# There's always room for improvement: the persistent benefits of a large-scale teacher evaluation system*

Simon Briole & Éric Maurin†

*Paris School of Economics, France*

July 6, 2021

**Abstract**

In France, secondary school teachers are evaluated every five to six years by senior experts from the Ministry of Education. These evaluations involve the supervision of one class session, a debriefing interview and the writing of an official evaluation report. Their results are used to determine teachers' career advancement. We show that these repeated evaluations help improve teacher effectiveness (as measured by their students' performance) at all stages of their career. The impact on student performance is particularly strong in priority education schools and remains significant several years after students leave middle school. Evaluators' feedback likely plays a key role in improving teacher effectiveness.

JEL classification: I20; I28; J24

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

*We would like to thank Sandra Black, Clément de Chaisemartin, Manon Garrouste, Marc Gurgand, Élise Huillery, Sandra McNally and Léa Palet for helpful comments on previous versions of this paper as well as participants of the IZA Economics of Education workshop 2020 and 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. Finally, 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

Evaluating employees' performance represents a major challenge for employers - both public and private - particularly in the education sector. The question is how to design evaluation systems that help employees develop their skills, reward their efforts and increase their motivation. The challenge is particularly crucial in the field of education, since teacher effectiveness is a key determinant of student performance and future human capital. Much recent work has attempted to shed light on this question and to identify the teacher evaluation systems most likely to lead to improvements in student performance.[1] One issue, however, is that most of this work is based on small-scale interventions, often with volunteer schools. This makes it very difficult to predict the impact of scaling up to all schools and teachers. This is of particular concern because it has long been established that the effects of small experimental programs often decrease considerably in size when these programs are implemented on a larger scale (Al-Ubaydli et al. (2017a,b); Banerjee et al. (2017)).

This paper overcomes this issue by investigating the effects of a low intensity, large scale evaluation program, namely the national system of teacher evaluation, which was progressively introduced in France over the nineteenth century. We build 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 in teachers' careers and their results have a direct impact on teachers' wages and promotion opportunities, as teachers with the highest 30% of evaluation scores are promoted through a fast track while the lowest 20% remain on a slow track. In this context, it is possible to compare student performance in the years just before and after teacher evaluations to test whether evaluations are followed by improved performance in national exams. Our data also allow us to track students over time, which enables us to explore whether teacher evaluations have a long-term impact on students' subsequent school career.

We first provide evidence that a math teacher's external evaluation is followed by a significant increase (of about 4.2% of a SD) in his or her students' math scores in national exams taken at the end of middle school (9th grade). The effect of math teachers' evaluations is observed for achievement in math but not in other subjects. This is consistent with the hypothesis that the increase in math performance is driven

[1] For evidence on the the importance of teacher quality see for example Rockoff (2004), Rivkin et al. (2005), Hanushek & Rivkin (2006), Aaronson et al. (2007), Hanushek & Rivkin (2010) or Chetty et al. (2014). For recent evidence on the impact of teacher evaluation see Taylor & Tyler (2012), Murphy et al. (2018) or Burgess et al. (2019).

by improved teaching practices of math teachers, not by changes in overall student quality or increased math workload (which would be detrimental to performance in other subjects). Furthermore, the increase in math teachers' effectiveness is observed not only at the end of the evaluation year, but also at the end of subsequent years. Such persistent effects on teacher effectiveness is consistent with the hypothesis that the visit of an evaluator is associated with an improvement in teachers' pedagogical skills, and not just a temporary increase in teachers' effort. Likewise, the influence of math teachers' evaluations on their students can still be seen several years later in high-school, as a greater proportion of their former students continue to study and successfully graduate in fields of study which involve taking math exams. These longer term effects on student outcomes give further suggestive evidence that external evaluations do not simply help math teachers "teach to the test"; rather they enable teachers to improve students' core skills as well as students' perception of the discipline.

Further analysis reveals that the effect of evaluations is even more significant for math teachers working in priority education schools (the 25% most deprived), in contexts where students' academic level is often very weak and teaching is more challenging. Also, the effect of an additional evaluation is about as strong for more experienced teachers as for less experienced ones. Finally, we extend our analysis to French language teachers. We find that evaluations have smaller effects on them 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 writing test scores (which likely capture the most advanced skills) but little effect on dictation/comprehension test scores.

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 the classes they have to teach. The fact that external evaluations in mathematics do not coincide with significant improvement in students' performance in other subjects (and vice versa) is consistent with our identification assumption. We have also checked that external evaluations do not coincide with teachers changing schools or with changes in the likelihood of teaching in priority education schools. Nor do they coincide with changes in the socio-demographic and academic characteristics of the classes to which teachers are assigned.

To further test the robustness of our results, we develop 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 student-to-teacher assignment within school. This approach confirms that variation in the proportion of teachers recently evaluated in a given school is followed by a parallel variation

in students' average math performance in that same school. In the end, we obtain school-level estimates that are very similar to teacher-level estimates.

Our paper contributes to the growing literature on the causal impact of policies aimed at improving teacher effectiveness. These include pay-for-performance programs that provide financial incentives to teachers, linked to their students' test scores or to the quality of their teaching practices (see for example Lavy (2009); Springer et al. (2010); Neal (2011); Fryer (2013); Dee & Wyckoff (2015); Lavy (2020)). They also include peer mentoring programs 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 programs 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 (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 find 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 early-career teachers.

In this context, one of the key features of the evaluation system studied in this paper is that it has been implemented at scale for a long time and covers all teachers (regardless of experience) and schools. Another important feature is that it combines a relative incentive scheme with an observation/feedback procedure conducted by an external evaluator with hierarchical authority. Our first contribution is to show that such a low-intensity, large-scale program is able to improve teachers' effectiveness at all stages of their career and in all school contexts. This is key, given the recurring difficulty of scaling up and sustaining policies that have proven to be cost-effective in local and experimental settings. Another contribution of the paper is to show that the evaluation of teacher practices can have an impact not only on their students' performance, but also on these students' educational choices in later years. Finally, taking advantage of our large sample size, we are able to document the heterogeneity of impacts across teachers and school contexts. The average impact of evaluations on student performance in math and French language appears to be significantly stronger in the most deprived areas. This result suggests that repeated external evaluations may be a simple way to reduce the achievement gap across school contexts.

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 effect 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 effect of evaluations. Building on a school-level approach, section 5 provides further evidence on the robustness of our results. Section 6 discusses the potential mechanisms driving the effect of evaluations. The final section concludes with a brief discussion on the implications of our results.

# 1 Institutional context

### Secondary school teachers

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.[2] Once recruited, they cannot be dismissed unless they are guilty of serious misconduct (e.g. physical abuse of students), which however happens very rarely. All teachers are paid according to a unique wage scale that is set at the national level, but the relative speed at which they move up the scale depends on the evaluation ratings they get all along their careers. External evaluations conducted every five to six years by senior experts of the Ministry of Education to assess teachers' pedagogical skills determine 60% of the total evaluation rating, while the remaining 40% is determined by internal evaluations conducted each year by school heads to assess teachers' general behavior (punctuality, attendance, participation in the life of the school, etc.).

The 30% of teachers who get the best evaluation ratings can access a fast promotion track (called *Grand Choix*), the next 50% best evaluated teachers are promoted through a regular track (*Choix*) and the 20% of teachers with the lowest ratings are promoted through a slow track, which corresponds to the minimal promotion speed based on experience. While evaluations cannot lead to formal sanctions (teachers cannot be dismissed or demoted due to bad ratings), promotion through the slow track is reserved for the teachers that evaluators have identified as the least competent. In practice, going from the first to the last level of the wage scale takes about 30 years through the slow track versus only 20 years through the fast track. After 20 years of teaching experience, teachers who always got bad evaluation ratings and were promoted solely based on experience (slow track) earn about 31,000 euros gross per year, while teachers who always got the best evaluation ratings and were promoted through the fastest track

[2]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 student performance on end-of-middle school exams, we will focus on CAPES recipients.

earn about 36,000 euros.[3] Generally speaking, the system under consideration rewards teachers' relative performance. This type of relative incentive scheme is difficult to manipulate (Neal (2011)) and has been shown to be effective in small controlled experiments (Lavy (2009, 2020)), but never studied on a large scale.

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 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*), who are teachers' hierarchical superiors. *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 one of the 26 education region by a centralized assignment system. 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 about 5 math *inspecteurs* per region and each one of them is responsible for about 450 math teachers (Table A2).

The external evaluation of individual teachers is the core business of *inspecteurs*, whose official mission is not only to monitor the quality of instruction provided by secondary school teachers, but also to act as advisors to the teachers for whom they are responsible.[4] Existing qualitative evidence supports the view that *inspecteurs* are making a real effort to support pedagogical innovations and help teachers improve their teaching practices, which they consider to be one of the most important parts of their work (Palet (2019)). In practice, *inspecteurs* perform about 60 evaluations per year on average.[5] As a consequence,

[3]Table A1 in the appendix provides further details on wage level and promotion speed at every stage of the career depending on evaluation results.

[4]BO, Note de service N°2005-089 du 17-6-2005 (https://www.education.gouv.fr/bo/2005/25/MEND0501225N.htm).

[5]About 350 math teachers are evaluated each year, in each region IGEN (2011). While the vast majority of evaluations are conducted by *inspecteurs* themselves, a small fraction is conducted by senior teachers temporarily appointed to help *inspecteurs*, who typically intend to take the exam to become *inspecteurs*. According to the same data, the proportion of external evaluations who are not conducted by *inspecteurs* vary across regions, but is never above 15%. Assuming that 85% of evaluations are conducted by *inspecteurs*, it means that each *inspecteur* conducts on average about 60 evaluations per year.

they develop an evaluation expertise and are recognized as pedagogical specialists by teachers, who trust their judgement and value their feedback (Albanel (2012)).

