Institute of Education

Teacher Screening, On the Job Evaluations and Performance

Asma Benhenda

Department of Quantitative Social Science
Working Paper No. 18-06
July 2018

Disclaimer

Any opinions expressed here are those of the author(s) and not those of the UCL Institute of Education. Research published in this series may include views on policy, but the institute itself takes no institutional policy positions.

DoQSS Workings Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

Department of Quantitative Social Science, UCL Institute of Education, University College London,

20 Bedford Way, London WC1H 0AL, UK

Teacher Screening, On the Job Evaluations and Performance

Asma Benhenda[1]

Abstract

I study the relationship between systematic and complementary screening and on-the-job teacher evaluations by their hierarchy, and teacher performance in secondary school. Using comprehensive French administrative data, I exploit within student across topics variations and I find that the classroom observation grade is the only evaluation grade significantly related to teacher performance. I then investigate whether the classroom observation has an impact on teacher performance and behavior during the year of evaluation and in subsequent years. An event study shows that the classroom observation has no statistically significant impact on student achievement. I find that teachers are more absent during the months following the evaluation, suggesting that this evaluation provokes a temporary change in teacher behavior.

JEL Codes: I2, J2, M51
Keywords:

Contact Details: Asma Benhenda (a.benhenda@ucl.ac.uk), Institute of Education, University College London

Acknowledgements: This paper was previously circulated with the title “How to Identify Good Teachers? Teacher Evaluations and Student Achievement". I am deeply grateful to my advisors Julien Grenet and Thomas Piketty for their invaluable guidance and support. Part of this paper was completed during my visit at Columbia University, I am grateful to Jonah Rockoff for insightful suggestions. I thank Marc Gurgand, seminar participants at Paris School of Economics, Columbia University and the French Ministry of Education for comments. I also thank the French Ministry of Education for help with the data. I acknowledge financial support from the Alliance Program of Columbia University.

[1] Institute of Education, University College London

1 Introduction

A lot of public resources are spent on evaluating teachers. In the United States for example, it can cost up to $4,000 per teacher each year [1]. Teacher evaluation is widespread in many developed countries: across OECD countries, more than 75 % of students are enrolled in schools evaluating their teachers (Isore, 2009). In spite of the importance of this practice, there is little empirical evidence on its efficiency with respect to its two main objectives: accountability and human capital formation, through training and incentives. The small set of existing papers focuses on the accountability dimension, and often in a controlled environment (Kane et al., 2013) or on a small sample of teachers (Jacob and Lefgren, 2008; Kane et al., 2011; Jacob et al., 2016; Bacher-Hicks et al., 2017). To my best knowledge, there are very few papers analysing the human capital formation aspect of teacher evaluation (Taylor and Tyler, 2012; Dee and Wyckoff, 2015). Yet most of them focus on a small sample of teachers and are unable to analyze the training and incentives dimension separately.

This paper analyzes the efficiency of teacher evaluation with respect to its two objectives using administrative data on 22,519 teachers and 502,302 students covering French public secondary schools from 2006-2015. How efficent are teacher evaluations in identifying good teachers ? Do teacher evaluations have an impact on subsequent teacher performance ?

In France, teacher skills are assessed by three different actors. First, before recruitment, candidates are assessed by the certification board, through the teacher certification grades. In France, contrary to other countries such as the United States, the certification process is competitive and designed to assess topic-specific content knowledge, with two levels of certifications, basic (CAPES) and advanced (*Agrégation*). For both certification levels, it is a two-stage process: candidates take a written exam, and then, those who pass take an oral exam. Second, teachers are assessed on the job by professional inspectors through a classroom observation designed to assess pedagogical skills. Inspectors also give teachers feedback. Third, teachers are also assessed on the job every year by their school principals, through the administrative grade designed to assess teacher behavior outside the classroom. Both the classroom observation and the administrative grade can have a small impact on teacher wage progression.

In this paper, I start by studying the screening/accountability objective of teacher evaluation. I analyze the relationship between the three teacher evaluations grades and

[1] This figure corresponds to the Cincinatti teacher evaluation system, see Taylor and Tyler (2012) for more details.

teacher effectiveness in raising student test scores. I exploit the fact that, in secondary school, teachers are topic-specific to identify the relationship between teacher evaluation grades and student achievement gains. More precisely, I exploit within student, across topics variations in teachers, and a fortiori in teachers' evaluation grades, to identify their relationship with teacher effectiveness in raising her students' test scores in 9th grade and 12th grade. I analyze these three evaluations grades separately, but also jointly, through "horse races", in order to study which of the evaluation grade is the most strongly related to student achievement gains. I find neither the certification level (CAPES vs. Agrégation), nor the certification grades (written nor oral) are associated with student achievement gains, whether analyzed separately or jointly in a horse race with the other evaluation grades. I also find that the administrative grade is not statistically associated with student achievement, whatever the specification. The only evaluation grade significantly associated with student achievement gains is the pedagogical grade. Both in 9th grade and 12th grade, one standard deviation increase in the pedagogical grade is associated with around two percent of a standard deviation increase in student achievement gains. In other words, one standard deviation increase in the pedagogical grade is on par with replacing an average teacher with a teacher at the 40th percentile of the teacher value-added distribution. I find that low income students are more sensitive to the pedagogical grade, especially in 12th grade.

