There's always room for improvement: the persistent benefits of repeated teacher evaluations*

Simon Briole & Éric Maurin†

Paris School of Economics, France

July 21, 2020

Abstract

In France, secondary school teachers are evaluated every five to six years by senior experts of the Ministry of Education. These evaluations involve the supervision of one class session, a debriefing interview and their results are used to determine teachers' career advancement. We show that these repeated evaluations contribute to improving teachers' effectiveness (as measured by their students' performance) at all stages of their career. Positive effects on students' outcomes persist over time and are particularly salient in education priority schools, in contexts where teaching is often very challenging.

JEL classification: I20; I28; J24

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

*We would like to thank Clément de Chaisemartin, Manon Garrouste, Marc Gurgand, Élise Huillery and Sandra McNally for helpful comments on previous versions of this paper as well as participants of the joint UCL-NHH-PSE workshop on human capital accumulation in Paris. We would also like to thank the French Ministry of Education (MEN-DEPP) for providing us with the administrative data exploited in this paper. Eventually, we acknowledge the support of the Norface Dynamics of Inequality Across the Life-course (DIAL) Joint Research Program (research Project file number 462-16-090, entitled Human capital and inequality during adolescence and working life) and the Agence Nationale pour la Recherche (project ANR-17-0004-01). This paper was previously circulated under the title "Does evaluating teachers make a difference?".

†E-mail address: eric.maurin@ens.fr.

Introduction

Teacher quality has long been identified as a key determinant of student achievement.[1] There is also a growing body of experimental evidence showing that teacher evaluation can be a way to improve teacher quality.[2] However, teacher evaluation still take many different forms across the world and there is little consensus on what a good evaluation system should be. In particular, it is not clear whether teacher evaluation has the same impact on all teachers in all school contexts or whether it is more effective to focus on the most challenging contexts and on teachers identified as having the most difficulty with their classes. Similarly, it is still not clear whether teachers can benefit from being evaluated several times during their career or whether it is preferable to focus on teachers at the early stages of their career. Answering these questions, however, is crucial for determining the best possible allocation of resources for teacher evaluation.

To shed light on these issues, this paper investigates the short and long-term effects of the long-established teacher evaluation system in France. It builds on administrative data with exhaustive information on the exact timing of teachers' evaluations in a context where all secondary school teachers are evaluated every five to six years by senior experts of the Ministry of Education who specialize in this task. Evaluations encompass the supervision of one class session, a debriefing interview and the writing of an official evaluation report which includes the teacher rating. They represent key stages of teachers' career and their results have a direct incidence on teachers' wage and promotion opportunities. In this context, it is possible to compare student performance just before and after teacher evaluations so as to test precisely whether evaluations are followed by an improvement in performance at national exams. Our data also makes it possible to track students over time in order to explore whether teacher evaluations have long term impact on student subsequent school career.

We first provide evidence that the visit of a math teacher by an external evaluator is followed by a significant increase (of about 4.4% of a SD) in his or her students' math scores on national examinations taken at the end of middle school (9th grade). The effect of math teachers' evaluations is observed on performance in math but not in other subjects, consistent with the assumption that increased performance in math are driven by improved teaching practices of math teachers, not by changes in students' overall quality or by an increase in math workload (which would be detrimental to performance in other subjects). Furthermore, math teachers' increased effectiveness is observed not only at the end of the evaluation

[1] See e.g., Rockoff (2004), Rivkin et al. (2005), Hanushek & Rivkin (2006), Aaronson et al. (2007), Hanushek & Rivkin (2010) or Chetty et al. (2014)

[2] See Taylor & Tyler (2012), Murphy et al. (2018) or Burgess et al. (2019). For international comparisons, see Isoré (2009), OECD (2013a,b) or Jackson et al. (2014)

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.

Further analysis reveal that the effect of evaluations is even more significant for math teachers assigned to education priority schools (the 25% most deprived), in contexts where students' academic level is often very weak and teaching more challenging. The impact is also stronger on the lowest performing teachers. On the other hand, the effect of an additional external evaluation appears to be about as strong for more experienced teachers as for less experienced ones. According to our preferred estimates, repeated external evaluations of math teachers explain on average about 30% of the increase in their teaching effectiveness over the course of their career. External evaluations have smaller effects on French language teachers than on math teachers, except in priority schools (+11% of a SD per additional evaluation). A closer look shows that the evaluations of French language teachers have a significant effect on the results of the writing tests (which likely capture the most advanced skills) but little effect on the results of the dictation/comprehension tests.

Generally speaking, our paper identifies the effect of teachers' external evaluations under the assumption that teachers are not able to manipulate the timing of evaluations or the composition of classes they have to teach. The fact that external evaluations in mathematics do not coincide with any significant improvement in students' performance in other subjects (and vice versa) is consistent with our identification assumption. Also, we checked that external evaluations do not coincide with teachers changing schools or with changes in the likelihood of teaching in priority education areas. Nor do they coincide with changes in the socio-demographic and academic characteristics of the classes to which teachers are assigned.

To further test our identification assumption, we also checked that our main findings were robust to focusing on the education regions where the timing of external evaluations is the most constrained and where teachers have the least leeway to make external evaluations coincide with improvement in their teaching context. More in general, it is possible to define (in each education region) a local modal spacing between successive external evaluations that cannot be manipulated by teachers and that accurately predicts the actual spacing. When identification is based on the predicted spacing rather than the actual

spacing, estimated effects of external evaluations tend to become even stronger. Eventually, we developed a school-level analysis so as to obtain estimates that are robust to potential spillover effects from evaluated to non-evaluated teachers as well as to potentially endogenous students-to-teacher assignment within school. This approach confirms that variations in the proportion of recently evaluated teachers in a school are followed by parallel variations in students' average math performance in the same school, the corresponding school-level impact being consistent with basic teacher-level estimates.

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 and low-skilled teachers (Rockoff (2008); Glazerman et al. (2008, 2010); Papay et al. (2020)) as well as programs of formal training and professional development (Angrist & Lavy (2001); Harris & Sass (2011)) and policies designed to evaluate and provide feedback to teachers (Weisberg et al. (2009); Allen et al. (2011); Taylor & Tyler (2012); Murphy et al. (2018); Burgess et al. (2019)). Another strand of this literature shows that teacher evaluation can influence the quality of the teaching workforce through its impact on teacher retention (Dee & Wyckoff (2015); Cullen et al. (2016); Sartain & Steinberg (2016); Dee et al. (2019)). Generally speaking, most existing papers focus on the effect of introducing new and local evaluation programs on teachers who had not been systematically evaluated before. They are suggestive that introducing such programs can have significant short-term effects, especially when they are high intensity or when they are targeted at voluntary low-achieving schools or at beginning/low-performing teachers. In this paper, we explore the effects of a long-established evaluation system implemented repeatedly and nationwide. Our first contribution is to show that such a system of repeated evaluations can produce both short-term and long-term effects at all stages of a teacher's career, and not just at the beginning. Our second contribution is to show that long-established teacher evaluation systems can produce significant benefits, even when they are implemented at scale and reach their long-term equilibrium. This is key given the recurrent difficulty of scaling-up and sustaining policies that have proven to be cost-effective in local and experimental settings (Todd & Wolpin (2003), Banerjee et al. (2017)). Our last contribution is to precisely document the heterogeneity of impacts across teachers and school contexts. The impact of evaluations on the average performance in math and French language appears to be significantly stronger in the most deprived areas as well as on the most struggling teachers. It has long been established that teacher effectiveness can vary considerably from one teacher to another[3] and our analysis shows that repeated

[3]For an overview of the methods used to explore teacher heterogeneity, see, for example Jacob & Lefgren (2008), Kane et al. (2011) or Bacher-Hicks et al. (2017).

external evaluations may be one way to alleviate this problem and to reduce the gap between the least and the most effective teachers.

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. Building on alternative identification strategies, section 5 provides further evidence on the robustness of our results. The final section concludes with a brief discussion on the implications of our results.

1 Institutional context

In France, secondary school teachers are civil servants, recruited through national competitive exams organized each year in each field of study by the Ministry of Education.[4] 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 five to six years by senior experts of the Ministry of Education. At every stage of their career, teachers who get the best evaluations are promoted at a much 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.[5] 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). Internal evaluations by school heads focus on teachers' general behavior (punctuality, attendance, participation in the life of the school, etc.) while external evaluations focus on teaching practices. Teachers' evaluations determine their career advancement, but play no role in their ability to change school or region. The request from teachers to change schools are processed through a centralized

[4]The vast majority (93%) is 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%) is 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.

[5]After 20 years of teaching experience, teachers who never got good evaluation grades and were promoted solely based on experience (regular track) earn about 31,000 euros gross per year, while teachers who got the best evaluation grades and were promoted through the fastest track earn about 36,000 euros. Table A1 in the appendix provides further details on promotion speed at every stage of the career depending on evaluation results.

system that takes into account their family situation and length of service in their current school, but gives the same opportunity to the best and worst rated teachers (Terrier (2014)).

External evaluators

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*). *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 A2 in the online appendix).

Once recruited, each *inspecteur* is assigned to a specific education region by a centralized assignment system. There are 26 education regions in mainland 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 2,300 math teachers (Table A2).[6] According to the same data source, about 350 math teachers are evaluated each year, in each region. While the vast majority of evaluations are conducted by *inspecteurs* themselves, a small fraction is conducted by senior teachers temporarily appointed to help *inspecteurs*.[7] Assuming that 85% of these evaluations are conducted by *inspecteurs*, it means that each *inspecteur* conducts on average about 60 evaluations per year. As a consequence, *inspecteurs* develop an evaluation expertise and are recognized as pedagogical specialists by teachers, who trust their judgement and value their feedback (Albanel (2012)).

