaching is very challenging, such as education priority schools. Reinforcing teacher evaluations in this type of schools thus appears as an appealing way to reduce educational inequalities.

References

Allen, J. P., Pianta, R. C., Gregory, A., Mikami, A. Y., & Lun, J. 2011. "An interaction-based approach to enhancing secondary school instruction and student achievement". *Science*, 333(6045):1034–1037.

Angrist, J. D. & Lavy, V. 2001. "Does teacher training affect pupil learning? Evidence from matched comparisons in Jerusalem public schools". *Journal of Labor Economics*, 19(2):343–369.

Bacher-Hicks, A., Chin, M. J., Kane, T. J., & Staiger, D. O. 2017. "An evaluation of bias in three measures of teacher quality: Value-added, classroom observations, and student surveys". National Bureau of Economic Research.

Eurydice. 2018. *Teaching Careers in Europe: Access, Progression and Support. Eurydice Report.* Eurydice Report. Luxembourg: Publications Office of the European Union.

Glazerman, S., Dolfin, S., Bleeker, M., Johnson, A., Isenberg, E., Lugo-Gil, J., Grider, M., Britton, E., & Ali, M. 2008. "Impacts of Comprehensive Teacher Induction: Results from the First Year of a Randomized Controlled Study. NCEE 2009-4034.". *National Center for Education Evaluation and Regional Assistance.*

Glazerman, S., Isenberg, E., Dolfin, S., Bleeker, M., Johnson, A., Grider, M., & Jacobus, M. 2010. "Impacts of Comprehensive Teacher Induction: Final Results from a Randomized Controlled Study. NCEE 2010-4027.". *National Center for Education Evaluation and Regional Assistance.*

Hanushek, E. A. & Rivkin, S. G. 2010. "Generalizations about using value-added measures of teacher quality". *American Economic Review*, 100(2):267–71.

Harris, D. N. & Sass, T. R. 2011. "Teacher training, teacher quality and student achievement". *Journal of Public Economics*, 95(7-8):798–812.

IGEN. 2011. "Mission sur le rôle et l'activité des inspecteurs pédagogiques du second degré, Note à Monsieur le ministre de l'Education nationale, de la jeunesse et de la vie associative". Note n 2011-02.

IGEN/IGAENR. 2006. "La contribution de l'éducation prioritaire à l'égalité des chances des élèves". Rapport n 2006-076.

IGEN/IGAENR. 2016. "Rôle et positionnement des inspecteurs du second degré en académie". Rapport n 2016-070.

Isoré, M. 2009. "Teacher evaluation: Current practices in OECD countries and a literature review". OECD Education Working Papers, No. 23, OECD Publishing, Paris.

Jackson, C. K., Rockoff, J. E., & Staiger, D. O. 2014. "Teacher effects and teacher-related policies". *Annu. Rev. Econ.*, 6(1):801–825.

Jacob, B. & Lefgren, L. 2008. "Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education". *Journal of Labor Economics*, 26(1):101–136.

Kane, T. J., Taylor, E. S., Tyler, J. H., & Wooten, A. L. 2011. "Identifying effective classroom practices using student achievement data". *Journal of Human Resources*, 46(3):587–613.

Lavy, V. 2009. "Performance pay and teachers' effort, productivity, and grading ethics". *American Economic Review*, 99(5):1979–2011.

Murphy, R., Weinhardt, F., & Wyness, G. 2018. "Who Teaches the Teachers? A RCT of Peer-to-Peer Observation and Feedback in 181 Schools". CEP Discussion Paper No 1565.

OECD. 2009. *Education at a glance 2009: OECD indicators*. OECD Publishing.

OECD. 2013a. *Synergies for Better Learning: An International Perspective on Evaluation and Assessment*. OECD Reviews of Evaluation and Assessment in Education, Editions OCDE, Paris.

OECD. 2013b. *Teachers for the 21st Century: Using Evaluation to Improve Teaching*. OECD Publishing.

Papay, J. P. & Kraft, M. A. 2015. "Productivity returns to experience in the teacher labor market: Methodological challenges and new evidence on long-term career improvement". *Journal of Public Economics*, 130:105–119.

Piketty, T. & Valdenaire, M. 2006. *L'impact de la taille des classes sur la réussite scolaire dans les écoles, collèges et lycées français: estimations à partir du panel primaire 1997 et du panel secondaire 1995*. Direction de l'évaluation et de la prospective.

Rockoff, J. E. 2008. "Does mentoring reduce turnover and improve skills of new employees? Evidence from teachers in New York City". National Bureau of Economic Research.

Taylor, E. S. & Tyler, J. H. 2012. "The effect of evaluation on teacher performance". *American Economic Review*, 102(7):3628–51.

Weisberg, D., Sexton, S., Mulhern, J., Keeling, D., Schunck, J., Palcisco, A., & Morgan, K. 2009. *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. ERIC.

Wiswall, M. 2013. "The dynamics of teacher quality". *Journal of Public Economics*, 100:61–78.

Main Tables

Table 1: 9th grade math teacher evaluation and student performance

	End of middle school test scores			
	Math	Humanities	Math	Humanities
	(1)	(2)	(3)	(4)
Evaluation	0.045**	0.004		
	(0.014)	(0.014)		
Evaluation on t			0.041**	0.006
			(0.014)	(0.014)
Evaluation before t			0.053**	-0.003
			(0.018)	(0.018)
Observations	30414	30414	30414	30414

Note: The table refers to our working sample of math teachers who teach 9th grade students between t_0=2008-2009 and t_1=2011-2012. Column (1) (column (2)) shows the result of regressing their students' average standardized score in math (humanities) at the end of year t on a dummy indicating that they underwent an external evaluation between t_0 and t. Column (3) (column (4)) shows the result of regressing the same dependent variable on a dummy indicating that they underwent an external evaluation on t and on a dummy indicating that they underwent an evaluation between t_0 and $t-1$. Models include a full set of teachers and year fixed effects as well as controls for students' average age, gender, family social background, German language study and Ancient language study. Standard errors are in parentheses. * p<0.10, ** p<0.05.