Second, I analyze the impact of the classroom observation on subsequent teacher performance. I focus on the classroom observation because i) the previous analysis shows its corresponding grade, the pedagogical grade, is the only one significantly related to teacher effectiveness; ii) contrary to the school principal evaluation, this evaluation does not occur every year. This allows me to conduct an event study. The classroom observation can impact teacher performance by improving teacher skills through the feedback they receive from inspectors. It can also impact performance by providing teachers incentives to exert effort. In order to distinguish between these two mechanisms, I exploit detailed data on teacher absence spells, which can be seen as a proxy for effort. I focus on 9th grade teachers, for which I have the best quality teacher absence data. I perform an event study in order to analyze the impact of the classroom observation on student test scores in 9th grade. I deal with endogeneity coming from non-random teacher - student matching with teacher and classroom-year fixed effects. In other words, I exploit within teacher, across year and within classroom-year, across teacher variations in the timing of the classroom observation. I find that the classroom observation has no statistically significant impact on student test scores in 9th grade. I then exploit the fact that I have precise date of the classroom observation and of the

teacher absence spells to perform a within year, across month event study. This within year approach enables me to compare teacher effort, as measured by the number of days of absence, within student. I find that teachers are more absent in the months following the evaluation than the month preceding the evaluation. On average, teachers are 0.35 days more absent in the month following the evaluation than the month just preceding it. This suggests that the classroom observation provokes a small and temporary change in teacher behavior.

The main contribution of this paper is i) to study a set of systematic evaluations, which occur both before recruitment and on the job, and which are aimed at measuring different types of teachers' skills; ii) to analyze both objectives of teacher evaluation : accountability and human capital formation. This paper contributes to several strands of the literature. First, this paper contributes to the literature analyzing on-the-job evaluations of teachers, such as classroom observations (Kane et al., 2010; Bacher-Hicks, 2017; Taylor and Tyler, 2012; Dee and Wyckoff, 2015) or principal evaluations (Jacob and Lefgren, 2008). This literature shows that both classroom observations and principal evaluations are significantly related to teacher effectiveness. I also take a step forward by comparing the two evaluations to one another, through "horse races". This approach is motivated by the intuition that teaching is a complex, multidimensional task and that each of these three evaluations targets a different dimension of this task. The existing studies are conducted in very specific contexts whith frequent, feedback intensive and high stake evaluations, which is not representative of most teacher evaluation systems (Steinberg and Donaldson, 2016). Furthermore, most of this literature analyzes the accountability objective of teacher evaluations. To my best knowledge, only a handful of papers analyze the incentive and training objectives of teacher evaluation (Taylor and Tyler, 2012; Dee and Wyckoff, 2015). Finally, this litterature focuses a small sample of teachers. In this paper, I analyze the whole universe of public secondary school teachers in France, which represent more than 300,000 teachers per year.

Second, it contributes to the literature on screening measures of effective teaching. This literature mostly focuses on teacher certification (Kane, Rockoff and Staiger, 2008) and finds that it is, at best, a very weak predictor of teacher quality. Contrary to the United States, candidates in France do not take college courses in K-12 education nor major in education. Furthermore, while teacher certification in the United States is not selective nor competitive (Koedel, 2011), the certification process in France is academically demanding and has very low passing rates. This is particularly the case for the higher level of certification, the *Agrégation*, which draws applicants from the

top French universities and has a passing rate of around 10 %. In that sense, this paper relates to the literature on Teach for America, a highly selective program which recruits college graduates from top US universities to serve as teachers in low income areas. These papers find positive effects of this program in Math (Boyd et al., 2006; Kane et al., 2008; Henry et al., 2014) . While Teach for America is an alternative certification program, concerning only a small fraction of candidates, the French certification process is government-run and the only way to become a tenured teacher. Furthermore, in this paper, I analyze not only the impact of the certification level, but also of the precise certification test scores, at both stages (written then oral) of the certification process. That relates this present paper to a recent paper which uses detailed data on teacher applications to an optional centralized multi-stage application process in Washington DC (Jacob et al., 2016). One of its main findings is that applicants mock interview scores strongly predict teacher effectiveness.