Timing of external evaluations

Inspecteurs are required to evaluate teachers once at the very beginning of their careers. Thereafter, the spacing between teacher's successive evaluations is decided by the *inspecteurs* of their education region, with teachers generally being notified a few weeks in advance. However, to the extent that external evaluations are necessary for the advancement of teachers, *inspecteurs* are required to evaluate each teacher as regularly as possible. They avoid as much as possible evaluating the same teacher at

[6] Overall, there were 135 math *inspecteurs* and 157 French language *inspecteurs* in France in 2008

[7] 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*.

close intervals or, conversely, no longer evaluating her at all. In practice, it is very rare for an external evaluation to occur either less than 4 years or more than 7 years after the previous one. In most regions, the majority of gaps between successive evaluations are either 4-5 years or 5-6 years and the date of each evaluation is strongly predicted by the date of the previous one. In the last section of the paper, we check that our main results are robust to using the predicted timing of evaluations rather than their actual timing as a source of identification.

Eventually, let us emphasize that the composition of classes for academic year t and the assignment of teachers to the different classes for the same year t, are decided by principals at the end of academic year *t-1*, at a time when it is impossible to know who will be evaluated on *t*. In this context, it is unlikely that principals can take evaluations into account when deciding on the composition of classes or on the assignment of teachers to the different classes.

Content of external evaluations

During each evaluation, *inspecteurs* have to conduct a number of activities defined by the legal framework regulating the teacher external evaluation protocol.[8] These activities mostly involves the supervision of one class session and a debriefing interview with the evaluated teacher, during which the *inspecteur* provides feedback and advice. *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., in order to avoid as much as possible to evaluate teachers on the basis of a single class-observation. 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. According to the survey conducted by Albanel (2012), about 17% of teachers receive a very good mark (with congratulations), about 56% receive a good mark (with encouragement to continue their efforts) and about 27% receive a lower mark (with reservations and advices to correct problems). In general, teachers are notified well in advance of the visit of the inspecteur, although the exact day of the visit remains generally unknown to them. 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

[8] Teachers have a legal right to contest the result of evaluations that do not comply to the protocol.

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.[9] 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)).[10] 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 60 evaluations, meaning about 350-500 euros per evaluation.[11] Given that there is only one evaluation every five or six years, the cost per teacher and year is about 80 euros.

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.[12] The 20% most underprivileged middle-schools benefit from education priority programs which provide them with additional resources.[13]

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

[9]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.

[10]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.

[11]In comparison, the Cincinnati program analyzed by Taylor & Tyler (2012) involves a total budget of about 7,500 dollars per evaluation, namely a cost that we estimate to be about 20 times more important than our program. As it happens, the Cincinnati program involves the observation of four classroom sessions during the year of the evaluation, three by an external expert and one by an internal one. Also, 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.

[12]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.

[13]As shown in table A3 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.

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.

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. The most prestigious field of study (and the one that opens up the most opportunities in higher education) is field S, which is also the field of study where the best students are concentrated. Another basic research question will be whether the effect of evaluations can still be seen one year later on field choice, especially on the probability of joining field S. The first year of high school (10th grade) is dedicated 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. 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, especially on students' ability to graduate in field S.

2 Data and samples

In this paper, we use administrative data with detailed information on secondary school teachers in mainland France for the period between academic years 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 extract from our main database the sample of math teachers who have less than 25 years of teaching experience, who were not evaluated in t_0[14] and who taught 9th grade students in t_0 as well as one additional time after t_0. This working sample includes 9,053 math teachers (i.e., about 75% of the total number of math teachers) and represents 29,156 (teacher x year) observations in total.

We provide some descriptive statistics in online Appendix A (see column (1) of Table A4)). Most of our empirical analysis will be conducted on this working sample. About half of teachers in this 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.[15]

Although a very low share of teachers in France leave the profession each year, one potential issue with this working sample is that external evaluations may have an impact on teachers' probability to teach 9th grade students one additional time 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 teachers who have less than 25 years of teaching experience and who were not evaluated in t_0 (N=10,140) 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 A5, 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 79% for non-evaluated teachers and only about 0.9 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 analysis on subsamples defined by school type (priority/non priority), teacher experience, teacher gender or teacher performance level in t_0 (as proxied by their students' performance in t_0, standardized by their number of years of teaching experience). In particular, the impact of one additional external evaluation between t_0 and t on attrition probability appears to be non significant regardless of whether we focus on teachers who fall in the bottom 30% of

[14] 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 (in 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. 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).

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

the teacher performance distribution or on teachers in the top 70% of this distribution.[16] Overall, we get an array of results suggesting that differential attrition is negligible. This result is consistent with the fact that teachers are civil servants from the central state, recruited on a life-long basis after difficult national competitive examinations, with a very protective status and very low quit rates (less than 0.1% in 2010-2011, according to the French Ministry of Education (2020)).

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.[17] 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

[16] Also, we do not find any evidence that teachers' behavior is influenced by the grade they got during evaluations taken between t_0 and t. In particular, we do not find evidence that teachers who obtain relatively weak grade at evaluations taken between t_0 and t are less likely to teach 9th grade students in t.

[17] A very small share of teachers in our sample (<1%) is evaluated twice during the period of observation (Table A4 in the Appendix). For these teachers, we only take into account the first evaluation occurring during this time period. All results presented in the paper are unchanged when we exclude teachers evaluated twice (available upon request).

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 this 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.[18] 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 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 the number of years they have been employed at their current school) and do not coincide with teachers moving to better schools. In

[18] 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.

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 section 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 two-way fixed effects 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..), as well as teacher j's time-varying characteristics. More precisely, it includes dummies controlling for each possible number of years of teaching experience in general and in the current school, 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 all regressions, error terms are clustered at the teacher level.

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 in 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 in t and a dummy (T_{2jt}) indicating that the evaluation took place between t_0 and $t-1$.[19]

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 A6 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). Taken together, these characteristics explain about 41% of the variance in math scores, but (consistent with our identifying assumption) the Table shows that none of them is related to the timing of external evaluations. 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.[20] 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 A7 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. Eventually, in the last section of the paper, we show that our results are robust to focusing on the education regions where the timing of external evaluations is the most constrained and where teachers have the least leeway to manipulate the date of their evaluations.

4.1 Main effects on math scores

The first column of Table 1 shows the basic effect of one additional evaluation on math teachers' effectiveness, as measured by their students' performance in math at end-of-middle school national exams.

[19] An alternative approach would be to estimate the effects of evaluations occurring at different stages of the school year. While we can precisely identify the (school) year of a given evaluation in our dataset, the month of evaluation is unfortunately missing or misspecified in most cases.

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

Consistent with our graphical analysis, it confirms that external evaluations are followed by a significant improvement in math score of about 4.4% of a SD. Given that about 58% of teachers were evaluated between 2008 and 2011 and that their students' performance increased on average by 8.3% of a SD during that same part of their career, we can estimate that repeated external evaluations explain about 30% of the improvement in math teachers' effectiveness over their career (where 0.3= 0.58 x 4.4/8.3). The second column shows the impact of external evaluations of math teachers on students' performance in humanities and, comfortingly, it shows no effect.[21] 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 (5.4% 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.

Some teachers in our working sample are externally evaluated in 2009, others are evaluated in 2010 or 2011 and others are never evaluated. The two-way fixed effect estimator used in this paper is a weighted average of all possible difference-in-difference (DD) estimators that compare these different groups of teachers to each other and over time (see e.g. Goodman-Bacon (2018)). Some elementary DD estimators compare teachers treated at a particular point in time with never treated teachers (where treatment=external evaluation) while others elementary DD compare groups of teachers treated at different points in time. One potential issue is that elementary weights can be negative which may bias two-way fixed effect estimates away from the sign of the true treatment effect. Building on Goodman-Bacon (2018), Table A8 in Appendix A focuses on the subsample of teachers observed in our working sample each year between t_0 and t_1 and shows the average effects and weights for the two basic types of DD used in this paper, namely those that compare treated and never treated teachers and those that compare groups of teachers treated at different point in time. Comfortingly, we obtain very similar estimates for both sources of identification and both weights are positive.

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,

[21] 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.

where 11 is the median number of years of experience in our sample), level of performance in t_0 (as proxied by whether the average performance of their students in t_0 falls or not in the bottom 30% of the performance distribution of teachers with a similar number of years of teaching experience), 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 (8.3% of a SD) than for teachers in non-priority schools (+3.4% 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.[22] Finally, the impact of external evaluations conducted between t_0 and t tends to be more significant on teachers whose performance level in t_0 is relatively weak. This last finding further suggests that external evaluations mainly help the most struggling teachers to catch up.

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 A6 and A7 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. 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

[22] 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)).

represent an increase of about 3% in this probability. Consistent with Table 2, the following columns show that this increase is more significant for teachers in education priority schools (+8%) as well as for teachers whose level of performance at t_0 is relatively weak (+6%). 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 in t_0, who are not evaluated in t_0 and who have less than 25 years of teaching experience in t_0. Tables 4 and 5 replicate Tables 1 and 2 using this working sample of French language teachers. These tables suggest that the evaluations of French language teachers are followed by improvements in student performance that are somewhat weaker than those that follow math teachers' evaluations (3.1% of a SD), except in priority education schools, where evaluations drive a 11% of a SD increase in student performance in French language. It can also be noted that the effects are much more significant for women than for men. However, this difference is difficult to interpret, as men represent only a very small and specific minority of French Language teachers (about 15%).