Table 2: 9th grade math teacher evaluation and student performance - by subgroups

	(1) All	(2) Female	(3) Male	(4) Low-exp	(5) High-exp	(6) Priority	(7) Non Priority
Math score	0.045**	0.038*	0.052**	0.054**	0.039**	0.094**	0.031**
	(0.014)	(0.020)	(0.020)	(0.020)	(0.020)	(0.029)	(0.016)
Humanities score	0.004	-0.000	0.008	0.007	0.004	0.008	0.006
	(0.014)	(0.019)	(0.020)	(0.020)	(0.019)	(0.031)	(0.015)
Observations	30414	15724	14690	15072	15342	6818	23596

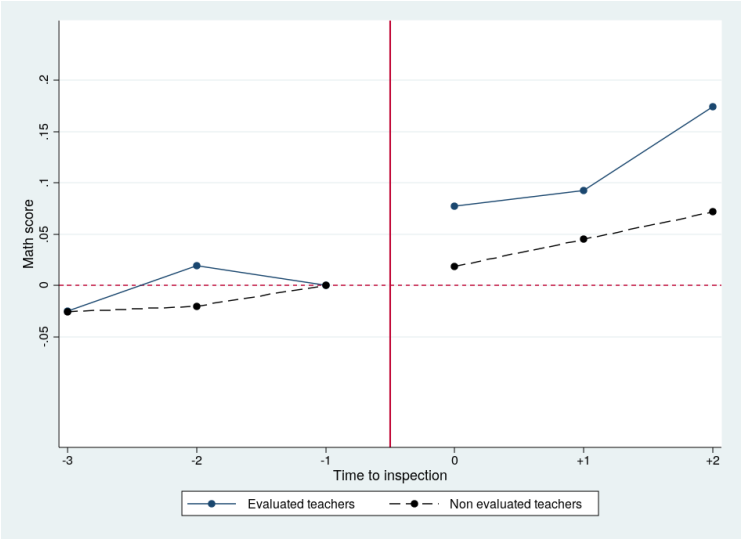
Note: The table refers to our working sample of math teachers who teach 9th grade students between t_0=2008-2009 and t_1=2011-2012. The first (second) row shows the results of regressing their students' average standardized score in math (humanities) at the end of year t on a dummy indicating that they underwent an external evaluation between t_0 and t. The first column refers to the full sample, whereas columns (2) and (3) refer to the subsamples of female and male teachers, columns (4) and (5) to the subsamples of teachers whose number of years of work experience is either above or below the median on t_0 (i.e., above or below 11 years), columns (6) and (7) to the subsample of teachers who were in education priority schools on t_0 and the subsample who were in non-priority schools. Models include a full set of teachers and year fixed effects as well as controls for students' average age, gender, family social background, German language study and Ancient language study. Standard errors are in parentheses. * p<0.10, ** p<0.05.

Table 3: 9th grade math teacher evaluation and student high school outcomes

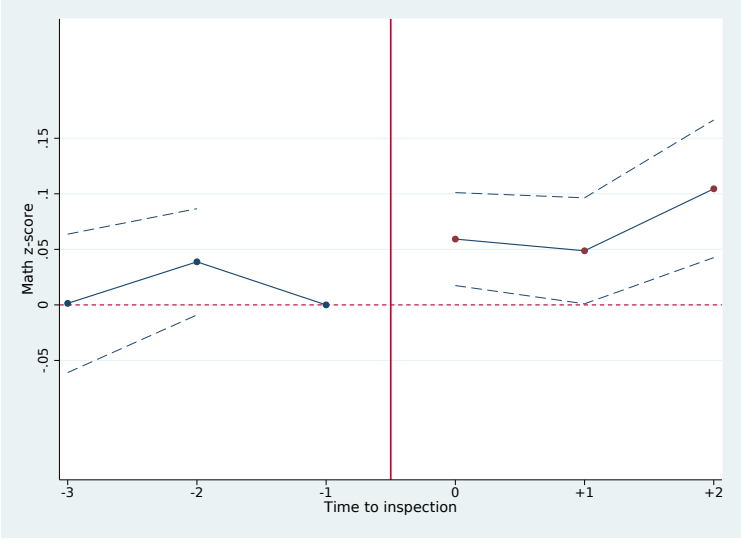
	(1) All	(2) Female	(3) Male	(4) Low-exp	(5) High-exp	(6) Priority	(7) Non Priority
Science as major field	0.005**	0.003	0.007**	0.008**	0.002	0.010**	0.003
	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)	(0.004)	(0.002)
	[0.176]	[0.183]	[0.169]	[0.161]	[0.191]	[0.123]	[0.192]
Graduation in Science	0.004**	0.002	0.007**	0.007**	0.003	0.009**	0.003
	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.002)
	[0.149]	[0.155]	[0.142]	[0.135]	[0.163]	[0.099]	[0.163]
Observations	30414	15724	14690	15072	15342	6818	23596

Note: The table refers to the working sample of math teachers who teach 9th grade students between t_0=2008-2009 and t_1=2011-2012. The first row shows the result of regressing the proportion of their 9th grade students who will choose science as major field of study at the end of 10th grade on a dummy indicating that they underwent an external evaluation between t_0 and t. The second row shows the result of regressing the proportion of their 9th grade students who will graduate in science at the end of 12th grade on the same independent variable. The first column refers to the full sample, whereas columns (2) to (7) refer to subsamples defined by teachers' gender, number of years of teaching experience on t_0 (above/below 11 years) and type of school attended on t_0 (priority/non priority). Models include a full set of teachers and year fixed effects as well as controls for students' average age, gender, family social background, German language study and Ancient language study. Standard errors are in parentheses. Sample means of the dependent variables are within square brackets. * p<0.10, ** p<0.05.