Finally, because this paper compares evaluations that are designed to measure two distinct type of skills, advanced content-knowledge and pedagogical skills, it relates to the very small economic literature on effective teaching practices Lavy (2015b). In particular, the fact that the certification grade is not related to teacher effectiveness contrary to the pedagogical grade, gives suggestive evidence that teacher pedagogical skills are more relevant than her content-knowledge.

The remainder of this paper proceeds as follows. Section 2 describes the three evaluation grades. Section 3 presents the data. Section 4 exposes the empirical approach. Section 5 analyzes the main results. The last section concludes.

2 Institutional Setting: Teacher Evaluations in France

2.1 The Certification Grade

In France, the teacher labor market is highly centralized: teachers are civil servants certified, recruited, paid and managed by the government. Teacher certification is obtained after passing a competitive national examination. This examination is taken after a year of intensive preparation at university departments specifically dedicated to teacher training. The examination for teaching in secondary school (*collège*) or high school (*lycée*) is subject-specific. There are two main certification levels for teachers teaching in secondary or high schools. The basic certification level is called *Certificat d'aptitude au professorat de l'enseignement du second degré* (CAPES). CAPES recipients (called *Certifiés* or *Capésiens)* are essentially meant to teach in secondary school

(which includes 9th grade) or in high school (which includes 12th grade). The advanced certification level is called *Agrégation*. *Agrégation* recipients (called *Agrégés*) are essentially meant to teach in the academic track of high school (which includes 12th grade) and sometimes in higher education, at the undergraduate level [2].

For both certification levels, the examination is composed of two successive stages: a written examination stage and an oral examination stage. First, candidates have to take written tests. For French literature and History, these tests are written essays. For mathematics, they consist in problem sets. Second, candidates who passed the written stage can take the oral tests. These tests are composed of three main parts. The first part consists in a lesson given in front of the selection board. The second part consists in an interview. The last part consists in a critical analysis of a text in French literature and in an exercise in mathematics [3]. Overall, the certification examinations are mostly academic exercises designed by public university to provide comprehensive assessments of advanced subject-specific content knowledge. This seems to be quite specific to the French context. In the United States for example, the certification examinations, called Praxis tests, are designed to assess equally both academic knowledge and pedagogical skills.

The selection board is essentially composed of experienced teachers and university professors. As an illustration, Table 2 reports individual characteristics of the selection board for the Capes certification exam in 2012 for both the Math and the French exam. [4] These boards are composed of more than a hundred members (128 for the Math exam, and 105 for the French exam), and half of them are Agrégés teachers. Those who are not Agrégés are either professional inspectors or university professors. Members of the selection board are on average very experienced: approximately 18 years of experience for both the Math and the French exam. The pedagogical and administrative grade of those who are teachers can be observed. On the one hand, they are situated in the top half of the distribution of the pedagogical grade. In Math, selection board members are on average at the 68th percentile of the pedagogical grade. In French, they are at the 55th percentile on average. On the other hand, they are situated in the bottom half of the distribution of the administrative grade : 33th percentile for Math, and 40th percentile for French.

I standardize the certification grade by certification level, subject and year. This

[2]The *Certifié* and *Agrégé* statuses are defined, respectively, by the Decree n°72-581 of July 4, 1972 and by the Decree n°72-580 of July 4, 1972.

[3]The distribution of the written exam and the oral exam grades are reported in Figure 1.

[4]The selection board of the Agrégation is overwhelmingly composed of university professors, for whom no administrative data is available.

standardization allows us to control for the differences in the selectivity of each subject-specific examination across years. [5]

Table 3 reports estimates from regressions of candidates' individual characteristics on the certification grades (written and oral exams). Estimates are reported for all candidates (columns 1 and 2) and only for admitted candidates (columns 3 and 4). The signs of several estimates changes when the estimation is restricted to admitted candidates. This suggests that the relationships between candidates individual characteristics and certification grades is not linear and depends on where candidates are in the distribution of the grades. For example, being a student rather than a certified teacher when taking the exam has a positive impact on the certification grades for all candidates, but a negative impact on the certification grades of admitted candidates. Furthermore, the relationship between the written exam grade and the oral exam grade is stronger for all candidates than for admitted candidates. For all candidates, a standard deviation increase in the written exam grade is associated with a 50 % of standard deviation increase in the oral exam grade whereas it is equal to only 9 % of a standard deviation for admitted candidates. This weak correlation suggests that, for admitted candidates, the written and the oral exams measure very different types of skills.