Generally speaking, these results are in line with the literature which shows that teacher effects tend to be 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)). To further explore the reasons for these weaker effects, we looked at the impact of French language teachers' evaluations separately on writing test scores and on dictation/reading comprehension test scores (see Table A9 in online appendix A).[23] We find that the impact of teacher evaluation is more significant on writing than on dictation/reading comprehension test scores. Specifically, the impact on writing test scores (+3.9% of a SD) is almost as significant as the impact on math test scores whereas the impact on dictation/reading comprehension test scores is not significantly different from zero at standard level.

[23]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.

These findings are consistent with the notion that open-ended exercises, such as writing exercises, have a better ability to detect students' progress than closed-ended exercises, such as reading comprehension exercises (see e.g. Kraft (2020)). They are also consistent with the literature showing that exercises are less likely to detect student progress when they focus on basic skills (such as spelling skills) rather than on advanced and recently acquired skills, such as math or writing skills (see Hopkins & Bracht (1975)). Overall, the lower impact observed on French language test scores may primarily reflect the fact that half of the French language exam deals with closed-ended questions and less advanced skills. The weaker effect on language exercises may also reflect that students learn language in many settings outside schools, so that the influence of teachers is diluted and distorted by that of other factors.

Eventually, we replicate our main graphical and regression analysis on the joint sample of math and French language teachers, so as to provide an estimate of the average effect of teachers' evaluations on end-of-middle school exams (see Table A10 as well as Figures A2(a) and A2(b)). Consistent with our previous findings, this analysis shows that the evaluation of a teacher is followed on average by a 4% of a SD increase in students' performance in the subject taught by the teacher (but has no effect on performance in the other subjects). This analysis also confirms that the average effect of teacher evaluations on students' performance is significantly stronger in education priority school (10% of a SD) than in non-priority ones (2% of a SD).

Not surprisingly, the effect shown in Table A10 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.[24]

5 Alternative strategies of identification

5.1 Predicted vs actual between-evaluation spacing

In the previous sections, the effect of external evaluations is identified under the assumption that teachers are not able to manipulate the timing of external evaluations. Specifically, we assume that

[24]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 80 euros (as discussed in section 2).

they are not able to make external evaluations coincide with improvement in the quality of students to whom they are required to teach or with improvement in their own teaching efficiency. A threat to our identification strategy would be that causality would go the other way, i.e. teachers would be able to request external evaluations and would request them only after they start teaching in good classes or only after they had gained enough teaching experience. Generally speaking, the fact that external evaluations in mathematics do not coincide with any significant improvement in students' performance in humanities (and vice versa) is consistent with our identification assumption. Also, it is reassuring that external evaluations do not coincide with changes in students' basic characteristics or with teachers changing schools or with changes in the probability of teaching in priority education areas. To take one step further, it is possible to rely on the above-mentioned fact that the spacing between two successive external evaluations is in practice strongly constrained by the obligation for *inspecteurs* to evaluate each teacher as regularly as possible.

In Appendix Table A11, we report (for each of the 26 mainland education regions) the main features of the distribution of the number of years elapsed since the previous external evaluation for teachers who were externally evaluated in 2008 and had been evaluated at least once before. The Table confirms that the spacing between two successive external evaluations is rarely less than 4 years. Also the Table shows that the local modal value of the distribution of between-evaluation spacing is generally 5 or 6 years and that the majority of spacings differ from the local modal value only by less than one year. In three regions, the variance of spacings is minimal with more than 75% of teachers being externally evaluated after exactly the same number of years since the last external evaluation.

When we focus on these three regions, the probability that a teacher gets evaluated between year t_0=2008 and year t is very directly determined by whether the number of years elapsed between her last evaluation and t is above or below the modal spacing between two evaluations in her region, namely very directly determined by a variable which cannot be manipulated. To be specific, if S_r represents the modal spacing between two external evaluations in region r and if t_i represents the year of the last pre-2008 evaluation for teacher i, the probability that i get evaluated between t_0 and t is very directly determined by a dummy variable Z_{it} indicating that the time elapsed between t_i and t is larger than $S_{r(i)}$ (where $r(i)$ is i's region). The first row of Table 6 shows the result of using Z_{it} as an instrumental variable to identify the effect of external evaluations in the three regions under consideration, focusing on teachers who have been externally evaluated at least once before 2008.

The first stage regression confirms that our instrument is very strong: the probability that a math teacher underwent an additional external evaluation between year t_0 and year t increases by more than

70 percentage points after the number of years elapsed between the last evaluation becomes larger than the local modal spacing between two evaluations. Columns 2 and 3 shows reduced-form impacts of the instrument on students' performance in math and humanities. Column 2 reveals that students' performance in maths at the end of middle school become about 10.1% of a SD stronger after the number of years elapsed since the last evaluation of their math teacher becomes larger than the local modal spacing between two evaluations. In contrast, we do not see any improvement in performance in humanities (Column 3).[25] Column 4 shows the corresponding instrumental variable estimate of the effect of math teachers' external evaluations on their students' performance in math. It suggests that each additional external evaluation induces an improvement of about 13.8% of a SD in students' performance in math. For the sake of comparison, column 5 shows that the estimated effect is about 8.2% of a SD when we use the same identification strategy (and OLS estimator) as in previous sections. Hence, the estimated effect of external evaluations on students' performance tends to be weaker when identification relies on the effective timing of external evaluations rather than on the timing predicted by the local spacing of evaluations (even though the difference between the IV and OLS estimates is not significant at standard levels).

One explanation for the weaker OLS estimate could be that, even in the three most disciplined regions, *inspecteurs* try to adapt the actual timing of evaluations to teachers' problems. The rare cases where evaluations occur earlier than normal would correspond to teachers who are most in difficulties with their classes (and need help) whereas the rare cases where evaluations occur later than normal would correspond to teachers who are most comfortable with their classes. Such a negative correlation between the actual timing of external evaluations and teachers' efficiency could explain why OLS estimates tend to underestimate the true effect of external evaluations and tend to be weaker than IV ones.

5.2 Is predictability a source of efficiency?

Whether one uses the actual timing of evaluations or the timing predicted by local modal spacing, the estimated effect of external evaluations appears to be stronger in the three regions where the timing is the most predictable. A first possible reason is that *inspecteurs* working in these regions are more focused on teachers' evaluations, which would explain why their evaluations are both more evenly distributed and more efficient than elsewhere. Another explanation could be that external evaluations tend to be all the

[25] To further check the validity of our instrument, we regressed math teachers' class average characteristics as well as school and colleagues' characteristics on our instrument and found no significant relationship (cf. Appendix Tables A12 and A13).

more effective if their timing is more predictable, with teachers being all the more motivated to improve their teaching practices when they know precisely when they will be evaluated.

To explore this issue, we extended the application of our IV strategy beyond the three regions where the timing of successive evaluations is the most predictable. In most of the other regions, the timing of successive evaluations is less predictable, but it is still possible to define thresholds in the distribution of between-evaluation spacings that correspond to clear jumps in the probability of being evaluated. Specifically, if P_{kr} denotes the probability that an external evaluation takes place k years after the previous one in region r, we observe in each region at least one value of k such that $P_{kr} - P_{k-1,r} > 0.10$, i.e., at least one value of k for which the increase in P_{kr} is greater than 0.10. If $k(r)$ represents the smallest such k in region r, the second row of Table 6 shows the results of an IV analysis for the full set of mainland regions, using as an instrument a variable (denoted Z_{it}) indicating whether the time elapsed since the last evaluation is above or below $k(r)$. Unsurprisingly, the first-stage effect is smaller (+14.5 percentage points) than the first-stage effect previously reported for the three regions with the most predictable timing, but it is still very significant from a statistical point of view. Most importantly, the corresponding IV estimate (+17% of a SD) is still much higher than the corresponding OLS estimates, even though it is less precisely estimated than the IV estimate that we get when we focus on the three regions were the spacing between evaluations is most predictable.

Overall, OLS estimates which rely on the actual timing of assessments (rather than on the timing predicted by local constraints) seem to be affected by an attenuation bias, and this bias seems in fact even stronger when focusing on regions where the timing is least constrained. These results are further consistent with the assumption that *inspecteurs* tend to evaluate at shorter intervals those teachers who have the most difficulty with their classes, so that evaluations relying on the actual timing of external evaluation likely underestimate the true impact of external evaluations on students' outcomes.

5.3 School-level analysis

So far, our analysis focuses on teachers who are continuously observed throughout the 2008-2011 period and relies on the assumption that the timing of their external evaluations is not related to the academic level of the classes to which they are assigned. We also implicitly assume that the external evaluation of one teacher has no influence on the effectiveness of other teachers in the same school. In this section, we develop a school level approach, so as to test the robustness of our main findings to alternative assumptions. Specifically, denoting Y_{st} the average performance in math of 9th grade students in school

s in year t, we consider the sample of schools observed throughout the 2008-2011 period and assume the following school level model:

$$Y_{st} = aT_{st} + bX_{st} + u_s + v_t + e_{st}$$

where T_{st} represents the proportion of 9th grade math teachers in school s in year t who benefited from an external evaluation between t_0 and t, while X_{st} is a vector of school level controls which captures 9th grade students' characteristics in school s in year t (proportion of female students, low income students, etc.). Finally, u_s and v_t represent the full set of school and year fixed effects. The parameter of interest a captures the direct effect of evaluations on evaluated teachers as well as potential spillover effects on teachers who are not evaluated over the period.