Although teacher evaluation represents the most important task assigned to *inspecteurs*, they are also in charge of many other aspects of the education policy, so that teacher evaluation is only one part of their activities. As a matter of fact, *inspecteurs* are also in charge of the design of the many national exams organized each year in France.[6] In each education region, *inspecteurs* also have to contribute to the conception and organization of teacher training and professional development programs. As regards human resources management, they are also expected to play a consulting role with teachers, namely they are expected to answer queries about both career advancement and teaching practices. More generally, *inspecteurs* are expected to supervise the actual enforcement of education policies in each education region and each school. Overall, according to surveys conducted by the Ministry of Education on the working condition of *inspecteurs*, the evaluation of teachers represents on average only between 20% and 30% of *inspecteurs*' activities (IGEN (2011); IGEN/IGAENR (2016)).[7]

**Timing of external evaluations**

*Inspecteurs* are required to evaluate teachers once at the very beginning of their careers. Thereafter, the spacing between teachers' 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 close intervals or, conversely, no longer evaluating her at all. Appendix Table A3 shows the distribution of between-evaluation intervals in the different regions and shows that these intervals are very rarely less than 4 years, but very often between 5 and 7 years, confirming that the intervals between visits are generally both constrained and difficult to predict exactly.[8]

Finally, 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 in year $t$ (CNESCO (2015)). In this context, it is unlikely that principals can take evaluations into account when deciding on class composition

[6] Most notably, they are in charge of the different types of end-of high school *Baccalauréat*, as well as the different types of end-of-middle school *Brevet*, the different *Certificat d'Aptitudes Professionnelles*, etc. However, it is important to note that *inspecteurs* do not mark these exams.

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

[8] As discussed in the last section of the paper, the effect of evaluations remains significant regardless of whether we focus on the regions in which the spacing between evaluation is the most predictable or in the other regions.

or on the assignment of teachers to individual classes.[9] In the next section, we provide ample empirical evidence to confirm that teachers are not assigned students with higher academic potential in the years in which they are evaluated. Furthermore, we also check that our main results are robust to a school-level approach that accounts for any potential within-school student sorting in the last section of the paper.

**Content of external evaluations**

For each evaluation, *inspecteurs* must follow a strict legal protocol that is defined at the national level.[10] The first mandatory step of the protocol consists in the supervision of one class session, during which *inspecteurs* evaluate teachers' (i) ability to organize the lesson, (ii) content-knowledge, (iii) capacity to engage students in the lesson, (iv) ability to adapt teaching to students' level and (v) use of various pedagogical resources. As part of this first step, *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 evaluating teachers on the basis of a single class observation.

The second phase of teacher evaluations consists in a debriefing interview with the evaluated teacher, which happens immediately after *inspecteurs*' classroom observation and typically lasts one hour. This is a key step of the evaluation process, during which *inspecteurs* provide feedback and advice. In practice, while the discussion primarily focuses on the observed lesson, *inspecteurs* and teachers can also exchange on broader teaching difficulties teachers may face or on their career prospects (Albanel (2012)).

To finalize the evaluation process, *inspecteurs* must produce a written report (the so-called *rapport d'inspection*) that is sent to both the evaluated teachers and the central education authority. This report includes the overall rating assigned to the teacher and provides a detailed justification of this rating based on classroom observation and interview with the teacher. It also formalizes the advice and feedback from the inspecteur to the teacher, which can include suggestions for specific training sessions that the teacher could attend to improve teaching or classroom management practices. During his investigations, Albanel (2012) had access to a sample of about 450 *rapport d'inspections*: 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 advice to correct problems).

[9] Even assuming that principals can predict the exact years of evaluations and manipulate class assignments so that evaluated teachers have the best classes in those years, it should be noted that it would be impossible for them to keep newly assessed teachers in the best classes in subsequent years, as these classes would have to be permanently reallocated to other teachers about to be assessed.

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

In general, teachers are notified a few weeks before the visit of the *inspecteur*. However, there is no legal constraint on notification delays.

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

An important feature of the French system is that students stay in the same class, in all subjects, (with the same teacher in each subject), throughout the school year. Classes are groups of about 25 students which represent, each year, very distinct entities. School principals assign students and teachers to classes before the beginning of the school year. In the remainder of this paper, we will mostly focus on teachers who teach 9th grade classes and our most basic measure of their effectiveness will be defined by the average performance of their students on 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. The content of the tests is set each year at the national level. In each education region, all tests are anonymized and graded by teachers from a different school than the student. Our analysis will draw on test scores standardized each year in each region using the distribution of all students taking the test.

After 9th grade, students enter high school, which runs from 10th grade to 12th grade. At the end of the first year of high school (10th grade), French students can either follow a general education or enroll in a technical or vocational education program. Those who follow general education must 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)

[11] 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.

[12] As shown in table A4 in online appendix A, the proportion of students from low-income families is twice bigger in priority education schools than in non-priority schools. Priority education 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.

is field S, which is also 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 choices, particularly on the probability of enrolling in the S field. While the first year of high school (10th grade) is devoted to exploring different subjects and choosing a field of specialization, the last two years of high school (11th and 12th grade) are devoted to preparing for the national high school leaving examination, the *Baccalauréat*, which is a prerequisite for entry into higher education. Students are required to take one exam per subject, and they graduate if their weighted average grade across subjects is 10/20 or higher, where the subjects taken and the weights are highly dependent on their field of specialization. A final 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, particularly on students' ability to graduate in the S field.

## 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 2008-2009 and 2011-2012. For each teacher $j$, this dataset gives information on whether (and when) $j$ underwent an external evaluation between 2008-2009 and 2011-2012. It also gives information on whether (and when) teacher $j$ taught 9th grade students and on the average performance of these students on exams taken at the end of 9th grade as well as on 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 2008-2009[13] and who taught 9th grade students in 2008-2009 as well as one additional time after 2008-2009. This working sample includes 9,053 math teachers (i.e., about 75% of the total number of math teachers) and represents 29,156 (teacher $\times$ year) observations in total.

We provide some descriptive statistics in online Appendix A (see column (1) of Table A5)). Most of our empirical analysis will be conducted on this working sample. About half of teachers in this sample

[13]We drop the small fraction of 9th grade teachers who are evaluated in 2008-2009 because the vast majority (about 96%) are not (re)evaluated before 2011-2012 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 2008-2009) 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 2008-2009 and 2011-2012. 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).

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.[14]

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 2008-2009, 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 2008-2009 (N=10,140) and we tested whether the probability to teach 9th grade students in year $t$ after 2008-2009 is different for teachers who are evaluated between 2008-2009 and $t$ and for those who are not evaluated in this time interval. As shown in online Appendix Table A6, we find no significant difference between the two groups of teachers. The probability to teach 9th grade students in a given year after 2008-2009 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 or teacher gender.[15] 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 next 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

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

[15] Also, we do not find any evidence that teachers' behavior is influenced by the grade they got during evaluations taken between 2008-2009 and $t$. In particular, we do not find evidence that teachers who obtain relatively weak grade at evaluations taken between 2008-2009 and $t$ are less likely to teach 9th grade students in $t$.

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 in 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.[16] 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 $e$ of evaluated math teachers defined by the year $t_e$ of their evaluation (with 2008-2009$< t_e \leq$2011-2012), let us consider $Y_{ed}$ the average performance in math of their 9th grade students on national exams taken at the end of year $t_e + d$ and $Y_{-ed}$ the average performance of the students of non-evaluated teachers at the end of the same year $t_e + d$. Denoting $Y_d$ and $Y_{-d}$ the average of $Y_{ed}$ and $Y_{-ed}$ across all possible evaluation year $t_e$, Figure 1(a) shows the evolution of $Y_d$ and $Y_{-d}$ when $d$ increases from $d$=-3 to $d$=+2 (i.e., the range of variation of $d$ in our sample). The figure reveals a marked increase in the average performance of students of evaluated teachers just after evaluations (i.e, for $d \geq 0$). The average performance of the evaluated and non-evaluated groups increases smoothly in the periods before and after the evaluation (reflecting returns to years of experience), but the gap between the two groups widens discontinuously just after the evaluation.

To take one step further, Figure 1(b) shows the estimated differences between evaluated and non-evaluated groups building on a two-way fixed effect model and using the last pre-evaluation year as a reference.[17] The figure 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. It should be noted that we obtain almost exactly the same figure when we limit ourselves to representing the basic differences between the two curves in Figure 1(a).

[16] A very small share of teachers in our sample (<1%) is evaluated twice during the period of observation (Table A5 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).

[17] To be specific, estimated differences in Figure 1(b) are obtained from regressing the average performance of the students of teacher $j$ in year $t$ on a full set of interactions between a dummy indicating whether teacher $j$ was evaluated between 2008 and 2011 and dummies indicating the number of years ($k$) between $t$ and the date of evaluation of teacher $j$ (with $k$=-3,...,3), controlling for a full set of teacher and year fixed effects, as well as for students' average characteristics (gender, age, family background, study of ancient languages, study of German Language) and teachers' number of years of teaching experience, type of school (priority/non priority) and education region.

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 the dependent variable, where scores in humanities are defined as the average of French language and history-geography scores.[18] These figures do not reveal any improvement in student performance in humanities after external evaluations of math teachers, in line with the assumption that external evaluations do not coincide with an overall improvement in the ability of students assigned to teachers. They are also consistent with the assumption that the increase in math performance is 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 performance is affected by the evaluation of non-math teachers. Figures 3(a) and 3(b) show that this is not the case, as these figures show no improvement in student math performance after the evaluation of French language teachers, which further suggests 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 particular, these figures show that external evaluations do not coincide with any change in teachers' probability to teach in priority education 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 on national exams taken in 2008, pre-treatment).

## 4 The effect of teacher evaluations: regression analysis

The previous section provides us with simple graphical evidence on the effect of external evaluations on the effectiveness of math teachers, as measured by their students' performance on external examinations. In this section, we explore the robustness of this finding - as well as the potential heterogeneity of effect

[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.

across teachers and schools - using more parsimonious regression models. Specifically, we keep on focusing on the same working sample of math teachers as in Figure 1(a) and we consider the following basic two-way fixed effects model:

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

where $Y_{jt}$ still represents the average standardized math score of teacher $j$'s students on exams taken at the end of year $t$, while $T_{jt}$ is a dummy indicating that an evaluation took place between 2008-2009 and $t$. Variable $X_{jt}$ represents a set of controls describing the average characteristics of students taught by teacher $j$ in year $t$ (gender, age, family background, study of ancient languages, study of German language). Variable $X_{jt}$ also includes teacher $j$'s time-varying characteristics, namely dummies controlling for each possible number of years of teaching experience in general and in the current school, a dummy indicating whether the teacher works in a priority education school and dummies indicating the education region. Finally, the $u_j$ and $\gamma_t$ parameters represent a comprehensive set of teacher and year fixed effects while $\epsilon_{jt}$ represent unobserved determinants of student performance. In all regressions, error terms are clustered at the teacher level.