2.2 The Pedagogical Grade

The pedagogical grade is a practice-based measure, evaluating teachers by directly observing them in their classroom (Attali and Bressoux, 2002; Isoré, 2009; IGEN, 2013b). It relies on a single classroom observation made by professional inspectors (called *Inspecteur d'académie - inspecteur pédagogique régional*). Professional inspectors are high ranked civil servants recruited through a national competitive examination restricted to experienced civil servants. Professional inspectors are former experienced teachers. In 2015 for example, inspectors are on average 52.43 years old and 54 % of them are males (Table 4). They have on average 20.22 years of experience as a teacher and 7.70 years of experience as an inspector. There are 3,295 inspectors, which means that, on average, there is approximately one inspector per 100 teachers.

Figure 4 shows the distribution of the pedagogical grade, by level of certification. Capésiens (teachers with the Capes) have a lower average certification grade than Agrégés (teachers with the Agrégation). The distribution is bell-shaped for Capésiens, whereas it has a higher right tail for Agrégés. Agrégés receive good grades more freqnently than bad grades, and more frequently than Capésiens do. Professional

[5] However, this standardization does not enable us to control for the difference in selectivity across subjects within a given year.

inspectors are asked to follow a national grading table, which depends on the teacher's certification level and ranking on the wage scale (Table 5). The aim of this grading scale is to make sure that there is enough variation within each notch of the wage scale [6] because, as we shall explain in detail below, this grade is used in the teacher promotion process. In Table 5, we mainly observe that the minimum and maximum grades increase with the ranking on the wage scale and the certification level. For example, the pedagogical grade of Capésiens whose rank on the wage scale is inferior to four must be between 32 and 47 points. This grading scale justifies in particular the standardization of the pedagogical grade by teachers' certification level and ranking on the wage scale.

The grading criteria are twofold. First, inspectors are required to check whether teachers follow the syllabus defined by the Ministry of Education. Second, inspectors evaluate pedagogical skills. Table 6 reports estimates of the regression of the standardized pedagogical grade on individual teacher characteristics such as gender, number of years of experience, teaching topic, certification level (Agrégé), and absence behavior (number of absence spells and number of days of absence). Male teachers, French teachers and Agrégés have slightly worse grades than other teachers, all other variables kept equal, both with and without school fixed effects. For example, being a male teacher is associated with a decrease of 5 - 8 percent of a standard deviation in the teachers' pedagogical grade. Teaching experience is positively associated with the pedagogical grade. Each additional year of experience is associated with an increase in the standardized pedagogical grade by three to six percent of a standard deviation. Teachers' absences and the pedagogical grade do not seem to be associated in a statistically significant way. Overall, the small magnitude of the correlations between teacher observable characteristics suggests that the pedagogical grade captures skills that are weakly correlated to teacher characteristics. In theory, novice teachers should be more frequently inspected: they should be systematically graded during their first year of teaching in order to get tenure and are inspected every three years throughout the beginning of their career (Suchaut, 2012). In practice, we observe in the data that, on average, teachers are inspected approximately every seven years, with variations across teaching topics (Figure 2). For French teachers, the average number of years between two inspections is 7.51 years, whereas for Math teachers it is 6.37 years and for Physics teachers it is 5.89 years. The inspection is more likely to happen at the beginning of the career than at the end. As shown in Figure 3, approximately 20 % of

[6] Memorandum n° 96-024 of January 9, 1996: " *L'objectif est[...] d'assurer [...] pour chaque échelon, une répartition bien étalée des notes pédagogiques.*"

inspections happen during the first five years of experience, with a peak of 8 percent during the third year of experience.

The pedagogical grade is designed to assess the efficiency of classroom practices. Inspectors are external observers and are therefore less likely to have preconceptions about the teachers they grade. However, the pedagogical grade can also be a noisy measure, especially because it is based on a single and two hours long classroom observation and because the evaluation criteria are not precisely defined (Bressoux, 2008). This raises the concern that inspectors characteristics may contaminate the outcome of the evaluation. In particular, it seems *a priori* that this evaluation system is less precise than another similar practice-based evaluation system, the Cincinnati Teacher Evaluation System, studied by Kane et al. (2010). First, in Cincinnati, teacher evaluation is based on four different classroom observations conducted periodically throughout the school year. Second, the Cincinnati Teacher Evaluation System uses a specified and research-based evaluation rubric (called the Danielson rubric). This rubric includes a very precise description of the practices and skills that effective teachers should possess and employ. As the Cincinnati Teacher Evaluation System seems more precise, we might expect the correlation between practice-based evaluation and teacher effectiveness to be stronger in Kane et al. (2010) than in the present study.

2.3 The Administrative Grade

Teachers are evaluated each year by their school principal through an administrative grade. School principals are teachers' immediate hierarchical superior. However, they are not in charge of hiring, promoting nor firing teachers. Their job is mostly to manage teachers on a daily basis. School principals are recruited through a national competitive examination restricted to experienced teachers (DGRH, 2010). Table 7 reports school principals' individual characteristics. On average, school principals are more likely to be males (55 percent of them are male). They have 18.6 years of experience as a teacher and 6.5 years of experience as principals. Their last pedagogical grade and administrative grades as a teacher is observed in the data. On average, they have a median pedagogical grade as they are at the 52th percentile of the distribution of the pedagogical grade. They are situated in the top half of the distribution of the administrative grade as they are at the 61th percentile of the distribution of this grade.