Main Graphs



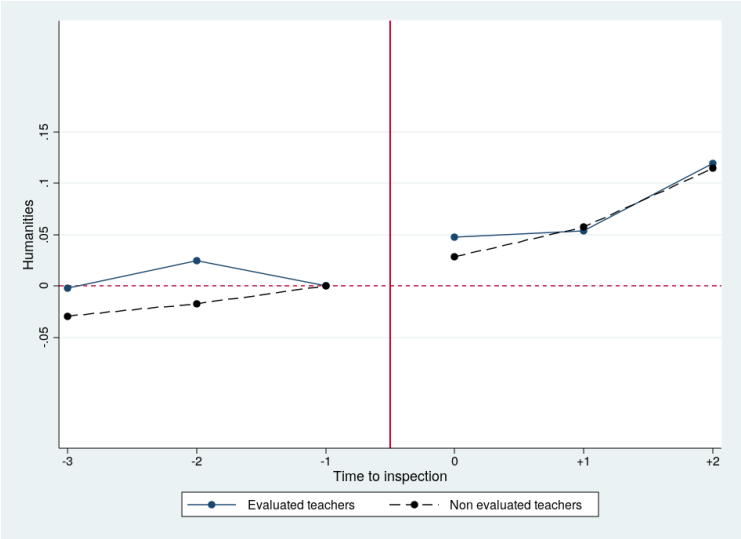
(a)



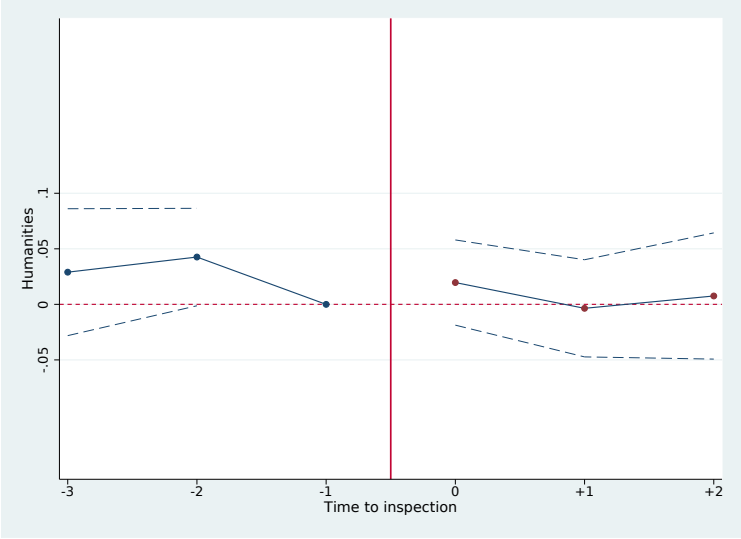
(b)

Figure 1: Math teacher evaluation and student performance in math

Note: The solid line in Figure 1 (a) shows math scores of students of evaluated math teachers before and after teachers' evaluations. The dotted line shows math scores of students of non-evaluated math teachers at exams taken on the same years. The solid line in Figure 1 (b) shows the difference in math scores between students of evaluated and non-evaluated math teachers before and after evaluations. The dotted lines show confidence intervals.



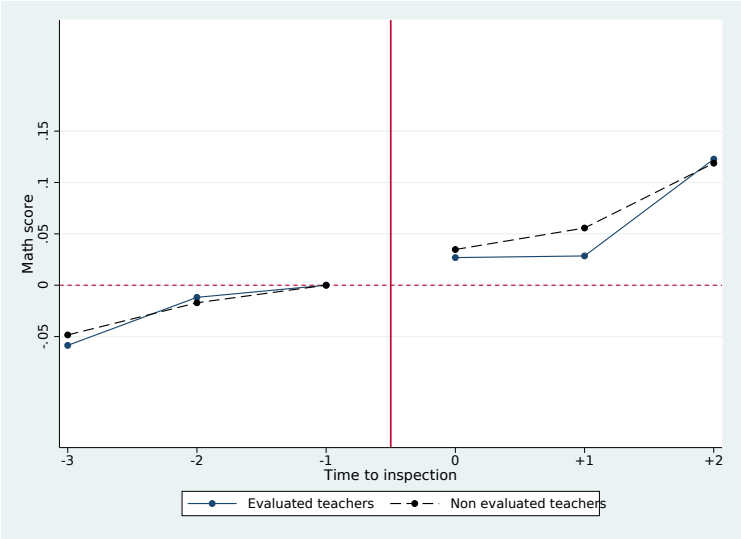
(a)