In this set-up, parameter $\beta$ can be interpreted as the effect of one additional external evaluation between year 2008-2009 and year $t$ on student performance at the end of year $t$. 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 2008-2009 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 2008-2009 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 student performance in math (as captured by changes in $\epsilon_{jt}$), namely the same identifying assumption as in the previous graphical analysis. This amounts to assuming that the change in effectiveness of evaluated and non-evaluated teachers would have been the same over the period under consideration, had the evaluated teachers not been evaluated. In essence, by comparing evaluated and non-evaluated teachers around evaluation dates, we identify how evaluations contribute to increasing teacher effectiveness beyond what

[19] It would be interesting to estimate the effect of evaluations occurring at different stages of the school year. Unfortunately, the exact month of evaluation is not available in our dataset.

is observed when only the accumulation of years of experience plays a role.[20] Tables A7 and A8 in the online appendix respectively show 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 $T_{jt}$ and on $T_{1jt}$ and $T_{2jt}$, using model (1). While these characteristics are strong predictors of students' math ability, the tables show that none of them are related to the timing of external evaluations, consistent with our identifying assumption. We have no information on students' pre-treatment test scores (8th grade scores for instance) and cannot provide a direct test that the treatment does not coincide with changes in this specific baseline variable. It should be noted, however, that, taken together, the observed baseline characteristics explain more than 40% of the variance in the average math scores of the groups of students assigned to the different teachers. Furthermore, the set of observed baseline characteristics includes the study of ancient language and the study of German language which are strongly correlated with test scores and represent the only characteristics which distribution across classes can be manipulated by French principals (they are not allowed to group students into classes based on their test scores). Given that the treatment doesn't coincide with any change in these baseline variables, it seems unlikely that it could coincide with systematic variations in baseline test score. 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.[21] These results hold true regardless of whether we use the full sample of math teachers or subsamples defined by teacher gender, level of experience, or type of schools. Tables A9 and A10 in the online appendix also confirm 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 tables also reveal 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 this was the case, evaluations would also mechanically coincide with assignment to classes with more senior and experienced colleagues.

[20] It should be noted that even if evaluations were undertaken exactly (say) every five years, the model would still separately identify the effect of years of experience and the effect of teaching evaluations by looking for specific discontinuities in the evolution of individual effectiveness every five years.

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

### 4.1 Main effect on math scores

The upper panel of Table 1 shows the basic effect of one additional evaluation on math teachers' effectiveness, as measured by their students' performance in math on end-of-middle school national exams. In line with our graphical analysis, coefficients presented in the first row of the table confirm that external evaluations are followed by a significant improvement in math score of about 4.2% of a SD. The second and third rows of the table show the results of 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. While the difference between the two effects is non-significant at standard level, the effect on exams taken at the end of the following years tends to be stronger (6.3% vs 3.4% of a SD), consistent with the fact that evaluations occur during the course of the year and can modify teaching practices for only part of the evaluation year. Columns 3 to 6 further show that these estimates are largely insensitive to the inclusion of teacher or student characteristics, which is consistent with our identification assumption. The lower part of the table replicates this regression analysis using student performance in humanities as the dependent variable. Again, consistent with our identification assumption, the table shows no effect of math teacher external evaluations on their students' performance in humanities, be it measured at the end of the evaluation year or later.[22]

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 for example 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 A11 in Appendix A focuses on the subsample of teachers observed in our working sample each year between 2008-2009 and 2011-2012 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. Reassuringly, we obtain very

[22] 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.

similar estimates for the two sources of identification and both weights are positive. This result is also consistent with the assumption that the effect of the treatment depends little on the date of treatment.

### 4.2 Heterogeneous effect

Table 2 shows the results of replicating our basic analysis separately on subsamples of math teachers defined by their gender, number of years of teaching experience (less than 11 years vs 11 years of more, where 11 is the median number of years of experience in our sample), or type of school (priority education 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 or lower level of work experience. By contrast, the impact appears to be significantly stronger for teachers in priority education 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.[23]

Consistent with our identifying assumption, Table 2 also confirms that external evaluations of math teachers have no significant effect on student performance in humanities, regardless of the subsample. As mentioned above, Tables A7 to A10 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. Finally, Table C3 in the online appendix further show that estimates are not sensitive to the inclusion of student characteristics as controls in the regressions.

### 4.3 Longer term effect

The previous sections suggest that external evaluations improve the effectiveness of math teachers, as measured by their ability to prepare their 9th grade students for end-of-middle school exams. However, it could be the case that evaluations only help teachers prepare their students for the very specific context of the end-of-middle school exams and not improve their overall math skills, which is often referred to as the "teaching to the test" effect. In this case, the benefits of external evaluations would fade over time and students would not perform better in math after 9th grade.

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

Table 3 shows that the influence of math teachers' evaluations 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 and 2 and looks at the probability that their students will subsequently choose science as their major field of study as well as at the probability that they will subsequently graduate 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 represents 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 priority education schools (+8%).[24] These longer term effects on students' choices and performance are suggestive that external evaluations do not simply help teachers "teach to the test", but make them able to improve students' core skills as well as students' perception of the discipline.

### 4.4 The effect 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 2008-2009, who are not evaluated in 2008-2009 and who have less than 25 years of teaching experience in 2008-2009. Tables 4 and 5 replicate Tables 1 and 2 using this working sample of French language teachers. These tables suggest that French language teachers' external evaluations are followed by improvements in student performance that are somewhat weaker than for math teachers, except in priority education schools, where evaluations drive a 11% of a SD increase in student performance in French language.[25] Consistent with the results obtained with math teachers, Table 4 also shows that the average effect of French language teachers' external evaluations on their students' performance in French language is weaker for exams taken at the end of the evaluation year (about 2% of a SD) than for exams taken in subsequent years (4% of a SD). Again, this is consistent with the assumption that evaluations drive an evolution in pedagogical practices from which students cannot immediately benefit over a full year.

[24] Again, Table C4 in the online appendix further show that estimates are not sensitive to the inclusion of student characteristics as controls in the regressions.

[25] It can also be noted that the effect is 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 for example 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 A12 in online appendix A).[26] 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.

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 for example 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.

### 4.5 Cost-benefit analysis

We replicate our main graphical and regression analyses 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 Tables A13 and A14 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 student performance in the subject taught by the teacher (but has no effect on performance in the other subjects), and that the effect is stronger on exams taken in years strictly after the evaluation year. This analysis also confirms that the average effect of teacher evaluations on student performance is significantly stronger in priority education school (10% of a SD) than in non-priority ones (2% of a SD).

[26]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.

By way of comparison, the teacher evaluation program implemented in Cincinnati generates an effect in mathematics on the order of 11% of a SD in the years following the evaluation (Taylor & Tyler (2012)), an impact about twice as high as the impact of the visit of an *inspecteur* on math performance. The costs involved in the Cincinnati program, however, are on the order of 7,500 dollars per evaluation, i.e., much higher than those involved by the evaluation program studied in this paper.

Specifically, as already pointed out, the evaluation of teachers represents on average between 20% and 30% of *inspecteurs*' activities. 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.

It would be interesting to further compare the cost-effectiveness of the system studied in this paper with those of lower intensity programs based on mentoring or teacher peer observation, like those studied by Burgess et al. (2019) or Papay et al. (2020). This exercise is made very difficult by the fact that these small-scale programs involve specific selections of disadvantaged schools, where effects tend to be more significant. In addition, the costs involved in these light touch interventions are mainly opportunity costs and are difficult to estimate.

Finally, we can point out that a teacher evaluation program like the one studied in this paper seems much more cost-effective than traditional class size reduction policies. As it happens, our estimated effect of teacher evaluation on student performance is of the same order of magnitude as the effect of a 5-student reduction in class size in French middle-schools (Piketty & Valdenaire (2006)). But reducing class size by 5 students involves costs that are out of proportion to those associated with the visit of an *inspecteur*. Specifically, 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 $0.20 \times 50{,}000$ euros, where 50,000 euros is a proxy for the total labor cost of a secondary school teacher. This results in a cost of about 10,000 euros per year, to be compared with a cost of less than 500 euros for an *inspecteur*'s evaluation, to be paid only every 5 or 6 years.

# 5 A 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 $\overline{Y}_{st}$ 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:

$$\overline{Y}_{st} = a\overline{T}_{st} + b\overline{X}_{st} + u_s + v_t + e_{st} \quad (2)$$

where $\overline{T}_{st}$ represents the cumulative proportion of 9th grade math teachers in school $s$ in year $t$ who benefited from an external evaluation between 2008-2009 and $t$, while $\overline{X}_{st}$ 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 year-on-year variations in $\overline{T}_{st}$) 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 A15 shows the result of regressing our different school level controls $\overline{X}_{st}$ on $\overline{T}_{st}$. These balancing checks confirm that there is no significant variation in students' average characteristics across years that correspond to different values of $\overline{T}_{st}$, consistent with our identifying assumption.

Table 6 first shows the results of our school-level regressions when we use the average performance on the end-of-middle school exams as the dependent variable. We find that the evaluation of 9th grade math teachers in a school is associated with a 4.3% 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 priority education 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 Discussion on potential mechanisms

In the previous sections, we provided evidence that the combination of a relative incentive scheme with a procedure of observing and evaluating teaching practices is able to sustainably improve teachers' effectiveness. Two main mechanisms may explain this effect. First, teachers may benefit from the feedback of *inspecteurs*, who have a long teaching experience and represent a pedagogical authority.[27] This may be particularly the case for teachers who are most in difficulty with their class. Second, the financial and symbolic stakes of *inspecteurs*' evaluations may encourage teachers to better prepare their courses and update their pedagogical practices. Teachers are notified of the visit a few weeks in advance and can use this time to best prepare the materials they will show the *inspecteur* and the content of the lesson they will teach during the visit.

During his or her visit, the *inspecteur* has access to the class book, where the detailed content of the lessons since the beginning of the year must be recorded, as well as the homework done and the assessments to which the students have been subjected (and the marks awarded). The *inspecteur* may ask to see some of the students' notebooks and graded tests. The teacher is judged in part on the accuracy and consistency of what is recorded in the notebook, as well as on the quality of the assessments, their appropriateness to the course content and to the level of the students (Albanel (2012)). To prepare for the *inspecteur*'s visit, teachers may try to improve the class book and make it as readable as possible. However, the teacher cannot change the content of past lessons or their chronology. Nor can she change the content or frequency of the tests she has given her students.

During the visit, the *inspecteur* also attends a lesson given by the teacher. The teacher is then judged for her performance in front of the students, the quality and clarity of her language, as well as her ability to capture the students' attention and make them active during the lesson. The teacher is also judged for her ability to build a session that is consistent with the national curriculum, with clearly stated objectives, a progression over the course of the lesson adapted to the level of the students and possibly a short summative evaluation at the end of the session. Finally, in addition to the meeting with the teacher, the *inspecteur* evaluates the quality of the teacher's involvement in the projects implemented at the school level with the help of the school principal.