In this framework, the identifying assumption is simply that year-on-year variations in the proportion of recently evaluated math teachers in a school (as measured by T_{st}) are unrelated to year-on-year variations in the academic level of students entering 9th grade in this school. Since schools have no control over student admissions, catchment areas, or over the timing of their teachers' evaluations, they have very little leeway to adjust the proportion of recently evaluated teachers to the variation in the academic level of their students. Online appendix Table A14 shows the result of regressing our different school level controls X_{st} on T_{st}. These balancing checks confirm that there is no significant variation in students' average characteristics across years that correspond to different values of T_{st}, consistent with our identifying assumption.

Table 7 first shows the results of our school-level regressions when we use the average performance at the end-of-middle school exams as dependent variable. We find that the evaluation of 9th grade math teachers in a school is associated with a 5% of a SD increase in 9th grade students' average performance in math in this school, namely about exactly the same impact as the one previously obtained with the teacher level analysis. The Table further shows that this school level effect is significantly stronger in education priority schools than in non-priority schools, which is again consistent with our results based on a teacher-level analysis. Finally, the Table confirms that the effect on students' outcomes persists in high school, since students who benefited from a higher proportion of evaluated teachers at the end of middle school also shows higher graduation rates in science at the end of high school.

6 Conclusion

Evaluating employees' performance represents a major challenge for employers, both public and private. It involves being able to motivate and better compensate employees whose performance improve the most over time. But it also involves being able to identify employees who are most likely to benefit from an outside view of their work. In this paper, we shed light on these issues based on an in-depth analysis of the long-established teacher evaluation system gradually introduced by the French state during the 19th century.

In this system, all teachers are evaluated every 5 or 6 years by a senior expert of the Ministry of Education. These external evaluations involve the supervision of one class session, a debriefing interview, the writing of an evaluation report and represent key stages of teachers' career. Building on exhaustive longitudinal administrative data, we first provide robust evidence that each additional evaluation produces lasting effects on teacher effectiveness and student achievement. The effects are 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. These effects are significantly stronger on advanced skills (math skill or writing skill) than on more basic skills (reading or spelling).

We further show that the impact of teacher evaluation on students' performance in math or French language is much stronger in priority education schools than in non-priority schools. Reinforcing teacher evaluation in deprived areas appears as a way to reduce inequalities across school contexts. Regardless of the school context, the impact is also stronger on the lowest performing teachers. It has long been established that there is a lot of variation in teacher quality and our findings suggest that repeated external evaluations can be a way to alleviate this problem. Eventually, the impact of external evaluations on more experienced teachers is about as significant as on less experienced ones, which suggests that it can be efficient to repeatedly evaluate teachers all along their career, not simply at the start, as it is often the case.

Taken together, our findings suggest that a low-intensity large-scale evaluation program can be highly cost effective, even when it is generalized to all teachers and even after it has reached its long-term equilibrium. They also suggest that such a program can help reduce inequalities in education quality across schools and classes, even if it is not its primary objective. Further research is needed to explore whether it would not be even more effective to focus evaluations more on the most disadvantaged schools and the most struggling teachers.

References

Aaronson, D., Barrow, L., & Sander, W. 2007. "Teachers and student achievement in the Chicago public high schools". *Journal of labor Economics*, 25(1):95–135.

Albanel, X. 2012. "Le travail d'évaluation. L'inspection des professeurs de l'enseignement secondaire". *Spirale-Revue de recherches en éducation*, 49(1):107–121.

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.

Banerjee, A., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukerji, S., Shotland, M., & Walton, M. 2017. "From proof of concept to scalable policies: Challenges and solutions, with an application". *Journal of Economic Perspectives*, 31(4):73–102.

Burgess, S., Rawal, S., & Taylor, E. S. 2019. "Teacher peer observation and student test scores: Evidence from a field experiment in English secondary schools.". Working Paper.

Chetty, R., Friedman, J. N., & Rockoff, J. E. 2014. "Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood". *American economic review*, 104(9):2633–79.

Cullen, J. B., Koedel, C., & Parsons, E. 2016. "The compositional effect of rigorous teacher evaluation on workforce quality". *Education Finance and Policy*, pages 1–85.

Dee, T. S. & Wyckoff, J. 2015. "Incentives, selection, and teacher performance: Evidence from IMPACT". *Journal of Policy Analysis and Management*, 34(2):267–297.

Dee, T. S., James, J., & Wyckoff, J. 2019. "Is Effective Teacher Evaluation Sustainable? Evidence from DCPS". *Education Finance and Policy*, pages 1–53.

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

Feuillet, P. 2020. "Le devenir des enseignants entre la rentrée 2017 et la rentrée 2018". *Note d'information n20.16, Avril 2020, DEPP.*

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.*

Goodman-Bacon, A. 2018. "Difference-in-differences with variation in treatment timing". National Bureau of Economic Research.

Hanushek, E. A. & Rivkin, S. G. 2006. "Teacher quality". *Handbook of the Economics of Education*, 2: 1051–1078.

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.

Hopkins, K. D. & Bracht, G. H. 1975. "Ten-year stability of verbal and nonverbal IQ scores". *American Educational Research Journal*, 12(4):469–477.

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.

Kraft, M. A. 2020. "Interpreting effect sizes of education interventions". *Educational Researcher*, 49(4): 241–253.

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.

Papay, J. P., Taylor, E. S., Tyler, J. H., & Laski, M. E. 2020. "Learning Job Skills from Colleagues at Work: Evidence from a Field Experiment Using Teacher Performance Data". *American Economic Journal: Economic Policy*, 12(1):359–88.

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.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. 2005. "Teachers, schools, and academic achievement". *Econometrica*, 73(2):417–458.

Rockoff, J. E. 2004. "The impact of individual teachers on student achievement: Evidence from panel data". *American economic review*, 94(2):247–252.

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.

Sartain, L. & Steinberg, M. P. 2016. "Teachers' labor market responses to performance evaluation reform: Experimental evidence from Chicago public schools". *Journal of Human Resources*, 51(3):615–655.

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

Terrier, C. 2014. "Matching Practices for secondary public school teachers–France". *Matching in Practice*.

Todd, P. E. & Wolpin, K. I. 2003. "On the specification and estimation of the production function for cognitive achievement". *The Economic Journal*, 113(485):F3–F33.

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	0.044**	0.009		
	(0.014)	(0.014)		
Evaluation in t			0.039**	0.010
			(0.014)	(0.014)
Evaluation before t			0.054**	0.001
			(0.018)	(0.018)
Observations	29156	29156	29156	29156

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 in t and on a dummy indicating that they underwent an evaluation between t_0 and $t-1$. Models include a full set of teachers, academic region and year fixed effects as well as controls for teachers' experience and seniority and for students' average age, gender, family social background, German language study and Ancient language study. Standard errors (in parentheses) are clustered at the teacher level. * p<0.10, ** p<0.05.

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

	(1) Female	(2) Male	(3) Low-exp	(4) High-exp	(5) Low-perf	(6) High-perf	(7) Priority	(8) Non Priority
Math score	0.033*	0.056**	0.053**	0.039**	0.075**	0.020	0.083**	0.034**
	(0.019)	(0.020)	(0.020)	(0.019)	(0.024)	(0.016)	(0.031)	(0.015)
Humanities score	-0.004	0.023	0.017	0.004	0.012	0.003	0.006	0.012
	(0.019)	(0.020)	(0.020)	(0.019)	(0.026)	(0.016)	(0.032)	(0.015)
Observations	15318	13838	14319	14837	9651	19212	6265	22891

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. Columns (1) and (2) refer to the subsamples of female and male teachers, columns (3) and (4) to the subsamples of teachers whose number of years of work experience is either above or below the median in t_0 (i.e., above or below 11 years), columns (5) and (6) to the subsamples of teachers defined by whether their students' performance in math in t_0 falls (or not) in the bottom 30% of the distribution, columns (7) and (8) to the subsample of teachers who were in education priority schools in t_0 and the subsample who were in non-priority schools. Models include a full set of teachers, academic region and year fixed effects as well as controls for teachers' experience and seniority and for students' average age, gender, family social background, German language study and Ancient language study. Standard errors (in parentheses) are clustered at the teacher level. * 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) Low-perf	(7) High-perf	(8) Priority	(9) Non Prior.
Science as	0.005**	0.002	0.007**	0.009**	0.001	0.005	0.004	0.008**	0.003
major field	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.004)	(0.002)
	[0.178]	[0.184]	[0.171]	[0.163]	[0.192]	[0.128]	[0.204]	[0.124]	[0.192]
Graduation	0.004**	0.001	0.008**	0.007**	0.002	0.006**	0.003	0.008**	0.003
in Science	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.002)	(0.004)	(0.002)
	[0.150]	[0.156]	[0.144]	[0.136]	[0.164]	[0.104]	[0.174]	[0.100]	[0.164]
Observations	29156	15318	13838	14319	14837	9651	19212	6265	22891

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 (9) refer to subsamples defined by teachers' gender, number of years of teaching experience in t_0 (above/below 11 years), performance in t_0 (as proxied by their students' average performance in math in t_0), and type of school attended in t_0 (priority/non priority). Models include a full set of teachers, academic region and year fixed effects as well as controls for teachers' experience and seniority and 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.