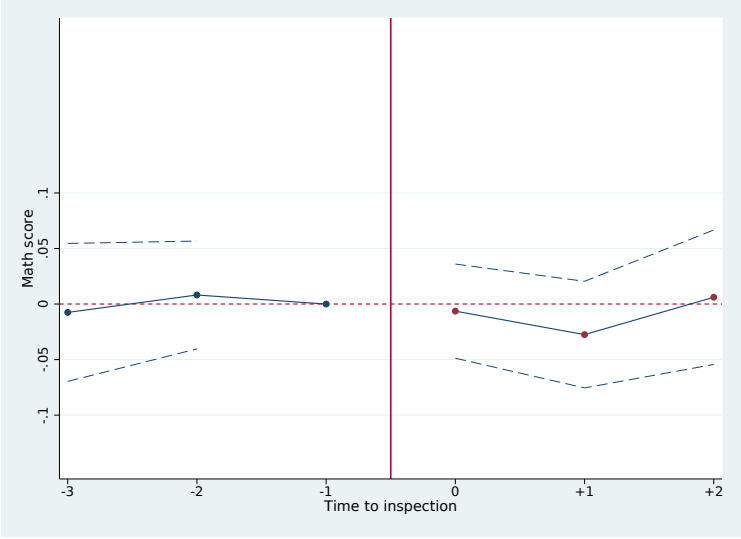
(b)

Figure 2: Math teacher evaluation and student performance in humanities

Note: The solid line in Figure 2 (a) shows humanities scores of students of evaluated math teachers before and after teachers' evaluations. The dotted line shows humanities scores of students of non-evaluated math teachers at exams taken on the same years. The solid line in Figure 2 (b) shows the difference in humanities scores between students of evaluated and non-evaluated math teachers before and after evaluations. The dotted lines show confidence intervals.



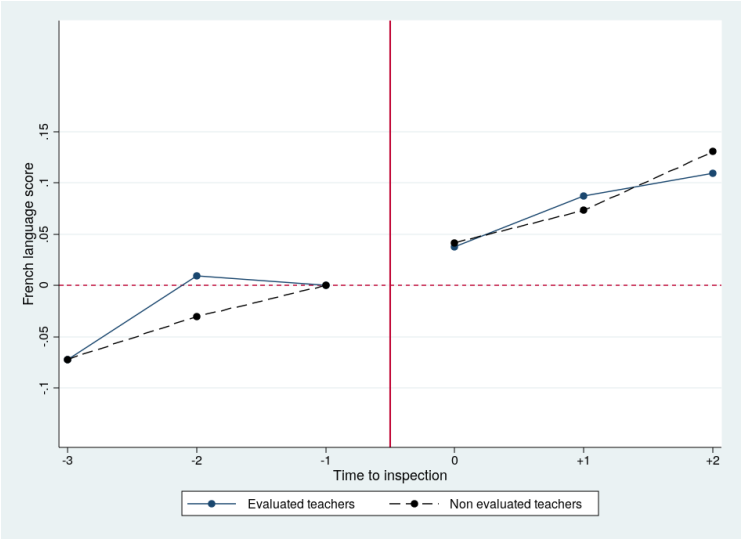
(a)



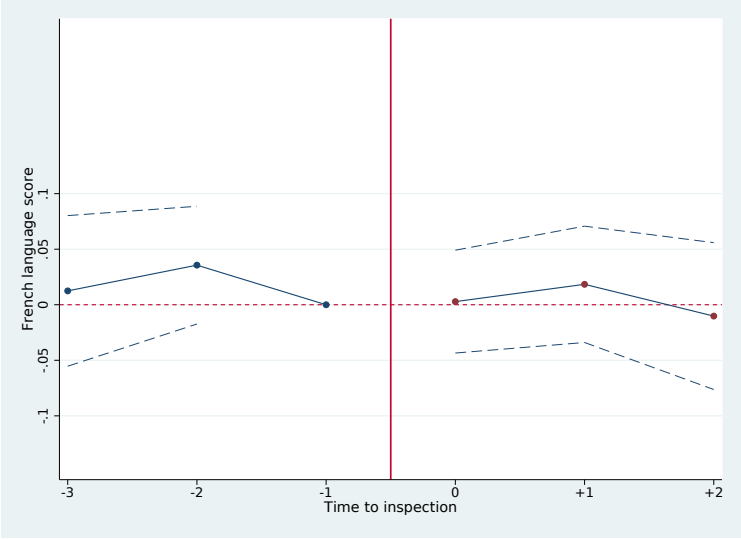
(b)

Figure 3: French language teacher evaluation and student performance in math

Note: The solid line in Figure 3 (a) shows math scores of students of evaluated French language teachers before and after teachers' evaluations. The dotted line shows math scores of students of non-evaluated French language teachers at exams taken on the same years. The solid line in Figure 3 (b) shows the difference in math scores between students of evaluated and non-evaluated French language teachers before and after evaluations. The dotted lines show confidence intervals.



(a)



(b)

Figure 4: French language teacher evaluation and student performance in French language

Note: The solid line in Figure 4 (a) shows French language scores of students of evaluated French language teachers before and after teachers' evaluations. The dotted line shows French language scores of students of non-evaluated French language teachers at exams taken on the same years. The solid line in Figure 4 (b) shows the difference in French language scores between students of evaluated and non-evaluated French language teachers before and after evaluations. The dotted lines show confidence intervals.

Appendix A - Additional Tables and Graphs

Descriptive statistics

Table A1: *Inspecteurs*' characteristics

	(1) Math	(2) French language
Inspecteurs' individual characteristics		
Age	51.50	53.37
	(7.42)	(7.20)
Experience as *inspecteur*	6.31	7.11
	(3.94)	(4.36)
Female	0.33	0.58
	(0.47)	(0.49)
Total nb of *inspecteurs*	142	165
Regional characteristics		
Nb of *inspecteurs* per region	4.7	5.5
	(2.5)	(3)
Nb of teachers per region	1676	2208
	(1013)	(1326)
Nb of evaluations per region	252	310
	(143)	(163)
Total nb of regions	31	31

Note: The table refers to the population of *inspecteurs* working for the Ministry of Education during academic year 2008-2009. The upper part of the table shows their average age, number of years of experience and gender, separately for math *inspecteurs* (column (1)) and French language *inspecteurs* (column (2)). The lower part of the table shows the average number of *inspecteurs*, teachers, evaluations per region (separately for math and French Language). Standard deviations are in parentheses.