In anticipation of the *inspecteur*'s visit, teachers usually make a special effort to prepare the lesson they will give in front of her/him. They can also warn their students and try to get them to be more disciplined and to participate more in the lesson on the day of the inspection (although this may not be

[27] Several recent studies support the idea that providing feedback to teachers can improve their efficiency (Taylor & Tyler (2012), Steinberg & Sartain (2015), Hussain (2015), Garet et al. (2017), Burgess et al. (2019), Papay et al. (2020)).

easy with classes of 25-30 teenagers). However, if there are problems with the way they talk to students or manage a classroom, they are unlikely to be able to solve the problems on their own in the days leading up to the *inspecteur*'s visit. More generally, if the inspection resulted in only a temporary and superficial increase in teachers' efforts before and during the *inspecteur*'s visit, the effect of the inspection would not persist in subsequent years.[28]

The persistence of the effect suggests that it is the consequence of deeper changes that only the *inspecteurs*' feedback is able to bring about. As a matter of fact, *inspecteurs*' feedback can be useful in several different ways. They can help teachers establish clear rules of conduct in the classroom, better manage their speech and that of their students, and construct more balanced lessons that are neither too lectured nor too unstructured. They can also help them find the balance between a friendly approach (to build confidence) and a self-control that inspires respect. Also, *inspecteurs* can help teachers improve the quality of the tests they use to assess their students. Designing good assessments is not easy and can have a large impact on student learning (see for example Black & Wiliam (1998); Muijs & Reynolds (2017)). This is an important area where the support and guidance of experienced teachers (as *inspecteurs* were before they became *inspecteurs*) can be very helpful.

Given the spacing between evaluations, the probability of being evaluated by the same *inspecteur* twice in a row is low. Under these conditions, one may wonder what could lead teachers to take their *inspecteur*'s feedback into account, since it is likely that they will no longer have to deal with him or her. One explanation is that the next *inspecteur* will necessarily have access to the official evaluation report written after each evaluation. If this report highlights weaknesses, it is these weaknesses that the new *inspecteur* will pay more attention to. If the teacher does not want to be badly graded on the next evaluation, then he or she had better follow the *inspecteur*'s key advice. As we measure it, the effect of evaluations may not be the same if there were no reports and if evaluations were not repeated over time.

Regardless of incentives, it is also likely that many teachers are motivated agents who would take up feedback even without future evaluations from the *inspecteurs* (Dixit (2002)). In addition, *inspecteurs*' tips for better classroom management do not necessarily require much effort to implement, while in return they can make the teacher's job easier.

[28] In three regions, the variance of between-visit intervals is minimal, with more than 75% of teachers being evaluated after exactly the same number of years since the last evaluation. To further explore the role of predictability, we compared the effect of evaluations in these three regions with those in other regions. The effect is significant in both groups of regions, but tends to be stronger in the three regions where the timing is the most predictable (7.8% vs 3.6% of a SD, see online Appendix Table A16). This result seems in line with the idea that evaluations are more effective when predictable. It cannot be ruled out, however, that this difference simply reflects a difference in the quality of *inspecteurs* across regions, some of whom may be both more efficient and more attentive to respecting an identical timing for all teachers.

*Inspecteurs* form a long established group of experts who specialize in the task of evaluating teachers and who acquire specific skills over time that allow them to provide high-quality feedback to teachers (Condette (2017)). They are also teachers' hierarchical superiors, which reinforces the weight of their recommendations. The repeated nature of interactions between teachers and *inspecteurs* (even if not always with the same one) makes it all the more difficult for the former to ignore the recommendations of the latter. These specific features of our setting help understand both why many teachers likely prepare their evaluations as best they can and, above all, why teachers likely take the feedback and advice of *inspecteurs* very seriously (as suggested by Albanel (2012)), these two mechanisms combining to contribute to a significant and persistent improvement in student achievement.

# 7 Conclusion

The evaluation of employees' performance represents a difficult challenge for employers. 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 a lasting effect on teacher effectiveness and student achievement. The effect is seen not only for students taught by the teacher the year of the evaluation but also for students taught by the same teacher the subsequent years, suggesting that evaluations improve teachers' core pedagogical skills. Teacher evaluations improve students' skills in a persistent manner, with a significantly greater effect on advanced skills (math or writing) than on more basic skills (reading or spelling).

We further show that the impact of teacher evaluation on student 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. Finally, 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.

Al-Ubaydli, O., List, J. A., LoRe, D., & Suskind, D. 2017a. "Scaling for economists: Lessons from the non-adherence problem in the medical literature". *Journal of Economic Perspectives*, 31(4):125–44.

Al-Ubaydli, O., List, J. A., & Suskind, D. L. 2017b. "What can we learn from experiments? Understanding the threats to the scalability of experimental results". *American Economic Review*, 107(5):282–86.

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.

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.

Black, P. & Wiliam, D. 1998. "Assessment and classroom learning". *Assessment in Education: principles, policy & practice*, 5(1):7–74.

Burgess, S., Rawal, S., & Taylor, E. S. 2019. "Teacher peer observation and student test scores: Evidence from a field experiment in English secondary schools.". EdWorkingPaper: 19-139.

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.

CNESCO. 2015. "La constitution des classes : pratiques et enjeux". Conseil National d'Évaluation du Système Scolaire.

Condette, J.-F. 2017. *Les personnels d'inspection: contrôler, évaluer, conseiller les enseignants: retour sur une histoire: France-Europe, XVIIe-XXe siècle.* Presses universitaires de Rennes.

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.

Dixit, A. 2002. "Incentives and organizations in the public sector: An interpretative review". *Journal of Human Resources*, pages 696–727.

Feuillet, P. 2020. "Le devenir des enseignants entre la rentrée 2017 et la rentrée 2018". *Note d'information n°20.16, Avril 2020, DEPP.*

Fryer, R. G. 2013. "Teacher incentives and student achievement: Evidence from New York City public schools". *Journal of Labor Economics*, 31(2):373–407.

Garet, M. S., Wayne, A. J., Brown, S., Rickles, J., Song, M., & Manzeske, D. 2017. "The Impact of Providing Performance Feedback to Teachers and Principals. NCEE 2018-4001.". *National Center for Education Evaluation and Regional Assistance.*

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.

Hussain, I. 2015. "Subjective performance evaluation in the public sector evidence from school inspections". *Journal of Human Resources*, 50(1):189–221.

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.

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

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.

Lavy, V. 2020. "Teachers' Pay for Performance in the Long-Run: The Dynamic Pattern of Treatment Effects on Students' Educational and Labour Market Outcomes in Adulthood". *The Review of Economic Studies*, 87(5):2322–2355.

Muijs, D. & Reynolds, D. 2017. *Effective teaching: Evidence and practice*. Sage.

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.

Neal, D. 2011. "The design of performance pay in education". In *Handbook of the Economics of Education*, volume 4, pages 495–550. Elsevier.

Palet, L. 2019. "De la qualification à la compétence: la fausse piste du mérite?". *Administration Education*, (3):129–137.

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.

Springer, M. G., Hamilton, L., McCaffrey, D. F., Ballou, D., Le, V.-N., Pepper, M., Lockwood, J., & Stecher, B. M. 2010. "Teacher pay for performance: Experimental evidence from the Project on Incentives in Teaching.". *National Center on Performance Incentives.*

Steinberg, M. P. & Sartain, L. 2015. "Does teacher evaluation improve school performance? Experimental evidence from Chicago's Excellence in Teaching project". *Education Finance and Policy*, 10(4):535–572.

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.*

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

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| *Panel A : math test scores* | | | | | | |
| Evaluation | 0.042** | | 0.041** | | 0.044** | |
| | (0.016) | | (0.016) | | (0.014) | |
| Evaluation in $t$ | | 0.034** | | 0.033** | | 0.039** |
| | | (0.016) | | (0.016) | | (0.014) |
| Evaluation before $t$ | | 0.063** | | 0.059** | | 0.054** |
| | | (0.020) | | (0.020) | | (0.018) |
| *Panel B : humanities test scores* | | | | | | |
| Evaluation | 0.004 | | 0.005 | | 0.009 | |
| | (0.016) | | (0.016) | | (0.014) | |
| Evaluation in $t$ | | 0.004 | | 0.005 | | 0.010 |
| | | (0.016) | | (0.016) | | (0.014) |
| Evaluation before $t$ | | 0.001 | | 0.001 | | 0.001 |
| | | (0.021) | | (0.021) | | (0.018) |
| Teacher controls | . | . | ✓ | ✓ | ✓ | ✓ |
| Student controls | . | . | . | . | ✓ | ✓ |
| Observations | 29156 | 29156 | 29156 | 29156 | 29156 | 29156 |

*Note*: The table refers to our working sample of math teachers who teach 9th grade students between 2008-2009 and 2011-2012. The first row of the upper (lower) panel 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 2008-2009 and $t$. Rows 2 and 3 respectively show 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 2008-2009 and $t-1$. All regressions include the full set of teacher, region and year fixed effects. Columns (3) and (4) further include dummies for teachers' number of years of experience and seniority level. Columns (5) and (6) further controls 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) Priority | (6) Non Priority |
|---|---|---|---|---|---|---|
| *Math score* | 0.033* | 0.056** | 0.053** | 0.039** | 0.083** | 0.034** |
| | (0.019) | (0.020) | (0.020) | (0.019) | (0.031) | (0.015) |
| *Humanities score* | -0.004 | 0.023 | 0.017 | 0.004 | 0.006 | 0.012 |
| | (0.019) | (0.020) | (0.020) | (0.019) | (0.032) | (0.015) |
| Observations | 15318 | 13838 | 14319 | 14837 | 6265 | 22891 |

*Note*: The table refers to our working sample of math teachers who teach 9th grade students between 2008-2009 and 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 2008-2009 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 2008-2009 (i.e., above or below 11 years), columns (5) and (6) to the subsample of teachers who were in priority education schools in 2008-2009 and the subsample who were in non-priority schools. Models include a full set of teachers, 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) Priority | (7) Non Prior. |
|---|---|---|---|---|---|---|---|
| *Science as* | 0.005** | 0.002 | 0.007** | 0.009** | 0.001 | 0.008** | 0.003 |
| *major field* | (0.002) | (0.003) | (0.003) | (0.003) | (0.003) | (0.004) | (0.002) |
| | [0.178] | [0.184] | [0.171] | [0.163] | [0.192] | [0.124] | [0.192] |
| *Graduation* | 0.004** | 0.001 | 0.008** | 0.007** | 0.002 | 0.008** | 0.003 |
| *in Science* | (0.002) | (0.003) | (0.003) | (0.003) | (0.003) | (0.004) | (0.002) |
| | [0.150] | [0.156] | [0.144] | [0.136] | [0.164] | [0.100] | [0.164] |
| Observations | 29156 | 15318 | 13838 | 14319 | 14837 | 6265 | 22891 |