Table 4: 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	0.031**	0.013		
	(0.015)	(0.015)		
Evaluation in t			0.020	0.015
			(0.016)	(0.016)
Evaluation before t			0.048**	0.007
			(0.019)	(0.019)
Observations	29507	29507	29507	29507

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 in t and on a dummy indicating that they underwent an evaluation between t_0 and $t-1$. Models include a full set of teachers, academic region and year fixed effects as well as controls for teachers' experience and seniority and for students' average age, gender, family social background, German language study and Ancient language study. Standard errors (in parentheses) are clustered at the teacher level. * p<0.10, ** p<0.05.

Table 5: 9th grade French language teacher evaluation and student performance - by subgroups

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Female	Male	Low-exp	High-exp	Low-perf	High-perf	Priority	Non Priority
French lang. score	0.037**	0.006	0.030	0.030	0.031	0.016	0.112**	0.006
	(0.016)	(0.038)	(0.022)	(0.020)	(0.028)	(0.017)	(0.035)	(0.016)
Mathematics score	0.014	0.011	0.003	0.021	0.025	-0.004	0.031	0.007
	(0.016)	(0.039)	(0.023)	(0.020)	(0.028)	(0.018)	(0.034)	(0.017)
Observations	24624	4883	13331	16176	9190	19256	6479	23028

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. Columns (1) and (2) refer to the subsamples of female and male teachers, columns (3) and (4) to the subsamples of teachers whose number of years of work experience is either above or below the median in t_0 (i.e., above or below 11 years), columns (5) and (6) to the subsamples of teachers defined by whether their students' performance in French language in t_0 falls (or not) in the bottom 30% of the distribution, columns (7) and (8) to the subsample of teachers who were in education priority schools in t_0 and the subsample who were in non-priority schools. Models include a full set of teachers, academic region and year fixed effects as well as controls for teachers' experience and seniority and for students' average age, gender, family social background, German language study and Ancient language study. Standard errors (in parentheses) are clustered at the teacher level. * p<0.10, ** p<0.05.

Table 6: 9th grade math teacher evaluation and students' performance: an IV approach

Panel A: Regions with exact timing	First stage (Eval.) (1)	Red-form (Math) (2)	Red-form (Humanities) (3)	IV (Math) (4)	OLS (Math) (5)
Z_{it}	0.727**	0.101**	-0.047	-	-
	(0.024)	(0.043)	(0.043)		
Actual evaluation	-	-	-	0.138**	0.082*
				(0.059)	(0.045)
Obs.	1967	1967	1967	1967	1967
Panel B: All regions	First stage (Eval.) (1)	Red-form (Math) (2)	Red-form (Humanities) (3)	IV (Math) (4)	OLS (Math) (5)
Z_{it}	0.144**	0.025*	0.002	-	-
	(0.008)	(0.014)	(0.014)		
Actual evaluation	-	-	-	0.171*	0.041**
				(0.101)	(0.015)
Obs.	26253	26253	26253	26253	26253

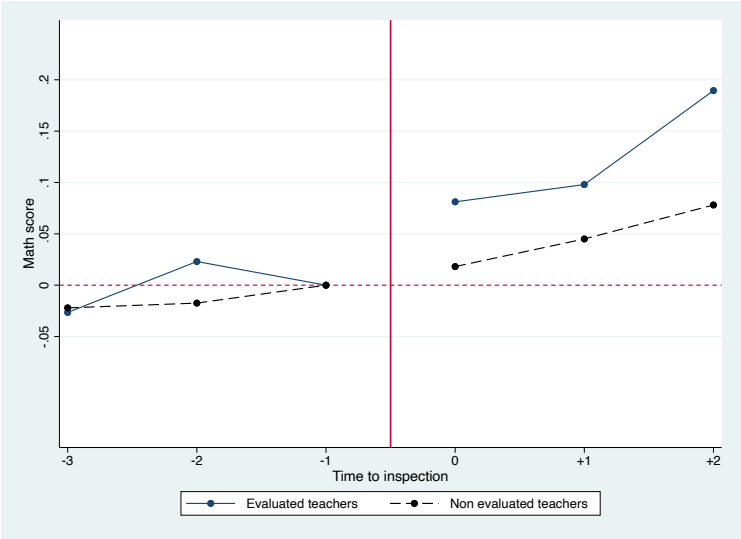
Note: the table shows the results of a regression analysis where we use Z_{it} as an instrumental variable, where Z_{it} is a dummy indicating that the number of years elapsed since the previous external evaluation is greater than the local minimum threshold defined in section 5. Panel A refers to the three education regions where the timing of external evaluations is the most predictable and where the local threshold used to define Z_{it} coincides with the modal value of the local distribution of between-evaluation spacing. Panel B refers to the full set of mainland regions. Column (1) shows the results of the (first stage) regression of the treatment (where treatment=one additional external evaluation between t_0 and t) on the instrument. Columns (2) and (3) show reduced-form regressions of math and humanities scores at the end of 9th grade on the instrument, column (4) shows the results of the corresponding IV regression and column (5) the results of the corresponding OLS regression. All Models include a full set of teachers, academic region and year fixed effects as well as controls for teachers' experience and seniority and for students' average age, gender, family social background, German language study and Ancient language study. Standard errors (in parentheses) are clustered at the teacher level. * p<0.10, ** p<0.05.

Table 7: Math teacher evaluation and student outcomes: school-level analysis

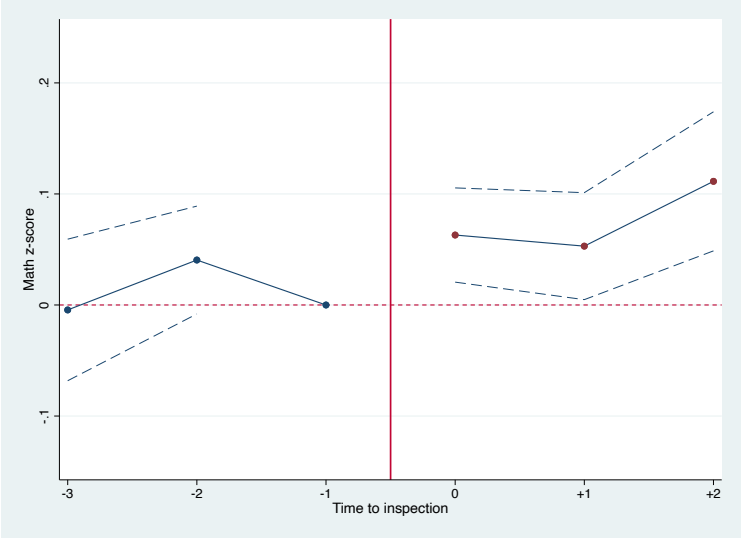
	(1) All schools	(2) Priority schools	(3) Regular schools
Math score	0.043**	0.107**	0.025
	(0.017)	(0.038)	(0.019)
Science as major field	0.0041*	0.0045	0.0042
	(0.0023)	(0.0042)	(0.0027)
Graduated in Science	0.0035*	0.0056	0.0034
	(0.0021)	(0.0037)	(0.0024)
Observations	19889	3683	16206

Note: The first row of the table shows the results of regressing the school average performance in math at end-of-9th grade national exams taken in t on the proportion of math teachers in this school who underwent an evaluation between t_0=2008-2009 and t. The second (third) row of the table shows the results of regressing the school proportion of students who will choose science as a major field (graduate in science) in high school one year (three years) later on the proportion of math teachers in this school who underwent an evaluation between t_0=2008-2009 and t. Standard errors are in parentheses. * p<0.10, ** p<0.05.

Main Graphs



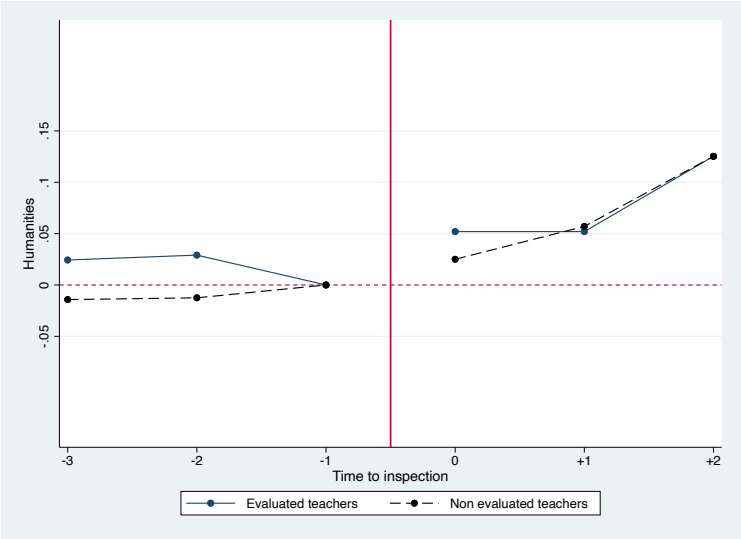
(a)



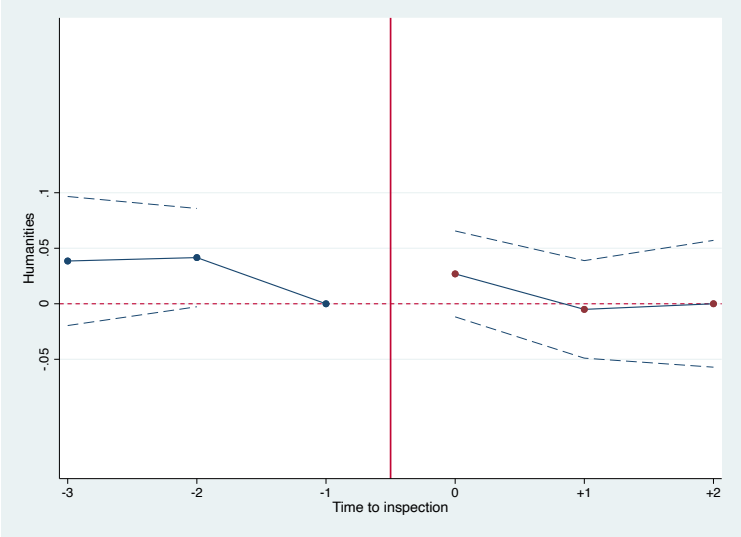
(b)

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 95% confidence intervals.