Table A2: Student characteristics - difference between priority and non priority schools

	Priority schools (1)	Non priority schools (2)	Difference (1) - (2)
Age	14.64	14.47	0.17**
	(0.24)	(0.18)	(0.01)
Female	0.51	0.51	-0.00
	(0.10)	(0.09)	(0.00)
Low-income	0.45	0.22	0.23**
	(0.20)	(0.14)	(0.01)
Average standardized test scores	-0.62	0.21	-0.83**
	(0.898)	(0.767)	(0.03)
Observations	1091	4144	5235

Note: The table shows the difference in students' average age as well as in the proportion of female students, low-income students and students' average scores at the end-of-middle school national exam, across priority and non-priority schools in 2008-2009. * p<0.10, ** p<0.05.

Table A3: Math teachers' evaluations and 9th grade teaching

	(1) All	(2) Female	(3) Male	(4) Low-exp	(5) High-exp	(6) Priority	(7) Non Priority
	0.008	0.013	0.002	0.009	0.006	0.010	0.009
	(0.006)	(0.009)	(0.009)	(0.009)	(0.009)	(0.014)	(0.007)
	[0.78]	[0.78]	[0.79]	[0.76]	[0.80]	[0.76]	[0.79]
Observations	39958	20757	19201	20450	19508	9246	30712

Note: the table refers to the sample of math teachers who teach 9th grade students on year t_0=2008-2009 and who are not evaluated during t_0. The table shows the result of regressing a dummy indicating that teachers teach 9th grade students on year t on a dummy indicating that teachers underwent an external evaluation between t_0 and t. Column (2) refers to the subsample of female teachers, column (3) to male teachers, column (4) to teachers whose number of years of teaching experience is below the median (i.e. above or below 11 years) and column (5) to teachers above this median. Eventually, columns (6) and (7) refer to teachers who were in education priority schools in 2008 and to those who were in non-priority schools in 2008, respectively. Standard errors are in parentheses. Sample means of the dependent variables are within square brackets. * p<0.10, ** p<0.05.

Table A4: Teachers' characteristics

	(1) Math	(2) French language
Experience (in 2008)	12.28	12.74
	(5.11)	(5.01)
Female teacher	0.52	0.83
	(0.50)	(0.37)
Priority schools (in 2008)	0.17	0.18
	(0.37)	(0.38)
Number of evaluations (N_e)		
$N_e = 0$	0.43	0.54
	(0.49)	(0.50)
$N_e = 1$	0.56	0.45
	(0.50)	(0.50)
$N_e > 1$	0.01	0.01
	(0.09)	(0.08)
Observations	30414	30779

Note: The table refers to our working sample of teachers who teach 9th grade students between t_0=2008-2009 and t_1=2011-2012. The table shows the mean characteristics of teachers in terms of number of years of teaching experience in 2008, gender and type of school in 2008, as well as the number of external evaluations that teachers underwent over the 4-year period under consideration. The first column refers to the subsample of math teachers whereas the second column refers to the subsample of French language teachers.

Balancing tests - tables and graphs

Table A5: Balancing test - 9th grade math teacher evaluation and student characteristics

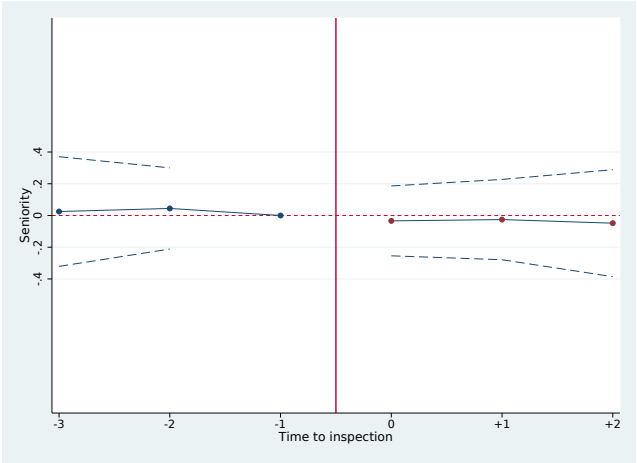
	(1) Age	(2) Female	(3) Low-income	(4) German	(5) Latin/Greek
All teachers (N=30414)					
Evaluation	0.004	-0.001	0.002	-0.000	0.002
	(0.004)	(0.003)	(0.003)	(0.004)	(0.003)
Female teachers (N=15724)					
Evaluation	0.010*	-0.004	0.005	-0.004	0.004
	(0.006)	(0.004)	(0.004)	(0.005)	(0.005)
Male teachers (N=14690)					
Evaluation	-0.001	0.002	-0.001	0.004	0.001
	(0.007)	(0.004)	(0.004)	(0.005)	(0.005)
Low-experience teachers (N=15072)					
Evaluation	0.005	0.002	0.003	-0.003	0.001
	(0.007)	(0.004)	(0.004)	(0.005)	(0.005)
High-experience teachers (N=15342)					
Evaluation	0.002	-0.003	-0.000	0.003	0.003
	(0.006)	(0.004)	(0.004)	(0.005)	(0.005)
Priority schools (N=6818)					
Evaluation	0.010	-0.010*	0.008	-0.000	-0.005
	(0.010)	(0.006)	(0.007)	(0.008)	(0.007)
Non Priority schools (N=23596)					
Evaluation	0.004	0.002	-0.000	0.000	0.005
	(0.005)	(0.003)	(0.003)	(0.004)	(0.004)

Note: the table shows the results of regressing 9th grade classes' average characteristics (average age of students, proportion of girls, proportion from low-income families, proportion studying German and proportion studying Latin or ancient Greek) on a dummy indicating that their math teacher underwent an evaluation between t_0=2008-2009 and t. The first row refers to the full working sample, whereas rows 2 to 7 refer to subsamples defined by teachers' gender, by teachers' number of years of experience (above or below 11 years) or by type of school attended (priority vs non-priority). Standard errors are in parentheses. * p<0.10, ** p<0.05.