Figure 5 plots the distribution of the administrative grade, by level of certification. For both level of certification, the distribution has a long left-tail, with a peak around 40, the maximum grade. The density at the peak is higher for Agrégés

than for Capésiens, indicating that Agrégés have more frequently the maximum grade than Capésiens. Like the pedagogical grade, the administrative grade depends on the teacher's certification level and ranking on the wage scale, according to a national grading table (Table 8). The structure of the national grading table for the administrative grade is, however, different from the pedagogical grade table. In the pedagogical grade table, for each rank on the wage scale, the intervals have approximately the same size (approximately 15 points). In the administrative grade table, the intervals become smaller as we go up in the wage scale (from 5 points to one point for the Capésiens). In the pedagogical grade table, the grades can go from 32 points to 60 points. In the administrative grade table, the range is much smaller and can go from 30 points to 40 points. Overall, this means that there is much less room for variations in the administrative grade.

The administrative grade is mainly designed to assess teacher' practices as a civil servant. Principals are explicitly instructed to exclude all pedagogical criteria from their evaluation [7]. More precisely, this grade evaluates teachers according to the following broad criteria: regular attendance, ponctuality, activity, influence [8]. Teachers have the right to access and challenge this grade. This may explain why the majority of teachers get the maximum grade. The administrative grade is therefore very different from the principal evaluations studied by Jacob and Lefgren (2008). In the latter, evaluations come from a survey made by the authors where they ask school principals to rate anonymously the teachers of their school. Principals were asked not only to provide an overall evaluation but also to assess specific teaching skills such as dedication, classroom management, parent satisfaction and ability to raise achievement. Table 9 reports estimates of the regression of the standardized administrative grade on individual teacher characteristics. Surprisingly, the correlation between teacher absences and the administrative grade is not statistically significant. This suggests that the administrative grade does not actually measure attendance. As the pedagogical grade, the administrative grade is correlated with teachers' experience: an additional year of experience is associated with an 7 - 9 percent increase in the standardized administrative grade.

[7] Circular of December 13, 2013: "*"appréciation sur la manière de servir de l'enseignant, en dehors d'appréciation à caractère pédagogique"*

[8] *assiduité, ponctualité, activité, rayonnement*

2.4 Impact of the Pedagogical and the Administrative Grade on Teachers' Careers

The two on the job evaluations are used directly in the teacher promotion procedure and indirectly in teacher mobility. Teacher salaries are determined by the Ministry of Education through a national wage scale. The main criteria for promotion is teaching experience. However, promotion can also be fostered by positive on the job evaluations. More precisely, teachers are ranked on a list for promotion (*tableau d'avancement*) according to the weighted average of their pedagogical grade (60 percent) and their administrative grade (40 percent). There are three ranking levels: high (*grand choix*), medium (*choix*) and low (*ancienneté*). Teachers ranked at the top of the list for promotion (*grand choix*) need less teaching experience to go up on the wage scale than teachers at the bottom of the list for promotion (*ancienneté*). For example, to go from the fifth notch to the sixth notch on the wage scale, teachers ranked at the top of the list for promotion need two years and six months of experience whereas teachers ranked at the bottom of the list for promotion need three years and six months of experience.

Teacher mobility is managed by an automated assignment mechanism, called Siam (*Système d'information et d'aide pour les mutations*), with two successive phases. First, teachers willing to leave their Local Educational Unity (*académie*) are assigned to a new Local Educational Unity. Second, (a) teachers assigned to a new Local Educational Unity and (b) teachers willing to switch school within their current Local Educational Unity are both assigned to their new school [9]. In both phases, priority between teachers is determined by their ranking on a bonus scale that mainly takes into account teachers' family situation, experience, seniority and ranking on the wage scale. *Agrégés* also receive a bonus for assignment in high school. Therefore, the two on the job evaluations are indirectly taken into account for mobility, *via* their effect on teachers' ranking on the wage scale.

3 Data and Summary Statistics

3.1 Description of the Data

This study relies on administrative data provided by the Statistical Department of the French Ministry of Education. Our set of data is composed of four main databases (also presented in Table 1):

[9] for a detailed description of teacher mobility procedures in France, see DEPP(2014) or the Matching in practice research network website (Terrier, 2014)

(i) individual data on certification examinations candidates including their name, their date of birth, their exam test scores and whether they passed or not. This database is extracted from the national OCEAN system. This data covers school years 2001-2002 to 2011-2012. However, the name variable is available only since the 2005-2006 school year.