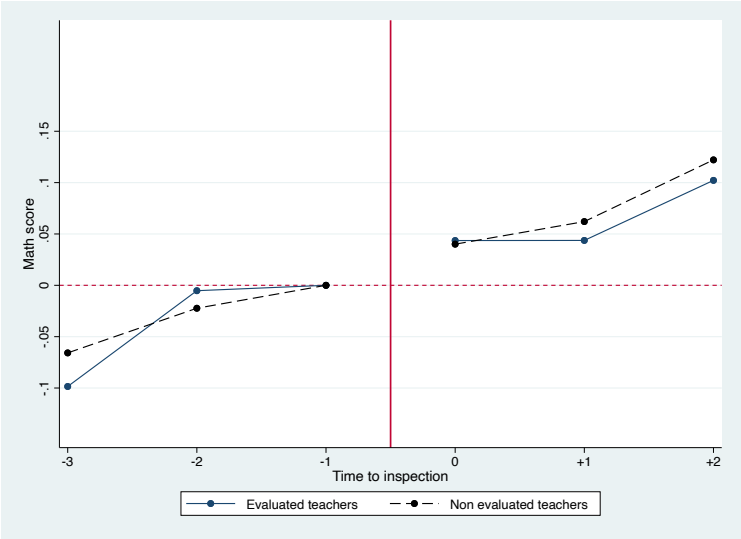
(a)



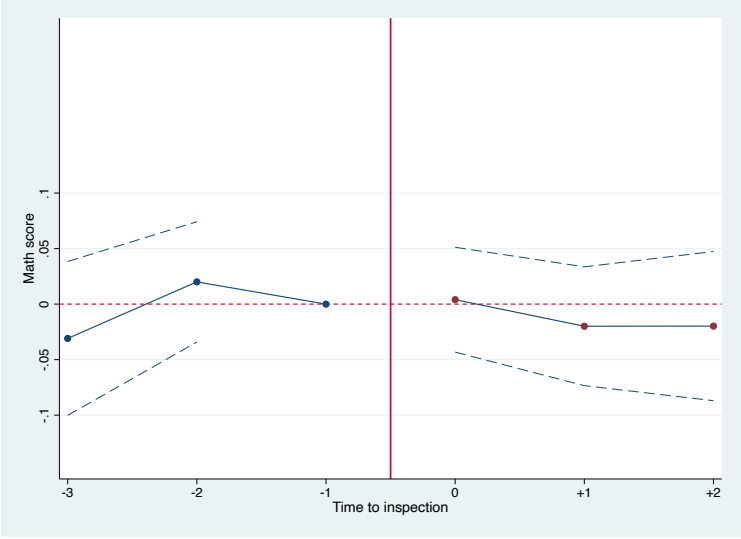
(b)

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 95% 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 95% confidence intervals.

Appendix A - Additional Tables and Graphs

Descriptive statistics

Table A1: Teacher promotion on the wage scale, by promotion track

Level	Gross annual wage in euros (2008)	Total number of years of teaching experience needed to reach the level		
		Regular track	Fast track (*Choix*)	Very fast track (*Grand Choix*)
1	19,141			
2	20,622	0.25	-	-
3	21,664	1	-	-
4	22,816	2	-	-
5	24,078	4.5	4.5	4
6	25,613	8	7.5	6.5
7	27,149	11.5	10.5	9
8	29,124	15	13.5	11.5
9	31,098	19.5	17.5	14
10	33,566	24.5	21.5	17
11	36,089	30	26	20

Note: The table shows teachers' gross annual wage in euros in 2008 for each possible position on the wage scale as well as the total number of years of teaching experience needed to reach each level by promotion track. The regular track corresponds to automatic promotion based on experience. The 30% of teachers who get the best evaluation ratings can access the very fast track (Grand Choix) while the next 50% best evaluated teachers can access the fast track (Choix). Source: *Décret n72-581 du 4 juillet 1972 relatif au statut particulier des professeurs certifiés.*

Table A2: *Inspecteurs*' characteristics

	(1) Math	(2) French language
Inspecteurs' individual characteristics		
Age	51.40	53.24
	(7.47)	(7.20)
Experience as *inspecteur*	6.32	7.07
	(3.98)	(4.36)
Female	0.34	0.58
	(0.47)	(0.49)
Total nb of *inspecteurs*	135	157
Regional characteristics		
Nb of *inspecteurs* per region	5.19	6.04
	(2.3)	(2.9)
Nb of teachers per region	2361	3101
	(1070)	(1421)
Nb of evaluations per region	346	414
	(136)	(139)
Total nb of regions	26	26

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 A3: Student characteristics - difference between priority and non priority schools

	Priority schools (1)	Non priority schools (2)	Difference (1) - (2)
Age	14.63	14.47	0.16**
	(0.23)	(0.17)	(0.01)
Female	0.51	0.51	-0.00
	(0.10)	(0.09)	(0.00)
Low-income	0.43	0.21	0.22**
	(0.19)	(0.13)	(0.01)
Average standardized test scores	-0.64	0.22	-0.86**
	(0.88)	(0.74)	(0.03)
Observations	1011	4037	5048

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 A4: Teachers' characteristics

	(1) Math	(2) French language
Experience (in 2008)	12.32	12.81
	(5.11)	(5.01)
Female teacher	0.53	0.83
	(0.50)	(0.37)
Priority schools (in 2008)	0.17	0.17
	(0.37)	(0.38)
Number of evaluations (N_e)		
$N_e = 0$	0.42	0.54
	(0.49)	(0.50)
$N_e = 1$	0.57	0.45
	(0.50)	(0.50)
$N_e > 1$	0.01	0.01
	(0.09)	(0.08)
Observations	29156	29507

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.

Table A5: Math teachers' evaluations and 9th grade teaching

	(1) All	(2) Female	(3) Male	(4) Low-exp	(5) High-exp	(6) Low-perf	(7) High-perf	(8) Priority	(9) Non Prior.
	0.009	0.013	0.004	0.010	0.007	0.003	0.015*	0.012	0.010
	(0.006)	(0.009)	(0.009)	(0.009)	(0.009)	(0.011)	(0.008)	(0.015)	(0.007)
	[0.79]	[0.78]	[0.79]	[0.76]	[0.81]	[0.78]	[0.82]	[0.77]	[0.79]
Obs.	38039	20139	17900	19283	18756	12820	24078	8418	29621

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), column (5) to teachers above this median, column (6) to teachers whose students' average performance in t_0 falls in the bottom 30% of the distribution, column (7) to teachers in the top 70% of this distribution. Eventually, columns (8) and (9) 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 (in parentheses) are clustered at the teacher level. Sample means of the dependent variables are within square brackets. * p<0.10, ** p<0.05.

Balancing tests - tables and graphs

Table A6: Balancing test - 9th grade math teacher evaluation and student characteristics

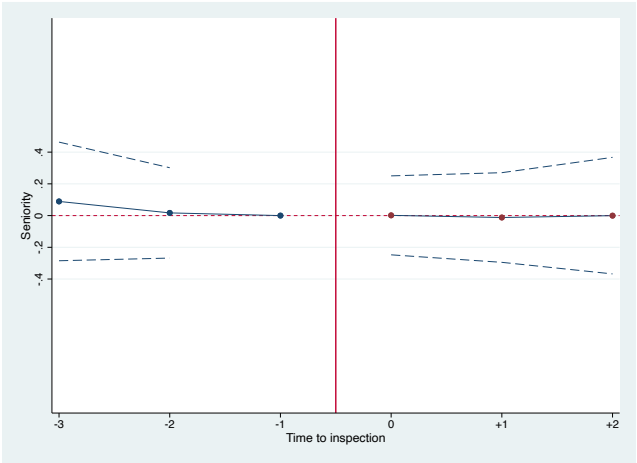
	(1) Age	(2) Female	(3) Low-income	(4) German	(5) Latin/Greek
All teachers (N=29156)	0.004	-0.001	0.002	0.000	0.002
	(0.004)	(0.003)	(0.003)	(0.004)	(0.003)
Female teachers (N=15318)	0.010*	-0.005	0.005	-0.004	0.005
	(0.006)	(0.004)	(0.004)	(0.005)	(0.005)
Male teachers (N=13838)	-0.002	0.002	-0.001	0.005	0.000
	(0.007)	(0.004)	(0.004)	(0.005)	(0.005)
Low-experience teachers (N=14319)	0.005	0.001	0.003	-0.003	0.002
	(0.007)	(0.004)	(0.004)	(0.005)	(0.005)
High-experience teachers (N=14837)	0.003	-0.003	0.001	0.003	0.003
	(0.006)	(0.004)	(0.004)	(0.005)	(0.005)
Low perf (N=9651)	0.008	0.003	0.003	-0.003	0.003
	(0.009)	(0.005)	(0.005)	(0.006)	(0.006)
High perf (N=19212)	0.002	-0.003	0.001	0.002	0.002
	(0.005)	(0.004)	(0.003)	(0.005)	(0.004)
Priority schools (N=6265)	0.010	-0.013**	0.008	0.000	-0.006
	(0.010)	(0.006)	(0.007)	(0.008)	(0.007)
Non Priority schools (N=22891)	0.003	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 9 refer to subsamples defined by teachers' gender, by teachers' number of years of experience (above or below 11 years), teachers' performance in t_0 (bottom 30% vs. top 70%) or by type of school attended (priority vs non-priority). Standard errors (in parentheses) are clustered at the teacher level. * p<0.10, ** p<0.05.