*Note*: The table refers to the working sample of math teachers who teach 9th grade students between 2008-2009 and 2011-2012. The first row shows the result of regressing the proportion of their 9th grade students who will choose science as their major field of study at the end of 10th grade on a dummy indicating that they underwent an external evaluation between 2008-2009 and $t$. The second row shows the result of regressing the proportion of their 9th grade students who will graduate in science at the end of 12th grade on the same independent variable. The first column refers to the full sample, whereas columns (2) to (7) refer to subsamples defined by teachers' gender, number of years of teaching experience in 2008-2009 (above/below 11 years), and type of school attended in 2008-2009 (priority/non priority). Models include a full set of teachers, 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

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| *Panel A : French language test scores* | | | | | | |
| Evaluation | 0.026 | | 0.028 | | 0.031** | |
| | (0.017) | | (0.018) | | (0.015) | |
| Evaluation in $t$ | | 0.017 | | 0.018 | | 0.020 |
| | | (0.019) | | (0.019) | | (0.016) |
| Evaluation before $t$ | | 0.040* | | 0.041* | | 0.048** |
| | | (0.022) | | (0.022) | | (0.019) |
| *Panel B : math test scores* | | | | | | |
| Evaluation | 0.009 | | 0.012 | | 0.013 | |
| | (0.017) | | (0.017) | | (0.015) | |
| Evaluation in $t$ | | 0.012 | | 0.014 | | 0.015 |
| | | (0.018) | | (0.018) | | (0.016) |
| Evaluation before $t$ | | -0.001 | | 0.004 | | 0.007 |
| | | (0.022) | | (0.022) | | (0.019) |
| Teacher controls | . | . | ✓ | ✓ | ✓ | ✓ |
| Student controls | . | . | . | . | ✓ | ✓ |
| Observations | 29507 | 29507 | 29507 | 29507 | 29507 | 29507 |

*Note*: The table refers to our working sample of French language teachers who teach 9th grade students between 2008-2009 and 2011-2012. The first row of the upper (lower) panel shows the result of regressing their students' average standardized score in French language (math) at the end of year $t$ on a dummy indicating that they underwent an external evaluation between 2008-2009 and $t$. Rows 2 and 3 respectively show 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 2008-2009 and $t-1$. All regressions include the full set of teacher, region and year fixed effects. Columns (3) and (4) further include dummies for teachers' number of years of experience and seniority level. Columns (5) and (6) further controls 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) Female | (2) Male | (3) Low-exp | (4) High-exp | (5) Priority | (6) Non Priority |
|---|---|---|---|---|---|---|
| *French lang. score* | 0.037** | 0.006 | 0.030 | 0.030 | 0.112** | 0.006 |
| | (0.016) | (0.038) | (0.022) | (0.020) | (0.035) | (0.016) |
| *Mathematics score* | 0.014 | 0.011 | 0.003 | 0.021 | 0.031 | 0.007 |
| | (0.016) | (0.039) | (0.023) | (0.020) | (0.034) | (0.017) |
| Observations | 24624 | 4883 | 13331 | 16176 | 6479 | 23028 |

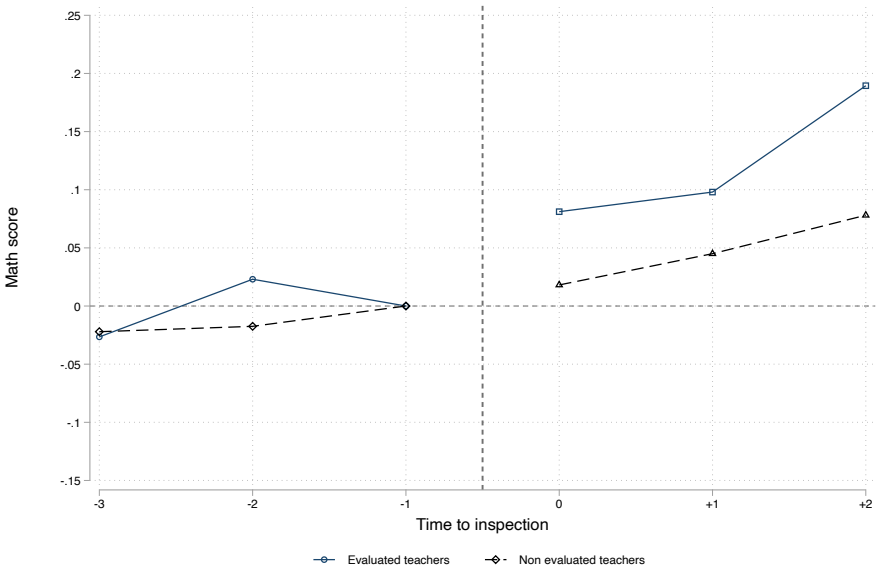
*Note*: The table refers to our working sample of French language teachers who teach 9th grade students between 2008-2009 and 2011-2012. The first (second) row shows the results of regressing their students' average score in French language (mathematics) at the end of year $t$ on a dummy indicating that they underwent an external evaluation between 2008-2009 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 2008-2009 (i.e., above or below 11 years), columns (5) and (6) to the subsample of teachers who were in priority education schools in 2008-2009 and the subsample who were in non-priority schools. Models include a full set of teachers, 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: Math teacher evaluation and student outcomes: school-level analysis

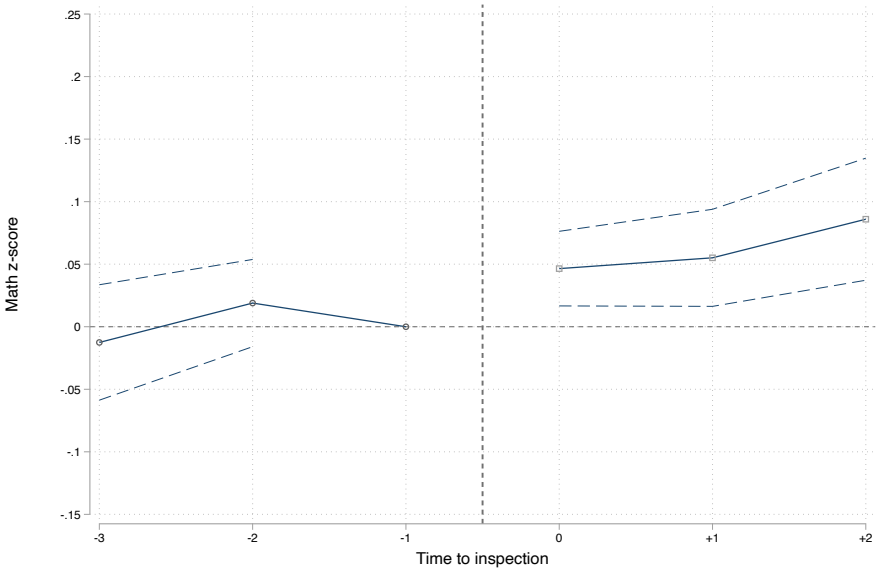
| | (1) All schools | (2) Priority schools | (3) Non priority schools |
|---|---|---|---|
| *Math score* | 0.043** | 0.107** | 0.025 |
| | (0.017) | (0.038) | (0.019) |
| *Science as major field* | 0.0041* | 0.0045 | 0.0042 |
| | (0.0023) | (0.0042) | (0.0027) |
| *Graduation in Science* | 0.0035* |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

*Note*: The table shows the results of estimating our school level model using three different dependent variables, namely 9th graders' average performance in math on the end-of-9th-grade national exams (row 1) the proportion of 9th graders who will choose science as a major field at the end of 10th grade (row 2) and the proportion of 9th graders who will graduate in science at the end of 12th grade (row 3). Standard errors are in parentheses. * p<0.10, ** p<0.05.

# Main Figures



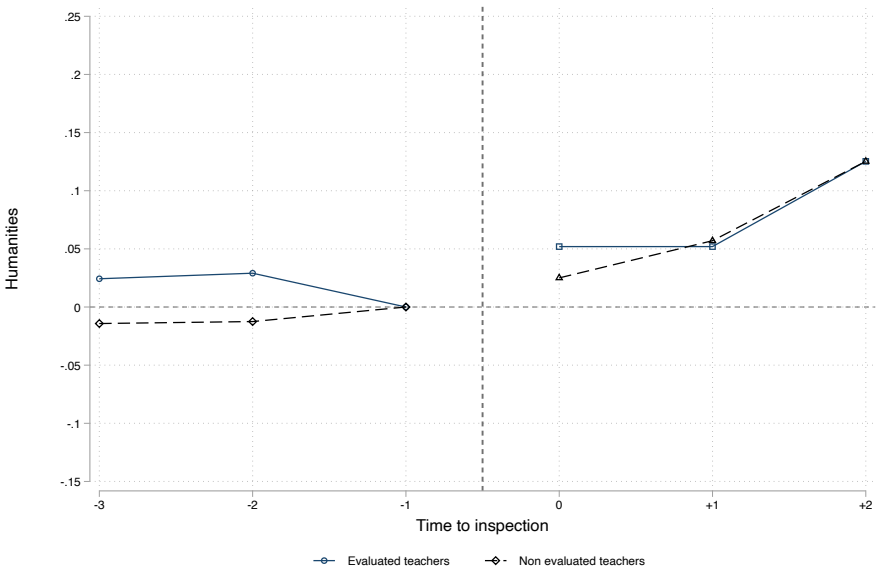

(a)



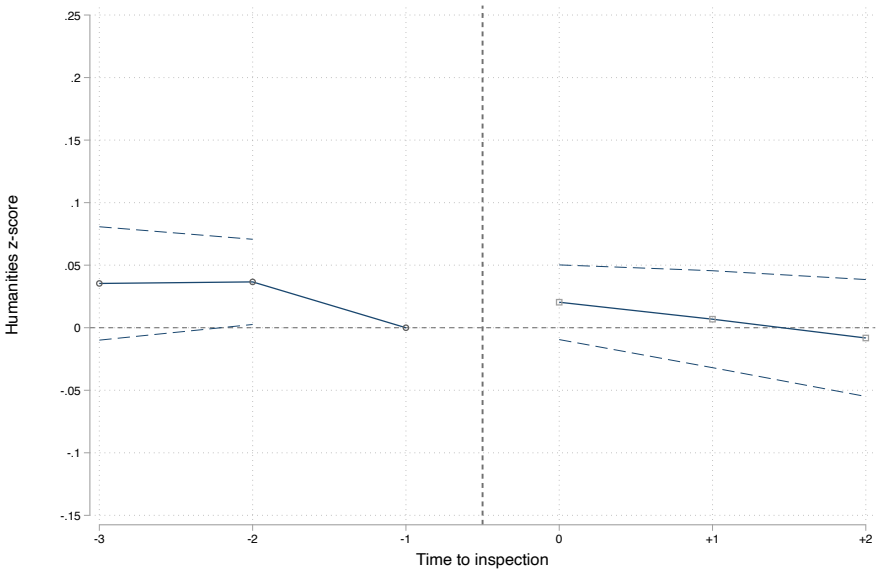

(b)

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 on exams taken in the same years. The solid line in Figure 1 (b) shows the estimated difference in math scores between students of evaluated and non-evaluated math teachers before and after evaluations. The dotted lines show 95% confidence intervals.