(ii) individual data on teachers, school principals and inspectors including their national identification number, their name, their date of birth, their personal characteristics. For teachers, the data includes their teaching subject(s), and, crucially, the identification number of the school and of the class in which they teach. All this information is mainly available in two databases, called *Annuaires* and *Relais*. These two databases cover school years 2001-2002 to 2014-2015. These two databases are merged with data on certification examinations based on the name, sex and date of birth variables.

(iii) individual data on students including socio-demographic characteristics such as gender and financial aid status [10] (*bourse sur critères sociaux*), an encrypted national identification number, their grades on the two national and externally grades examinations taken in the final year of 9th grade (the *Diplôme national du brevet* – hereafter DNB) and in the final year of 12th grade (*Baccalauréat*), the identification number of their school and of their class. These two latter variables enable us to match each teacher to her students. All this information is collected at the regional level (in databases called *Bases élève académique*) and gathered in a single national database by the Statistical department of the Ministry of Education. This database covers school years 2005-2006 to 2014-2015.

iv individual data on teacher absence spells for 9th grade teachers including the detailed dates of the absence spells. This datasets is mergeed with the other teacher data through teacher's individual identifier.

The construction of the final samples required numerous and sometimes delicate merges between the different databases. The main merging procedures and their outcomes are described in detail in the data appendix.

[10] The financial aid status is not reliable in the student database commonly used in France (*Base centrale scolarité*). This is because the *Base centrale scolarité* is a beginning of the school year photography. At the beginning of the school year, the information on students' financial status is still incomplete. The database we are using here is an end of the school year of photography. At the end of the school year, the information on students' financial status is complete. Therefore, the financial aid status variable we are using is reliable.

3.2 Summary Statistics

Table 10 reports a number of summary statistics for teacher characteristics. In order to discuss the external validity of the samples, we also report statistics for all secondary school teachers teaching between 2006-2007 and 2011-2012. Sampled teachers are significantly younger and less experienced than all teachers. The average age difference between all teachers and sampled teachers is equal to 11.2 years and is significant at the one percent level. This large difference can be explained by the fact that our sample is composed of teachers who had passed the certification examination from 2006 to 2011. On average, teachers in the sample have around three years of experience. Sampled teachers are more likely to teach in the Parisian suburbs (Créteil and Versailles *académies*). The average difference in the proportion of teachers teaching in the Parisian suburbs is equal to 22 percentage points and is significant at the 1 percent level. These areas are the most unattractive areas for teachers, as evidenced by the fact that half of the teacher transfer requests comes from these areas (DEPP, 2014). Consequently, the fact that sampled teachers are over-represented in these areas can be explained by the fact that the main criteria for mobility in the national assignment mechanism is experience (see section 1). Therefore, on average, the samples over-represent young and inexperienced teachers teaching in unattractive areas.

Table 11 reports average student characteristics for all students and for sampled students. Low-income students (identified by their financial aid status) and low achievers are over-represented in the samples. For example, 21 percent of all students are financial aid recipients against 31 percent of sampled students. The difference is significant at the 1 percent level. This confirms the fact that our samples over-represent unattractive areas.

Finally, we study the correlation between the three evaluation grades for sampled teachers in order to get a grasp of the relationships between them and the underlying teaching skills they each measure (Table 12). The magnitude of the correlation coefficients cannot be directly interpreted. For direct interpretation of the correlation between evaluation grades, we use the coefficient of determination (R^2), which is equal to the square of the correlation coefficient. As suggested by Table 3, the correlation between the oral certification grade and the written certification grade is weak for admitted candidates. For both 9th and 12th grades teachers, the correlation coefficient is equal to 0.07 and is statistically significant at the one percent level. The R^2 is equal to 0.005, which means that 0.5 percent of the oral certification grade can be explained by the variation in the written certification grade (and reciprocally). The pedagogical grade is significantly correlated to both certification grades at the one percent level.

For 9th grade for example, five percent of the variation in the pedagogical grade can be explained by the variation in the certification grades (both written and oral). Finally, the administrative grade is very weakly correlated to the certification grades: the R^2 is equal to 0.04 percent in 9th grade and to 0.3-0.05 percent in 12th grade. The administrative grade is mildly correlated to the pedagogical grade, with a R^2 equal to 16-25 percent. Overall, the fact that all the evaluation grades are weakly to mildly correlated to each other suggests that these grades do not duplicate each other and measure different types of skills.