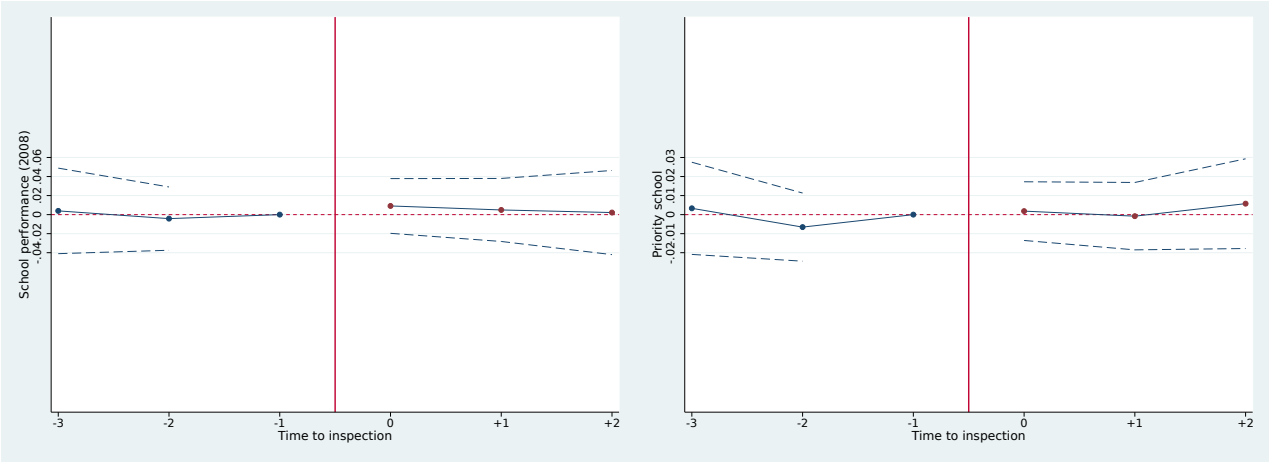
Table A6: Balancing test - 9th grade math teacher evaluation, teacher mobility and colleagues' characteristics

	(1) Teacher seniority	(2) Priority schools	(3) School performance	(4) Colleagues' experience	(5) Colleagues' seniority
All teachers (N=30414)					
Evaluation	0.035	0.004	-0.002	-0.039	-0.004
	(0.030)	(0.003)	(0.004)	(0.093)	(0.084)
Female teachers (N=15724)					
Evaluation	0.058	0.003	-0.002	-0.106	-0.109
	(0.042)	(0.003)	(0.006)	(0.131)	(0.118)
Male teachers (N=14690)					
Evaluation	-0.002	0.004	-0.000	0.042	0.120
	(0.043)	(0.004)	(0.006)	(0.133)	(0.121)
Low-exp (N=15072)					
Evaluation	0.071**	0.006	-0.007	-0.085	0.011
	(0.035)	(0.005)	(0.007)	(0.131)	(0.119)
High-exp (N=15342)					
Evaluation	-0.008	0.002	0.004	-0.007	-0.024
	(0.048)	(0.002)	(0.004)	(0.133)	(0.120)
Priority schools (N=6818)					
Evaluation	0.107	0.007	0.000	-0.048	0.179
	(0.081)	(0.009)	(0.014)	(0.193)	(0.172)
Non priority schools (N=23596)					
Evaluation	0.016	0.003*	-0.002	-0.015	-0.046
	(0.031)	(0.002)	(0.004)	(0.107)	(0.097)

Note: the table shows the results of regressing teacher seniority, school characteristics (priority school, school performance) and colleagues' characteristics (experience, seniority) on a dummy indicating that the math teacher underwent an evaluation between t_0=2008-2009 and t. School performance in column (3) is the average math test score in 2008 of the school in which the math teacher teaches in year t. Eventually, colleagues' experience and seniority in columns (4) and (5) refer to the average characteristics of the 9th grade French language and history teachers who teach the same 9th grade students as the math teacher in year t. Standard errors are in parentheses. * p<0.10, ** p<0.05.



(a) Seniority



(b) School performance

(c) Priority school

Figure A1: Math teacher evaluation and teacher mobility

Note: The solid lines in Figure A1 (a) to A1 (c) show the difference between evaluated and non-evaluated math teachers before and after evaluations in terms of teacher seniority (a), school performance as measured by the school average math test scores in 2008 (b) and teacher probability to teach in a priority school (c). The dotted lines show confidence intervals.

French teachers evaluation and student performance

Table A7: 9th grade French language teacher evaluation and student performance

	End of middle school test scores			
	French language	Math	French language	Math
	(1)	(2)	(3)	(4)
Evaluation	0.016	0.015		
	(0.016)	(0.015)		
Evaluation on t			0.006	0.015
			(0.016)	(0.016)
Evaluation before t			0.028	0.010
			(0.020)	(0.020)
Observations	30779	30779	30779	30779

Note: The table refers to our working sample of French language teachers who teach 9th grade students between t_0=2008-2009 and t_1=2011-2012. Column (1) (column (2)) shows the result of regressing their students' average score in French language (mathematics) at the end of year t on a dummy indicating that they underwent an external evaluation between t_0 and t. Column (3) (column (4)) shows the result of regressing the same dependent variable on a dummy indicating that they underwent an external evaluation on t and on a dummy indicating that they underwent an evaluation between t_0 and $t-1$. Models include a full set of teachers and year fixed effects as well as controls for students' average age, gender, family social background, German language study and Ancient language study. Standard errors are in parentheses. * p<0.10, ** p<0.05.