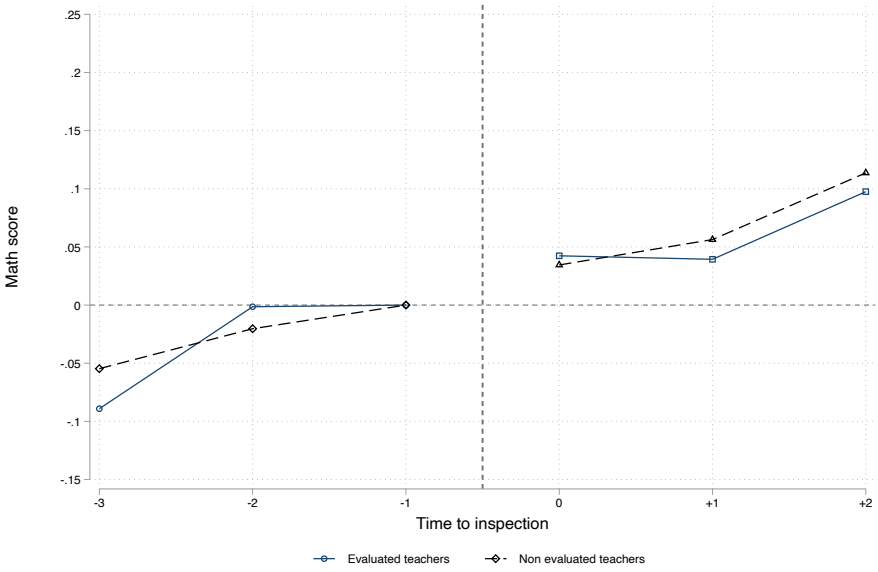

(a)



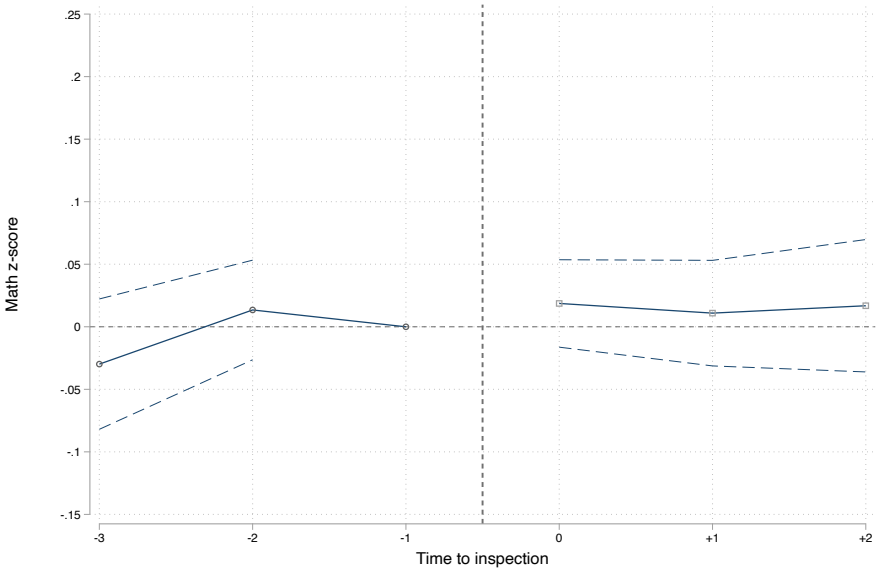

(b)

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 on exams taken in the same years. The solid line in Figure 2 (b) shows the estimated 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 on exams taken in the same years. The solid line in Figure 3 (b) shows the estimated 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 Figures

## 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 | | |
|---|---|---|---|---|
| | | Slow track | Regular track (*Choix*) | 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 30% of teachers who get the best evaluation ratings can access the fast track (Grand Choix). The next 50% best evaluated teachers are promoted through the regular track (Choix). The 20% of teachers with the lowest ratings are promoted through the slow track, which corresponds to the minimal promotion speed based on experience. Source: *Décret n°72-581 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: 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 A4: 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 on the end-of-middle school national exam, across priority and non-priority schools in 2008-2009. * p<0.10, ** p<0.05.

Table A5: 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 2008-2009 and 2011-2012. It shows teachers' average number of years of teaching experience in 2008, proportion of female, type of school in 2008 and average number of external evaluations undertaken 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 A6: Math teachers' evaluations and 9th grade teaching

| | (1) All | (2) Female | (3) Male | (4) Low-exp | (5) High-exp | (6) Priority | (7) Non Prior. |
|---|---|---|---|---|---|---|---|
| | 0.009 | 0.013 | 0.004 | 0.010 | 0.007 | 0.012 | 0.010 |
| | (0.006) | (0.009) | (0.009) | (0.009) | (0.009) | (0.015) | (0.007) |
| | [0.79] | [0.78] | [0.79] | [0.76] | [0.81] | [0.77] | [0.79] |
| Obs. | 38039 | 20139 | 17900 | 19283 | 18756 | 8418 | 29621 |

*Note*: The table refers to the sample of math teachers who teach 9th grade students in 2008-2009 and who are not evaluated during 2008-2009. It shows the result of regressing a dummy indicating that teachers teach 9th grade students in year $t$ on a dummy indicating that teachers underwent an external evaluation between 2008-2009 and $t$. Column (2) refers to the subsample of female teachers, column (3) to male teachers, column (4) and (5) to teachers whose number of years of teaching experience is below or above the median (i.e. above or below 11 years), column (6) and (7) to teachers who were in priority education 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 Figures

Table A7: 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) |
| *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 2008-2009 and $t$. The first row refers to the full working sample, whereas rows 2 to 7 refer to subsamples defined by teachers' gender, by teachers' number of years of experience (above or below 11 years), or by type of school attended (priority vs non-priority). Standard errors (in parentheses) are clustered at the teacher level. * $p<0.10$, ** $p<0.05$.

Table A8: Balancing test - 9th grade math teacher evaluation and student characteristics II

| | (1) Age | (2) Female | (3) Low-income | (4) German | (5) Latin/Greek |
|---|---|---|---|---|---|
| *All teachers* (N=29156) | | | | | |
| Evaluation in $t$ | 0.007 | 0.001 | 0.003 | 0.000 | 0.001 |
| | (0.004) | (0.003) | (0.003) | (0.004) | (0.004) |
| Evaluation before $t$ | -0.001 | -0.007* | 0.002 | 0.002 | 0.004 |
| | (0.006) | (0.004) | (0.004) | (0.005) | (0.005) |
| *Female teachers* (N=15318) | | | | | |
| Evaluation in $t$ | 0.013** | -0.003 | 0.004 | -0.002 | 0.003 |
| | (0.006) | (0.004) | (0.004) | (0.005) | (0.005) |
| Evaluation before $t$ | 0.001 | -0.012** | 0.008 | -0.007 | 0.010 |
| | (0.007) | (0.005) | (0.005) | (0.007) | (0.007) |
| *Male teachers* (N=13838) | | | | | |
| Evaluation in $t$ | -0.001 | 0.004 | 0.001 | 0.004 | 0.000 |
| | (0.007) | (0.005) | (0.005) | (0.005) | (0.005) |
| Evaluation before $t$ | -0.002 | -0.002 | -0.004 | 0.012* | -0.002 |
| | (0.010) | (0.005) | (0.005) | (0.007) | (0.007) |
| *Low-experience teachers* (N=14319) | | | | | |
| Evaluation in $t$ | 0.008 | 0.003 | 0.004 | -0.004 | 0.001 |
| | (0.007) | (0.005) | (0.004) | (0.005) | (0.005) |
| Evaluation before $t$ | -0.006 | -0.005 | -0.000 | 0.004 | 0.006 |
| | (0.010) | (0.005) | (0.005) | (0.007) | (0.007) |
| *High-experience teachers* (N=14837) | | | | | |
| Evaluation in $t$ | 0.005 | -0.001 | 0.000 | 0.005 | 0.002 |
| | (0.006) | (0.005) | (0.004) | (0.006) | (0.005) |
| Evaluation before $t$ | 0.002 | -0.008 | 0.003 | 0.001 | 0.002 |
| | (0.007) | (0.005) | (0.005) | (0.007) | (0.007) |
| *Priority schools* (N=6265) | | | | | |
| Evaluation in $t$ | 0.012 | -0.012* | 0.011 | -0.002 | -0.008 |
| | (0.010) | (0.006) | (0.007) | (0.008) | (0.007) |
| Evaluation before $t$ | 0.004 | -0.016** | 0.005 | 0.007 | 0.001 |
| | (0.013) | (0.008) | (0.009) | (0.011) | (0.010) |
| *Non Priority schools* (N=22891) | | | | | |
| Evaluation in $t$ | 0.006 | 0.004 | -0.000 | 0.001 | 0.004 |
| | (0.005) | (0.004) | (0.003) | (0.004) | (0.004) |
| Evaluation before $t$ | -0.002 | -0.005 | 0.000 | 0.001 | 0.005 |
| | (0.007) | (0.004) | (0.004) | (0.005) | (0.005) |

*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 external evaluation in $t$ and on a dummy indicating that they underwent an evaluation between 2008-2009 and $t-1$. Standard errors (in parentheses) are clustered at the teacher level. * p<0.10, ** p<0.05.