4 Relationship between Teacher Evaluation Grade and Student Achievement

4.1 Empirical Strategy

The objective is to estimate the causal relationship between teacher evaluation grades and student achievement. The main identification issue stems from the non-random teacher-student matching: if, for example, teachers with higher evaluation grades tend to be systematically assigned to better students, a naive cross-section regression would lead to upward-biased estimates of the relationship between teacher evaluation grades and student achievement gains. Figures 6 and 7 plots the average share of *Agrégés*(ordered by percentile rank), the average percentile rank administrative, certification and pedagogical grades by the share of financial aid student per school (ordered by percentile rank). They both suggest non-random teacher-student matching. For example, in 12th grade, schools with the largest share of *Agrégés* are those with the smallest share of financial aid students. Another source of concern is that the two on the job evaluations, the pedagogical and the administrative grades, can be contaminated by students' behavior. For example, a teacher assigned to dissipated and slow students will face the risk to have his pedagogical skills underestimated by the professional inspector during her classroom observation. In that case, a naive cross-section would lead to downward-biased estimates of the relationship between the pedagogical grade and student achievement gains.

In the literature, this identification issue is usually addressed with panel data methods. Most studies rely on longitudinal data that includes student test scores for each student across multiple years (Rockoff, 2004; Rivkin, Hanushek and Kain, 2005). Following each student across multiple years allows for the inclusion of their previous year test scores. Consequently, the basic idea of the standard identification strategy in the

literature is to exploit within student variations in teacher credentials *across years*.

In the present study, we address the teacher-student sorting identification issue with student fixed effects. We do not exploit within student variations in teacher evaluation grades *across years* but within student variations in teacher evaluation grades *across topics*. This method has been previously implemented by two main studies. The first is a paper by Lavy (2010) in which he analyzes the effect of student classroom instructional time per week on student achievement. This paper mainly uses the 2006 edition of the OECD international survey PISA (Programme for International Student Assessment). Lavy's identification strategy relies on within student, across topics variations in student instructional time. The second is a paper by Clotfelter, Ladd and Vigdor (2010), in which they analyze the relationship between teacher credentials such as experience, certification status, educational level, certification test scores etc. and student achievement. This paper analyzes four cohorts of North Carolina (United States) tenth grade students between school years 1999-2000 and 2002-2003, mainly focusing on their statewide end-of-course test scores in algebra, English and science. Their identification strategy relies on within student, across topics variations in teacher credentials. Even if these two studies focus on different research questions, they both rely on cross-section or repeated cross-section data. This is why it seems that the within student, across topics identification strategy is very appropriate to our context and data. Indeed – contrary to American K-12 students for example, who are regularly externally evaluated throughout their studies – French K-12 students mainly take only two externally graded examinations throughout their studies[11]. Therefore, we only have two reliable student achievement measures (the DNB and the *Baccalauréat*), with a three year interval. Thus, like Lavy (2010) and Clotfelter, Ladd and Vigdor (2010), we are not able to follow each student across multiple years and to exploit within student, across years variations. However, the fact that both externally graded examinations take place in secondary school, where, contrary to elementary school, students have several and topic-specific teachers, fully allows us to exploit within student, across topics variations in teacher evaluation grades to identify the relationship between teacher evaluation grades and student achievement gains.

Formally, the model we consider is the following:

$$A_{i,s,k,t} = T_{j(i,s,k,t)}\beta + \theta_i + \theta_s + \theta_k + \theta_t + e_{i,s,k,t} \tag{1}$$

[11] sixth grade students (*élèves de sixième*) used to take a national examination in Math and reading at the beginning of their school year but we do not have access to the corresponding data yet.

where:

- $A_{i,s,k,t}$ the achievement of student i in subject s, in school k and in school year t;
- The function $j(i,s,k,t)$ returns the identity of the unique teacher teaching student i, in subject s, in school k and in school year t. $T_{j(i,s,k,t)}$ is a vector of this teacher evaluation grades;
- θ_i student i fixed effect;
- θ_s subject s fixed effect[12];
- θ_k school k fixed effect;
- θ_t school year t fixed effect;
- $e_{i,s,k,t}$ is a student-by-subject specific error term.

Student fixed effects θ_i capture time-invariant student confounding factors such as student family background, ability, etc. Note that by controlling for these student fixed effects, we also control for the school fixed effect θ_k. Consequently, exploiting within-student variation allows for the controlling of a number of sources of potential biases related to unobserved characteristics of the school, the student or their interaction.