Table A8: 9th grade French language teacher evaluation and student performance - by subgroups

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	All	Female	Male	Low-exp	High-exp	Priority	Non Priority
French language score	0.016	0.023	-0.003	0.010	0.019	0.076**	-0.003
	(0.016)	(0.017)	(0.039)	(0.023)	(0.021)	(0.035)	(0.017)
Mathematics score	0.015	0.016	0.010	0.011	0.018	0.035	0.008
	(0.015)	(0.017)	(0.038)	(0.022)	(0.021)	(0.032)	(0.017)
Observations	30779	25601	5178	14135	16644	7027	23752

Note: The table refers to our working sample of French language teachers who teach 9th grade students between t_0=2008-2009 and t_1=2011-2012. The first (second) row shows the results of regressing their students' average score in French language (mathematics) at the end of year t on a dummy indicating that they underwent an external evaluation between t_0 and t. The first column refers to the full sample, whereas columns (2) and (3) refer to the subsamples of female and male teachers, columns (4) and (5) to the subsamples of teachers whose number of years of work experience is either above or below the median on t_0 (i.e., above or below 11 years), columns (6) and (7) to the subsample of teachers who were in education priority schools on t_0 and the subsample who were in non-priority schools. Models include a full set of teachers and year fixed effects as well as controls for students' average age, gender, family social background, German language study and Ancient language study. Standard errors are in parentheses. * p<0.10, ** p<0.05.

Table A9: 9th grade French language teacher evaluation and student performance by French language subtopic test scores and by subgroups

	(1) All	(2) Female	(3) Male	(4) Low-exp	(5) High-exp	(6) Priority	(7) Non Priority
Reading test scores	0.008	0.013	-0.009	-0.005	0.017	0.066*	-0.010
	(0.015)	(0.016)	(0.038)	(0.023)	(0.020)	(0.034)	(0.017)
Writing test scores	0.028	0.036*	0.005	0.038	0.017	0.077*	0.014
	(0.019)	(0.020)	(0.047)	(0.028)	(0.025)	(0.043)	(0.020)
Observations	30778	25600	5178	14135	16643	7027	23751

Note: The table refers to our working sample of French language teachers who teach 9th grade students between t_0=2008-2009 and t_1=2011-2012. The first (second) row shows the results of regressing their students' average score in reading (writing) at the end of year t on a dummy indicating that they underwent an external evaluation between t_0 and t. The first column refers to the full sample, whereas columns (2) and (3) refer to the subsamples of female and male teachers, columns (4) and (5) to the subsamples of teachers whose number of years of work experience is either above or below the median on t_0 (i.e., above or below 11 years), columns (6) and (7) to the subsample of teachers who were in education priority schools on t_0 and the subsample who were in non-priority schools. Models include a full set of teachers and year fixed effects as well as controls for students' average age, gender, family social background, German language study and Ancient language study. Standard errors are in parentheses. * p<0.10, ** p<0.05.

Math and French language teachers' external evaluations and student performance

Table A10: Math and French language teachers' evaluations and student performance - by subgroups

	(1) All	(2) Female	(3) Male	(4) Low-exp	(5) High-exp	(6) Priority	(7) Non Priority
Score in the subject	0.030**	0.028**	0.038**	0.031**	0.029**	0.083**	0.014
	(0.010)	(0.013)	(0.018)	(0.015)	(0.014)	(0.023)	(0.012)
Score in other subjects	0.008	0.007	0.010	0.009	0.008	0.015	0.007
	(0.010)	(0.013)	(0.017)	(0.015)	(0.014)	(0.022)	(0.011)
Observations	61187	41321	19866	29204	31983	13842	47345

Note: The table refers to joint sample of math and French language teachers who teach 9th grade students between t_0=2008-2009 and t_1=2011-2012. The first (second) row shows the results of regressing their students' average score in the subject they teach (subjects they don't teach) at the end of year t on a dummy indicating that they underwent an external evaluation between t_0 and t. The first column refers to the full sample, whereas columns (2) and (3) refer to the subsamples of female and male teachers, columns (4) and (5) to the subsamples of teachers whose number of years of work experience is either above or below the median on t_0 (i.e., above or below 11 years), columns (6) and (7) to the subsample of teachers who were in education priority schools on t_0 and the subsample who were in non-priority schools. Models include a full set of teachers and year fixed effects as well as controls for students' average age, gender, family social background, German language study and Ancient language study. Standard errors are in parentheses. * p<0.10, ** p<0.05.

Appendix B - Data construction

This paper uses an administrative database with detailed information on secondary school teachers for the period between t_0=2008-2009 to t_1=2011-2012. For each teacher j, this dataset gives information on whether (and when) j underwent an external evaluation between t_0 and t_1. It also gives information on whether (and when) teacher j taught 9th grade students and on the average performance of these students at exams taken at the end of 9th grade as well as at exams taken subsequently at end of high school. In this appendix, we explain how we build this database.