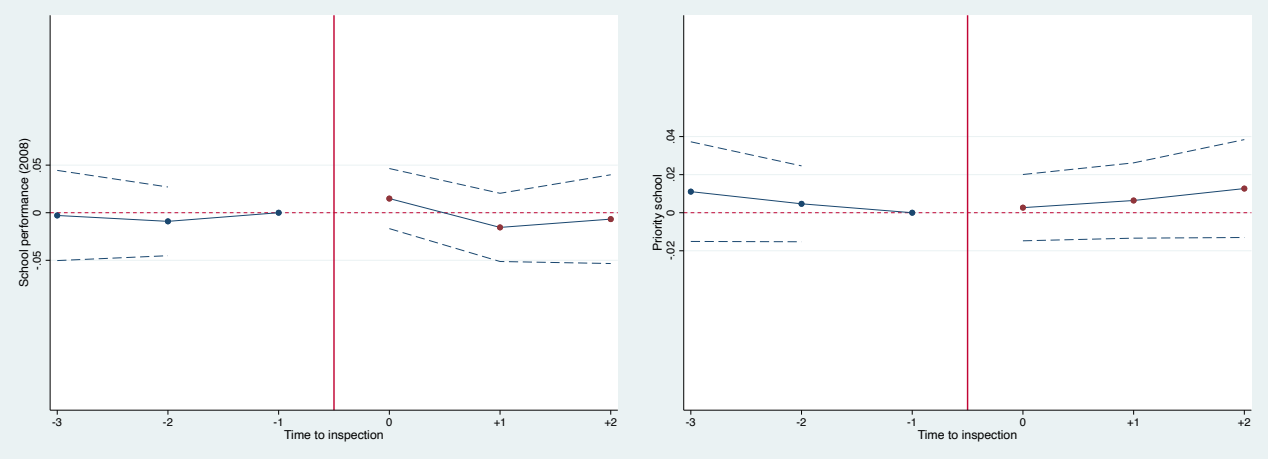
Table A7: Balancing test - 9th grade math teacher evaluation, teacher mobility and colleagues' characteristics

	(1) Teacher seniority	(2) Priority schools	(3) School performance	(4) Colleagues' experience	(5) Colleagues' seniority
All teachers (N=29156)	0.033	0.004	-0.002	-0.027	0.008
	(0.033)	(0.003)	(0.005)	(0.096)	(0.088)
Female teachers (N=15318)	0.060	0.002	-0.003	-0.120	-0.129
	(0.044)	(0.004)	(0.006)	(0.132)	(0.120)
Male teachers (N=13838)	-0.008	0.003	0.001	0.083	0.174
	(0.049)	(0.004)	(0.007)	(0.140)	(0.129)
Low-exp (N=14319)	0.077**	0.007	-0.007	-0.085	0.022
	(0.037)	(0.005)	(0.008)	(0.136)	(0.122)
High-exp (N=14837)	-0.010	-0.000	0.004	0.025	-0.002
	(0.053)	(0.003)	(0.005)	(0.137)	(0.127)
Low-perf (N=9651)	0.065	0.006	-0.001	-0.087	-0.058
	(0.071)	(0.006)	(0.011)	(0.164)	(0.149)
High-perf (N=19212)	0.021	0.002	-0.002	-0.001	0.040
	(0.034)	(0.003)	(0.004)	(0.120)	(0.110)
Priority schools (N=6265)	0.107	0.007	-0.004	-0.053	0.189
	(0.094)	(0.010)	(0.016)	(0.209)	(0.187)
Non priority schools (N=22891)	0.018	0.003*	-0.002	-0.005	-0.040
	(0.032)	(0.002)	(0.004)	(0.109)	(0.100)

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 (in parentheses) are clustered at the teacher level. * 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 95% confidence intervals.

Robustness check - Goodman-Bacon Decomposition

Table A8: Robustness check - Goodman-Bacon Decomposition

	(1) DD coeff	(2) weights
Overall DD coefficient	.039**	-
	(0.016)	
Decomposition		
Timing groups	0.038	0.324
Treated vs Untreated groups	0.048	0.658
Within residual	-0.248	0.019
Observations	17828	17828

Note: This Table shows the average effects and weights for the two basic types of diff-in-diff (DD) variations used in this paper, namely those that compare treated and never treated teachers and those that compare groups of teachers treated at different point in time, using a Goodman-Bacon (2018) decomposition. The Table refers to the subsample of teachers who are observed at all periods between t_0=2008 and t_1=2011. Models include a full set of teachers, academic region and year fixed effects as well as controls for teachers' experience and seniority and for students' average age, gender, family social background, German language study and Ancient language study. Standard errors (in parentheses) are clustered at the teacher level. * p<0.10, ** p<0.05.

French language teacher evaluation and student performance

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) Low-perf	(7) High-perf	(8) Priority	(9) Non Prior.
Reading scores	0.023	0.028*	0.000	0.015	0.028	0.014	0.012	0.096**	0.000
	(0.014)	(0.016)	(0.036)	(0.021)	(0.019)	(0.027)	(0.016)	(0.034)	(0.016)
Writing scores	0.039**	0.048**	0.009	0.054**	0.023	0.058*	0.020	0.114**	0.017
	(0.018)	(0.020)	(0.046)	(0.027)	(0.025)	(0.034)	(0.021)	(0.044)	(0.020)
Observations	29507	24624	4883	13331	16176	8726	19754	6479	23028

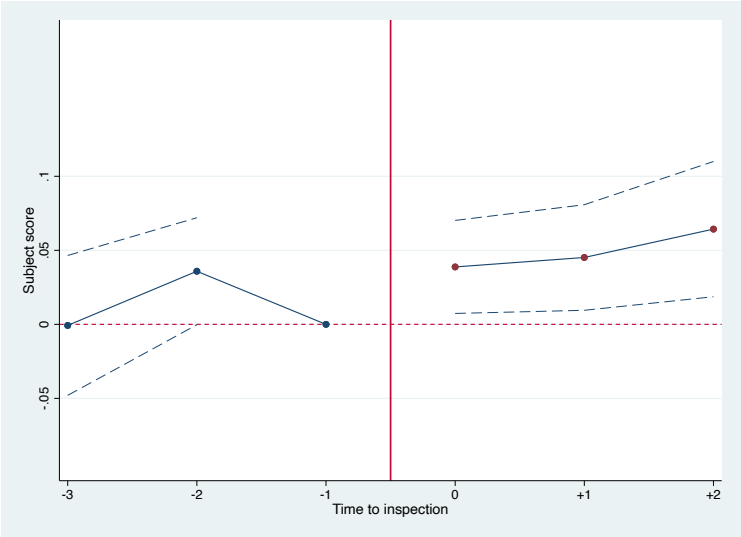
Note: The table refers to our working sample of French language teachers who teach 9th grade students between t_0=2008-2009 and t_1=2011-2012. The first (second) row shows the results of regressing their students' average score in reading (writing) at the end of year t on a dummy indicating that they underwent an external evaluation between t_0 and t. 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 in t_0 (i.e., above or below 11 years), columns (6) and (7) to the subsamples of teachers defined by whether their students' average performance in French language in t_0 falls (or not) in the bottom 30% of the distribution, columns (8) and (9) to the subsample of teachers who were in education priority schools in t_0 and the subsample who were in non-priority schools. Models include a full set of teachers, academic region and year fixed effects as well as controls for teachers' experience and seniority and for students' average age, gender, family social background, German language study and Ancient language study. Standard errors (in parentheses) are clustered at the teacher level. * p<0.10, ** p<0.05.