Table A9: Balancing test - 9th grade math teacher evaluation, teacher mobility and colleagues' characteristics

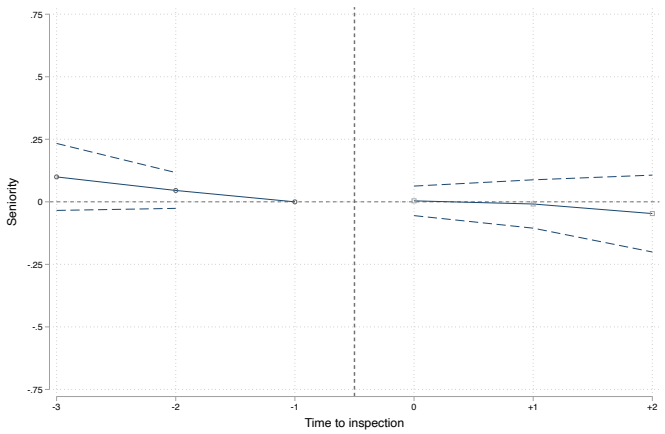
| | (1) Teacher seniority | (2) Priority schools | (3) School performance | (4) Colleagues' experience | (5) Colleagues' seniority |
|---|---|---|---|---|---|
| *All teachers* (N=29156) | 0.033 | 0.004 | -0.002 | -0.027 | 0.008 |
| | (0.033) | (0.003) | (0.005) | (0.096) | (0.088) |
| *Female teachers* (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) |
| *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 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 $t$. 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.

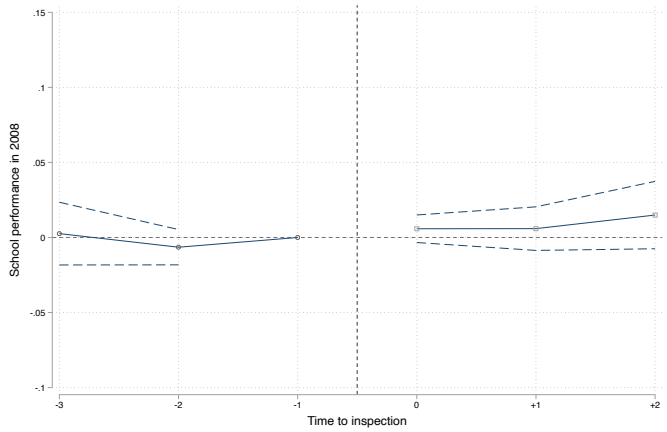
Table A10: Balancing test - 9th grade math teacher evaluation, teacher mobility and colleagues' characteristics II

| | (1) Teacher seniority | (2) Priority schools | (3) School performance | (4) Colleagues' experience | (5) Colleagues' seniority |
|---|---|---|---|---|---|
| *All teachers* (N=29156) | | | | | |
| Evaluation in $t$ | 0.022 | 0.003 | -0.002 | -0.048 | -0.008 |
| | (0.031) | (0.003) | (0.004) | (0.097) | (0.088) |
| Evaluation before $t$ | 0.045 | 0.007* | -0.008 | -0.015 | 0.014 |
| | (0.050) | (0.004) | (0.007) | (0.132) | (0.121) |
| *Female teachers* (N=15318) | | | | | |
| Evaluation in $t$ | 0.064 | 0.002 | -0.001 | -0.113 | -0.095 |
| | (0.041) | (0.003) | (0.006) | (0.134) | (0.120) |
| Evaluation before $t$ | 0.065 | 0.004 | -0.009 | -0.110 | -0.219 |
| | (0.070) | (0.005) | (0.009) | (0.185) | (0.169) |
| *Male teachers* (N=13838) | | | | | |
| Evaluation in $t$ | -0.032 | 0.002 | -0.001 | 0.034 | 0.104 |
| | (0.046) | (0.004) | (0.006) | (0.142) | (0.130) |
| Evaluation before $t$ | 0.012 | 0.008 | -0.003 | 0.098 | 0.294* |
| | (0.071) | (0.006) | (0.010) | (0.189) | (0.174) |
| *Low-exp* (N=14319) | | | | | |
| Evaluation in $t$ | 0.066* | 0.006 | -0.007 | -0.108 | 0.003 |
| | (0.035) | (0.005) | (0.007) | (0.137) | (0.121) |
| Evaluation before $t$ | 0.109* | 0.012 | -0.015 | -0.081 | 0.010 |
| | (0.057) | (0.007) | (0.012) | (0.185) | (0.169) |
| *High-exp* (N=14837) | | | | | |
| Evaluation in $t$ | -0.017 | -0.001 | 0.004 | 0.007 | -0.012 |
| | (0.049) | (0.003) | (0.005) | (0.139) | (0.129) |
| Evaluation before $t$ | -0.006 | 0.003 | 0.001 | 0.038 | 0.030 |
| | (0.079) | (0.004) | (0.007) | (0.190) | (0.173) |
| *Priority schools* (N=6265) | | | | | |
| Evaluation in $t$ | 0.065 | 0.003 | -0.002 | -0.057 | 0.154 |
| | (0.089) | (0.010) | (0.015) | (0.209) | (0.187) |
| Evaluation before $t$ | 0.229* | 0.020 | -0.013 | -0.072 | 0.360 |
| | (0.136) | (0.014) | (0.024) | (0.281) | (0.250) |
| *Non priority schools* (N=22891) | | | | | |
| Evaluation in $t$ | 0.018 | 0.003** | -0.002 | -0.032 | -0.047 |
| | (0.030) | (0.002) | (0.004) | (0.110) | (0.101) |
| Evaluation before $t$ | -0.004 | 0.002 | -0.005 | 0.011 | -0.082 |
| | (0.051) | (0.002) | (0.006) | (0.150) | (0.138) |

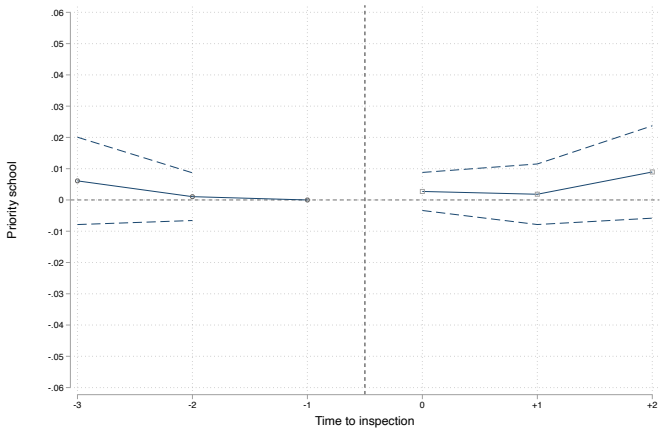
*Note*: The table shows the results of regressing teacher seniority, school characteristics (priority school, school performance) and colleagues' characteristics (experience, seniority) on a dummy indicating that the math teacher underwent an external evaluation in $t$ and on a dummy indicating that she underwent an evaluation between 2008-2009 and $t-1$. School performance in column (3) is the average math test score in 2008 of the school in which the math teacher teaches in year $t$. Finally, 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 Figures A1 (a) to A1 (c) show the estimated 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 A11: 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 2008-2009 and 2011-2012. Models include a full set of teachers, 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 A12: 9th grade French language teacher evaluation and student performance by French language subtopic test scores and by subgroups

| | (1) All | (2) Female | (3) Male | (4) Low-exp | (5) High-exp | (6) Priority | (7) Non Prior. |
|---|---|---|---|---|---|---|---|
| *Reading scores* | 0.023 | 0.028* | 0.000 | 0.015 | 0.028 | 0.096** | 0.000 |
| | (0.014) | (0.016) | (0.036) | (0.021) | (0.019) | (0.034) | (0.016) |
| *Writing scores* | 0.039** | 0.048** | 0.009 | 0.054** | 0.023 | 0.114** | 0.017 |
| | (0.018) | (0.020) | (0.046) | (0.027) | (0.025) | (0.044) | (0.020) |
| Observations | 29507 | 24624 | 4883 | 13331 | 16176 | 6479 | 23028 |

*Note*: The table refers to our working sample of French language teachers who teach 9th grade students between 2008-2009 and 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 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 2008-2009 (i.e., above or below 11 years), columns (6) and (7) to the subsample of teachers who were in priority education schools in 2008-2009 and the subsample who were in non-priority schools. Models include a full set of teachers, 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 A13: 9th grade math and French language teacher evaluation and student performance

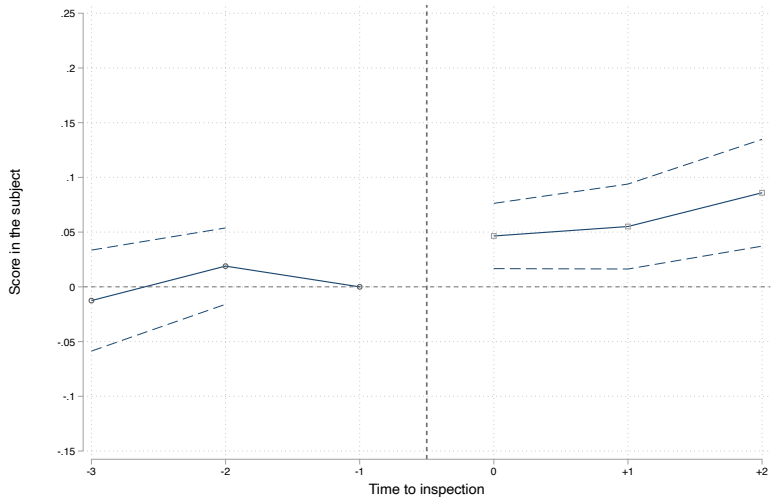
| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| *Panel A : Score in the subject* | | | | | | |
| Evaluation | 0.038** | | 0.038** | | 0.040** | |
| | (0.012) | | (0.012) | | (0.010) | |
| Evaluation in $t$ | | 0.029** | | 0.029** | | 0.032** |
| | | (0.012) | | (0.012) | | (0.011) |
| Evaluation before $t$ | | 0.057** | | 0.055** | | 0.055** |
| | | (0.015) | | (0.015) | | (0.013) |
| *Panel B : Score in other subjects* | | | | | | |
| Evaluation | 0.006 | | 0.008 | | 0.010 | |
| | (0.012) | | (0.012) | | (0.010) | |
| Evaluation in $t$ | | 0.007 | | 0.008 | | 0.011 |
| | | (0.012) | | (0.012) | | (0.011) |
| Evaluation before $t$ | | 0.001 | | 0.003 | | 0.002 |
| | | (0.015) | | (0.015) | | (0.013) |
| Teacher controls | . | . | ✓ | ✓ | ✓ | ✓ |
| Student controls | . | . | . | . | ✓ | ✓ |
| Observations | 58657 | 58657 | 58657 | 58657 | 58657 | 58657 |

*Note*: the table refers to the joint sample of math teachers and French language teachers who teach 9th grade students between 2008-2009 and 2011-2012. The first row of the upper (lower) panel shows the result of regressing their students' average standardized 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 2008-2009 and $t$. Rows 2 and 3 respectively show 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 2008-2009 and $t-1$. All regressions include the full set of teacher, region and year fixed effects. Columns (3) and (4) further include dummies for teachers' number of years of experience and seniority level. Columns (5) and (6) further controls 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.