To construct this working file, we use three exhaustive administrative databases. The first one is the *Fichier Anonymisé d'Élèves pour la Recherche et les Études* (herafter, FAERE). For each academic year, it provides information on all secondary school students, including their socio-demographic characteristics, their ID number, the ID number of their class, their choice of field of study at the end of 10th grade as well as their results at the (externally set and marked) national exams taken at the end of middle school (9th grade) or at the end of high-school (12th grade). The exam taken at the end of middle school involves three written tests (in math, French language and history-geography) and we know students' scores at these different tests. We also know whether students choose science as major field of study at the end of 10th grade and whether they graduated in science at the end of 12th grade.

Using this individual level database, it is possible to build a class level database providing for each 9th grade class observed between 2008-2009 and 2011-2012 (a) the ID of the class and the academic year when the class is observed, (b) the average scores of the students of the class in math and humanities at exams taken at the end of the academic year (i.e. at the end of 9th grade), (c) the proportion of students of the class who will subsequently choose science as major field of study at the end of 10th grade (d) the proportion of students who subsequently succeed in graduating in science at the end of 12th grade.

The second database is an administrative dataset - called base *Relais* - which provides for each class observed between 2008-2009 and 2011-2012 the ID number of the class and the ID number of its teachers. This dataset makes it possible to augment our class-level database with information on the IDs of the math and French language teachers of each 9th grade class.

Eventually, we used the *Annuaire du Personnel du Secondaire Public* (herafter APSP). For each academic year, it provides information on the background characteristics of all teachers from public secondary schools (ID number, age, gender, level of experience, qualifications). For each teacher j and each academic year t, we also know whether j is evaluated during t. This dataset makes it possible to augment the class level database with information on math and French language teachers, and most

notably with information on whether (and when) they underwent an external evaluation between 2008-2009 and 2011-2012[19].

Overall, we get a class level database covering the period from 2008-2009 to 2011-2012 and providing for each 9th grade class observed during this 4-year period (a) the ID number of the class and the academic year when it is observed, (b) the ID number and socio-demographic characteristics of its math and French language teachers, (c) the date of the external evaluations that its math and French language teachers underwent during this 4-year period and (d) the average outcomes of its students at the end of 9th grade as well as their subsequent outcomes at the end of 10th grade or 12th grade.

Eventually, by averaging the variables of this database at the teacher x year level, we build a database which makes it possible to explore the extent to which teachers' external evaluations are followed by an improvement in their effectiveness, as measured by their ability to prepare 9th grade students for the end-of-middle school exams or by their ability to induce 9th grade student to choose science as major field of study in high school and to graduate in science.

[19]For each education region r and each academic year t, the APSP also provide background information on *inspecteurs* assigned to region r during t, namely information on their age, gender, level of experience as well as on their previous position within the French administration. Note, however, that we have no information on the specific teachers that were evaluated by each specific *inspecteurs*. It is not possible to match specific teacher's evaluations with specific inspecteurs.

Appendix C - Robustness checks

Table C1: Robustness checks - 9th grade math teacher evaluation and student performance - by subgroups

	(1) All	(2) Female	(3) Male	(4) Low-exp	(5) High-exp	(6) Priority	(7) Non Priority
Math score	0.043**	0.035*	0.053**	0.054**	0.036*	0.091**	0.030*
	(0.014)	(0.019)	(0.020)	(0.020)	(0.019)	(0.029)	(0.015)
Humanities score	0.005	-0.001	0.010	0.007	0.005	0.009	0.006
	(0.013)	(0.019)	(0.020)	(0.020)	(0.018)	(0.031)	(0.015)
Observations	32379	16906	15473	15072	17307	7029	25350

Note: The table refers to the same working sample of math teachers as Table 1, augmented by teachers with more than 25 years of teaching experience. The first (second) row shows the results of regressing their students' average score in math (humanities) at the end of year t on a dummy indicating that they underwent an external evaluation between t_0=2008-2009 and t. The first column refers to the full sample, whereas columns (2) and (3) refer to the subsamples of female and male teachers, columns (4) and (5) to the subsamples of teachers whose number of years of work experience is either above or below the median (i.e., above or below 11 years), columns (6) and (7) to the subsample of teachers who were in education priority schools on t_0 and the subsample who were in non-priority schools. Models include a full set of teachers and year fixed effects as well as controls for students' average age, gender, family social background, German language study and Ancient language study. * p<0.10, ** p<0.05.

Table C2: Robustness check - 9th grade math teacher evaluation and student high school outcomes

	(1) All	(2) Female	(3) Male	(4) Low-exp	(5) High-exp	(6) Priority	(7) Non Priority
Science as major field	0.004**	0.002	0.007**	0.008**	0.001	0.009**	0.003
	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)	(0.004)	(0.002)
	[0.179]	[0.187]	[0.170]	[0.161]	[0.194]	[0.124]	[0.194]
Graduation in science	0.004**	0.001	0.008**	0.007**	0.002	0.009**	0.002
	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.002)
	[0.151]	[0.158]	[0.143]	[0.135]	[0.166]	[0.100]	[0.165]
Observations	32379	16906	15473	15072	17307	7029	25350

Note: The table refers to the same working sample of math teachers as Table 1, augmented by teachers with more than 25 years of teaching experience. The first row shows the result of regressing the proportion of their 9th grade students who will choose science as major field of study at the end of 10th grade on a dummy indicating that they underwent an external evaluation between t_0 and t. The second row shows the result of regressing the proportion of their 9th grade students who will graduate in science at the end of 12th grade on the same independent variable. The first column refers to the full sample, whereas columns (2) to (7) refer to subsamples defined by teachers' gender, number of years of teaching experience (above/below 11 years), type of school attended (priority/non priority). Models include a full set of teachers and year fixed effects as well as controls for students' average age, gender, family social background, German language study and Ancient language study. Sample means of the dependent variables are within square brackets. * p<0.10, ** p<0.05.