Math and French language teachers' external evaluations and student performance

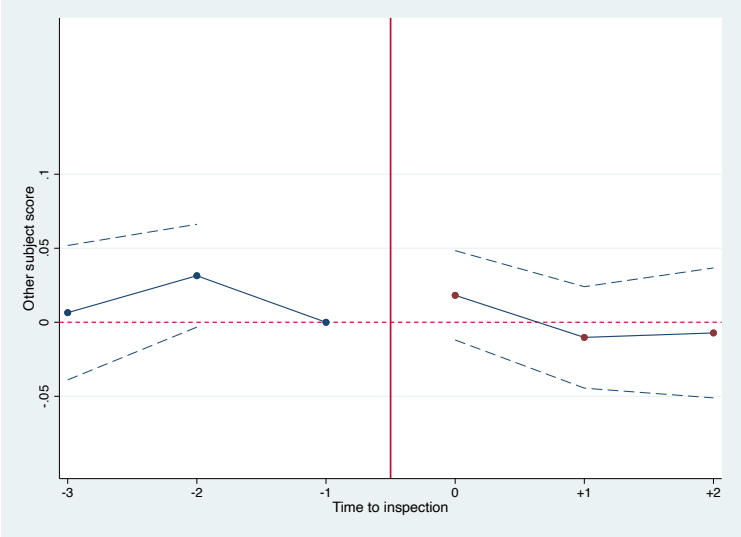
Table A10: Math and French language teachers' evaluations and student performance - by subgroups

	(1) All	(2) Female	(3) Male	(4) Low-exp	(5) High-exp	(6) Low-perf	(7) High-perf	(8) Priority	(9) Non Prior.
Score in the subject	0.040** (0.010)	0.038** (0.012)	0.047** (0.018)	0.045** (0.015)	0.036** (0.014)	0.051** (0.018)	0.021* (0.012)	0.102** (0.023)	0.022** (0.011)
Score in other subjects	0.010 (0.010)	0.004 (0.012)	0.022 (0.018)	0.011 (0.015)	0.009 (0.014)	0.020 (0.019)	-0.004 (0.012)	0.011 (0.023)	0.010 (0.011)
Observations	58657	39938	18719	27647	31010	18374	38963	12741	45916

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) to (9) refer to subsamples defined by teachers' gender, number of years of teaching experience in t_0 (above/below 11 years), performance in t_0 (as proxied by their students' average performance in t_0), and type of school attended in t_0 (priority/non priority). Models include a full set of teachers, academic region and year fixed effects as well as controls for teachers' experience and seniority and for students' average age, gender, family social background, German language study and Ancient language study. Standard errors (in parentheses) are clustered at the teacher level. * p<0.10, ** p<0.05.



(a)



(b)

Figure A2: Math and French language teacher evaluation and student performance

Note: The solid line in Figure A2 (a) shows the difference in test scores between students of evaluated and non-evaluated math and French language teachers before and after evaluations, in the subject taught by the teacher. The solid line in Figure A2 (b) shows the same difference with student test scores in subjects not taught by the teacher. The dotted lines show 95% confidence intervals.

Predicted timing of external evaluations

Table A11: Distribution of between-evaluation spacing, by education region

(1) Region	(2) N	(3) mode	(4) % mode	(5) % mode +/- 1 year	(6) % < 4 years
1	374	4	0.76	0.78	0.07
2	159	4	0.33	0.50	0.01
3	262	5	0.46	0.67	0.05
4	278	5	0.55	0.72	0.01
5	232	5	0.31	0.56	0.01
6	347	5	0.32	0.41	0.09
7	281	5	0.84	0.84	0.01
8	397	5	0.40	0.69	0.11
9	444	5	0.29	0.43	0.10
10	303	5	0.31	0.57	0.01
11	366	5	0.58	0.65	0.05
12	69	5	0.41	0.64	0.00
13	529	5	0.45	0.66	0.06
14	231	5	0.57	0.71	0.00
15	368	5	0.32	0.53	0.03
16	492	5	0.46	0.74	0.07
17	367	6	0.43	0.69	0.09
18	269	6	0.23	0.39	0.04
19	388	6	0.24	0.47	0.05
20	170	6	0.87	0.67	0.02
21	75	7	0.42	0.43	0.02
22	343	7	0.48	0.51	0.02
23	723	7	0.23	0.50	0.09
24	588	8	0.18	0.22	0.09
25	625	8	0.23	0.30	0.01
26	301	9	0.20	0.30	0.04

Note: For each mainland education region j (with j=1 to 26), this table shows the main features of the distribution of the number of years elapsed since the previous external evaluation for math teachers who were evaluated in 2008 and had been evaluated at least once before. Column (2) shows the number of observations, column (3) shows the local modal value of the distribution, column (4) shows the proportion of observations that correspond to the modal value, column (5) shows the proportion of observations that fall in the interval [modal value – 1 year; modal value + 1 year], column (6) shows the proportion of evaluations that occur less than 4 years after the previous one. To ensure anonymity of regions, the number displayed in column (1) doesn't correspond to any official classification.

Table A12: Testing instrument exogeneity using students' characteristics

	(1) Age	(2) Female	(3) Low-income	(4) German	(5) Latin/Greek
Regions with exact timing (N=1967)	0.003	-0.000	0.005	-0.002	0.015
	(0.015)	(0.012)	(0.011)	(0.020)	(0.019)
All regions (N=26253)	0.002	-0.004	0.002	-0.007	-0.005
	(0.005)	(0.004)	(0.003)	(0.005)	(0.004)

Note: the table shows the results of regressing 9th grade classes' average characteristics on the instrument Z_{it} as defined in section 5, namely on a dummy indicating that the time elapsed between the teacher's last evaluation and t is greater than $k(r)$, with $k(r)$ being the minimal spacing preceding a 10% increase in the probability of being evaluated in region r. The first row refers to the three education regions with exact timing (in which case $k(r)$ coincides with the local modal value of between-evaluation spacing) whereas the second row refers to the full set of mainland education regions. Standard errors (in parentheses) are clustered at the teacher level. * p<0.10, ** p<0.05.

Table A13: Testing instrument exogeneity using teachers' characteristics

	(1) Teacher seniority	(2) Priority schools	(3) School performance	(4) Colleagues' experience	(5) Colleagues' seniority
Regions with exact timing (N=1967)	0.108	0.010	0.003	-0.330	-0.136
	(0.153)	(0.012)	(0.017)	(0.447)	(0.424)
All regions (N=26253)	-0.038	0.004	-0.005	-0.024	0.029
	(0.044)	(0.003)	(0.006)	(0.122)	(0.110)

Note: the table shows the results of regressing teacher seniority, school characteristics (priority school, school performance) and colleagues' characteristics (experience, seniority) on the instrument Z_{it} as defined in section 5, namely on a dummy indicating that the time elapsed between the teacher's last evaluation and t is greater than $k(r)$, with $k(r)$ being the minimal spacing preceding a 10% increase in the probability of being evaluated in region r. The first row refers to the three education regions with exact timing (in which case $k(r)$ coincides with the local modal value of between-evaluation spacing) whereas the second row refers to the full set of mainland education regions. 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 (in parentheses) are clustered at the teacher level. * p<0.10, ** p<0.05.

School-level analysis

Table A14: Balancing test - 9th grade math teacher evaluation and student characteristics, school level

	(1) Age	(2) Female	(3) Low-income	(4) German	(5) Latin/Greek
All schools (N=19934)	0.003	-0.006	0.007	0.004	-0.001
	(0.008)	(0.006)	(0.006)	(0.005)	(0.005)
Priority schools (N=3691)	0.003	-0.026*	0.014	0.001	0.007
	(0.020)	(0.015)	(0.015)	(0.012)	(0.012)
Regular schools (N=16243)	0.005	-0.003	0.008	0.004	-0.004
	(0.008)	(0.007)	(0.006)	(0.006)	(0.005)

Note: the table shows the results of regressing 9th grade students' school 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 the school proportion of math teachers who underwent an evaluation between t_0=2008-2009 and t. The first row refers to the full working sample, whereas rows 2 and 3 refer to subsamples defined by type of school (priority vs non-priority). 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[26].

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.

[26]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 - Additional 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) Low-perf	(7) High-perf	(8) Priority	(9) Non Prior.
Math	0.042**	0.030	0.057**	0.053**	0.035*	0.073**	0.019	0.079**	0.032**
	(0.014)	(0.019)	(0.020)	(0.020)	(0.019)	(0.023)	(0.016)	(0.030)	(0.015)
Humanities	0.009	-0.005	0.025	0.017	0.004	0.008	0.005	0.007	0.012
	(0.013)	(0.018)	(0.020)	(0.020)	(0.018)	(0.025)	(0.015)	(0.032)	(0.015)
Observations	31102	16492	14610	14319	16783	10221	20576	6475	24627

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. 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 in t_0 (i.e., above or below 11 years), columns (6) and (7) to the subsamples of teachers defined by whether their students' performance in math in t_0 falls (or not) in the bottom 30% of the distribution, columns (8) and (9) to the subsample of teachers who were in education priority schools in t_0 and the subsample who were in non-priority schools. Models include a full set of teachers, academic region and year fixed effects as well as controls for teachers' experience and seniority and for students' average age, gender, family social background, German language study and Ancient language study. Standard errors (in parentheses) are clustered at the teacher level. * 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) Low-perf	(7) High-perf	(8) Priority	(9) Non Prior.
Science as	0.004**	0.000	0.008**	0.007**	0.001	0.005*	0.003	0.008**	0.003
major field	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.002)	(0.004)	(0.002)
	[0.180]	[0.188]	[0.172]	[0.163]	[0.195]	[0.177]	[0.183]	[0.125]	[0.195]
Graduation	0.004**	0.001	0.008**	0.009**	0.001	0.005	0.003	0.008**	0.003
in science	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.004)	(0.002)
	[0.153]	[0.159]	[0.145]	[0.136]	[0.167]	[0.149]	[0.155]	[0.101]	[0.166]
Observations	31102	16492	14610	14319	16783	10221	20576	6475	24627

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 (9) refer to subsamples defined by teachers' gender, number of years of teaching experience in t_0 (above/below 11 years), performance in t_0 (as proxied by their students' average performance in math in t_0), and type of school attended in t_0 (priority/non priority). Models include a full set of teachers, academic region and year fixed effects as well as controls for teachers' experience and seniority and 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.