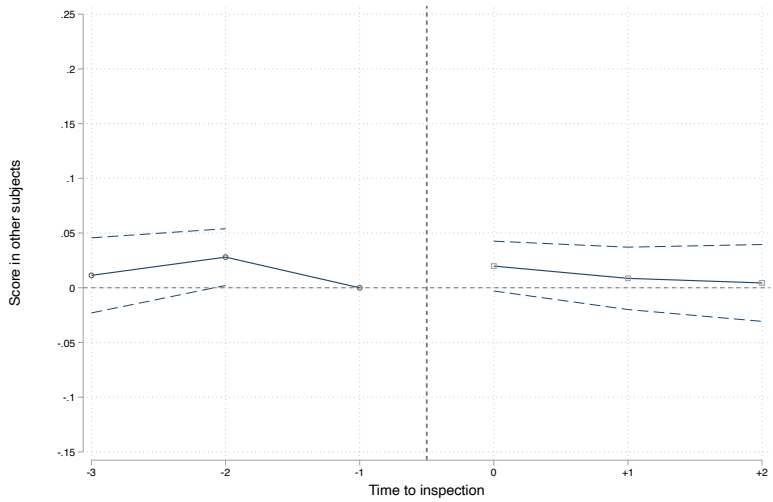
Table A14: Math and French language teachers' evaluations and student performance - by subgroups

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

*Note*: The table refers to the joint sample of math and French language teachers who teach 9th grade students between 2008-2009 and 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 2008-2009 and $t$. The first column refers to the full sample, whereas columns (2) to (7) refer to subsamples defined by teachers' gender, number of years of teaching experience in 2008-2009 (above/below 11 years), and type of school attended in 2008-2009 (priority/non priority). Models include a full set of teachers, 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 estimated 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.

## School-level analysis

Table A15: 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) |
| *Non priority 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 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.

## Between-evaluation spacing and the effect of evaluation

Table A16: 9th grade math teacher evaluation and student math performance, by group of educational regions

| | (1) | (2) |
|---|---|---|
| *Panel A : regions with exact timing (N=1997)* | | |
| Evaluation | 0.079* | |
| | (0.046) | |
| Evaluation in $t$ | | 0.081* |
| | | (0.045) |
| Evaluation before $t$ | | 0.074 |
| | | (0.063) |
| *Panel B : other regions (N=27159)* | | |
| Evaluation | 0.036** | |
| | (0.015) | |
| Evaluation in $t$ | | 0.031** |
| | | (0.015) |
| Evaluation before $t$ | | 0.049** |
| | | (0.019) |
| Teacher controls | ✓ | ✓ |
| Student controls | ✓ | ✓ |

*Note*: The table shows the result of regressing 9th grade math teachers students' average standardized score in math at the end of year $t$ on a dummy indicating that they underwent an external evaluation between 2008-2009 and $t$. Rows 2 and 3 respectively show 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 2008-2009 and $t-1$. The upper panel of the table refers to the 3 educational regions where the variance in between-evaluation spacing is minimal, and the lower panel refers to all other regions. All regressions include the full set of teacher, region and year fixed effects, controls for teachers' experience and seniority level, as well as controls 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.

## Appendix B - Data construction

This paper uses an administrative database with detailed information on secondary school teachers for the period between 2008-2009 and 2011-2012. For each teacher $j$, this dataset provides information on whether (and when) $j$ underwent an external evaluation between 2008-2009 and 2011-2012. It also provides information on whether (and when) teacher $j$ taught 9th grade students and on the average performance of these students on exams taken at the end of 9th grade as well as on exams taken subsequently at the 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 on (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 on these different tests. We also know whether students choose science as their major field of study at the end of 10th grade and whether they graduate 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 on 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 their major field of study at the end of 10th grade (d) the proportion of students who subsequently graduate 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.

Finally, 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[29].

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.

Finally, by averaging the variables of this database at the teacher $\times$ 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 students to choose science as their major field of study in high school and to graduate in science.

[29]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) Priority | (7) Non Prior. |
|---|---|---|---|---|---|---|---|
| *Math* | 0.042** | 0.030 | 0.057** | 0.053** | 0.035* | 0.079** | 0.032** |
| | (0.014) | (0.019) | (0.020) | (0.020) | (0.019) | (0.030) | (0.015) |
| *Humanities* | 0.009 | -0.005 | 0.025 | 0.017 | 0.004 | 0.007 | 0.012 |
| | (0.013) | (0.018) | (0.020) | (0.020) | (0.018) | (0.032) | (0.015) |
| Observations | 31102 | 16492 | 14610 | 14319 | 16783 | 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 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 2008-2009 (i.e., above or below 11 years), columns (6) and (7) to the subsample of teachers who were in priority education schools in 2008-2009 and the subsample who were in non-priority schools. Models include a full set of teachers, 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) Priority | (7) Non Prior. |
|---|---|---|---|---|---|---|---|
| *Science as* | 0.004** | 0.000 | 0.008** | 0.007** | 0.001 | 0.008** | 0.003 |
| *major field* | (0.002) | (0.003) | (0.003) | (0.003) | (0.003) | (0.004) | (0.002) |
| | [0.180] | [0.188] | [0.172] | [0.163] | [0.195] | [0.125] | [0.195] |
| *Graduation* | 0.004** | 0.001 | 0.008** | 0.009** | 0.001 | 0.008** | 0.003 |
| *in science* | (0.002) | (0.003) | (0.003) | (0.003) | (0.003) | (0.004) | (0.002) |
| | [0.153] | [0.159] | [0.145] | [0.136] | [0.167] | [0.101] | [0.166] |
| Observations | 31102 | 16492 | 14610 | 14319 | 16783 | 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 their major field of study at the end of 10th grade on a dummy indicating that they underwent an external evaluation between 2008-2009 and $t$. The second row shows the result of regressing the proportion of their 9th grade students who will graduate in science at the end of 12th grade on the same independent variable. The first column refers to the full sample, whereas columns (2) to (7) refer to subsamples defined by teachers' gender, number of years of teaching experience in 2008-2009 (above/below 11 years), and type of school attended in 2008-2009 (priority/non priority). Models include a full set of teachers, 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 C3: Robustness check - 9th grade math teacher evaluation and student performance by subgroups, without student controls

| | (1) Female | (2) Male | (3) Low-exp | (4) High-exp | (5) Priority | (6) Non Priority |
|---|---|---|---|---|---|---|
| *Math score* | 0.024 | 0.060** | 0.048** | 0.039* | 0.064* | 0.036** |
| | (0.022) | (0.023) | (0.022) | (0.022) | (0.034) | (0.018) |
| *Humanities score* | -0.017 | 0.028 | 0.014 | 0.003 | -0.024 | 0.016 |
| | (0.022) | (0.023) | (0.024) | (0.021) | (0.037) | (0.018) |
| Observations | 15318 | 13838 | 14319 | 14837 | 6265 | 22891 |

*Note*: The table refers to our working sample of math teachers who teach 9th grade students between 2008-2009 and 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 2008-2009 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 2008-2009 (i.e., above or below 11 years), columns (5) and (6) to the subsample of teachers who were in priority education schools in 2008-2009 and the subsample who were in non-priority schools. Models include a full set of teacher, region and year fixed effects as well as controls for teachers' experience and seniority. Standard errors (in parentheses) are clustered at the teacher level. * p<0.10, ** p<0.05.

Table C4: Robustness check - 9th grade math teacher evaluation and student high school outcomes, without student controls

| | (1) All | (2) Female | (3) Male | (4) Low-exp | (5) High-exp | (6) Priority | (7) Non Prior. |
|---|---|---|---|---|---|---|---|
| *Science as* | 0.004* | 0.000 | 0.009** | 0.006** | 0.002 | 0.006 | 0.004 |
| *major field* | (0.002) | (0.003) | (0.003) | (0.003) | (0.003) | (0.004) | (0.002) |
| | [0.178] | [0.184] | [0.171] | [0.163] | [0.192] | [0.124] | [0.192] |
| *Graduation* | 0.004* | 0.001 | 0.008** | 0.008** | 0.001 | 0.006 | 0.004 |
| *in Science* | (0.002) | (0.003) | (0.003) | (0.003) | (0.003) | (0.004) | (0.003) |
| | [0.150] | [0.156] | [0.144] | [0.136] | [0.164] | [0.100] | [0.164] |
| Observations | 29156 | 15318 | 13838 | 14319 | 14837 | 6265 | 22891 |

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

Table C5: Robustness check - 9th grade French language teacher evaluation and student performance by subgroups, without student controls

| | (1) Female | (2) Male | (3) Low-exp | (4) High-exp | (5) Priority | (6) Non Priority |
|---|---|---|---|---|---|---|
| *French lang. score* | 0.029 | 0.038 | 0.033 | 0.025 | 0.136** | -0.003 |
| | (0.019) | (0.044) | (0.026) | (0.024) | (0.043) | (0.019) |
| *Mathematics score* | 0.009 | 0.040 | 0.007 | 0.019 | 0.045 | 0.002 |
| | (0.018) | (0.043) | (0.025) | (0.023) | (0.038) | (0.019) |
| Observations | 24624 | 4883 | 13331 | 16176 | 6479 | 23028 |

*Note*: The table refers to our working sample of French language teachers who teach 9th grade students between 2008-2009 and 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 2008-2009 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 2008-2009 (i.e., above or below 11 years), columns (5) and (6) to the subsample of teachers who were in priority education schools in 2008-2009 and the subsample who were in non-priority schools. Models include a full set of teacher, region and year fixed effects as well as controls for teachers' experience and seniority. Standard errors (in parentheses) are clustered at the teacher level. * p<0.10, ** p<0.05.