To fully grasp the identification hypothesis, equation 1 can be transformed into the following equation:

$$A_{i,s,k,t} - A_{i,s',k,t} = (\theta_s - \theta_{s'}) + (T_{j(i,s,k,t)} - T_{j(i,s',k,t)})\beta + (e_{i,s,k,t} - e_{i,s',k,t}) \qquad (2)$$

Student achievement in subject s is not measured in absolute terms but *relative* to her achievement in subject s'. Similarly, teacher evaluation grades are measured relative to the evaluation grades of the teacher teaching another subject to the same student. Therefore, the identification hypothesis of this strategy formally writes:

$$\mathbb{E}[e_{i,s,k,t} - e_{i,s',k,t} | T_{j(i,s,k,t)} - T_{j(i,s',k,t]} \;=\; 0 \qquad (3)$$

This hypothesis means that the unobservable determinants of students differential achievement across topics are uncorrelated with the corresponding differences in their teachers' evaluation grades. Intuitively, this identification hypothesis would be violated if students who are unobservably *relatively* more able in some subject ($(e_{i,s,k,t} - e_{i,s',k,t}) > 0$) were systematically assigned to teacher with stronger credentials ($T_{J(i,s,k,t)} -$

[12]For Senior high school, this subject fixed effect also allows us to take into account the fact that the student examination in French is not taken the same year as the student examination in Math

$T_{J(i,s',k,t)} > 0$). In that situation, we could not disentangle the effect of teacher credentials from the fact that some students are intrinsically high-achievers in some subjects: our results would overestimate the effect of teacher credentials. *A priori*, it seems that this identification hypothesis is more plausible for 9th grade than for 12th grade. The main reason is that, contrary to 9th grade, there are several tracks in 12th grade, corresponding to subject major (science (*série Scientifique*) and humanities (*série Économique et social*).[13] Jackson (2012) for example shows that there is a positive teacher-student assortative mating across high school tracks. In the US context, it means that the best teachers are assigned to the best tracks (*i.e.* those offering the best college opportunities), which are chosen by the best students. In the French context, it possibly means that the best Math teachers are assigned to the science track. This would lead to the *relative* positive teacher-student assortative mating that threatens the validity of our identification hypothesis if, for example, students who are relatively better in Math than in French chose the science track rather than the literature track. This is why, following Jackson (2012), the analysis of 12th grade is done by track.

4.2 Results

4.2.1 Baseline Results

Emerging from the analysis of the raw correlations between the three evaluation grades done in section 2 is that the evaluation grades are correlated but not duplicate to each other. Therefore, we are not only willing to know how the teacher skills captured by the evaluations measures influence student achievement but also the relative strength of this influence. we run three separate regressions, each of them including a single evaluation grade as an explanatory variable. These three separate regressions are reported in the first line of each table. Then, we run a "horse race" between the evaluation grades by including them jointly in the same regression. This enables us to test the relative strength of the relationship between evaluation grades and student achievement gains. The "horse race" estimates are reported in the last three column of each table. The fourth column of each table reports estimates for the certification grade; the fifth column for the pedagogical grade; the sixth and last column for the administrative grade. To get a grasp of the sorting bias coming from the non-random teacher-student matching, we report estimates both with and without student fixed

[13]There is a third track, which concentrate a minority students, called the literary track (série litéraire) that we do not study in the paper due to the low quality of the data for this track.

effects. All regressions include year fixed effects, topic fixed effects, the interaction between year fixed effects and topic fixed effects.

Ninth Grade. Without student fixed effects (Table 13), and when included separately (first line), being Agrégé, the oral certification grade and the administrative grade are positively correlated with student achievement gains. Being *Agrégé* (rather than *Capésien*) is associated with a 9 percent of standard deviation in student achievement gain in 9th grade. A standard deviation increase in the administrative grade is associated with a 0.99 percent increase in student achievement gains. The estimates do not vary when all the evaluations are included jointly, except for the oral certification grade estimate which becomes statistically insignificant. With student fixed effects (Table 14), whatever the specification, only the estimate associated with the pedagogical grade is statistically significant. These estimates were not statistically significant in the naive specification without student fixed effects which suggests that teachers with low achieving students face a negative bias during their evaluation. A standard deviation increase in the pedagogical grade is associated with a 1.4-1.6 percent of a standard deviation increase in student achievement gain. In other words, one standard deviation increase in the pedagogical grade is on par with replacing an average teacher with a teacher at the 40th percentile of the teacher value-added distribution [14].

This result is consistent with two interpretations. First, the pedagogical grade captures better what makes a good teacher than the other teacher evaluations. This can be because professional inspectors are more efficient at identifying good teachers than other actors such as the selection board of the certification examination or school principals. This can also be because classroom observations are more revealing situations. Second, it can be because the underlying skills the pedagogical grade is meant to measure, i.e. pedagogical skills, are more relevant to teacher quality than the underlying skills measured by the other evaluation grades.

Twelfth Grade. Whether student fixed effects are included or not (Table 15 and Table 16), the pedagogical grade is positively and significantly associated with student achievement gain in the humanities track. With student fixed effects (Table 16) and when the evaluations are included jointly, a standard deviation increase in the pedagogical grade is associated with a three percent of a standard deviation increase in student achievement. For the science track, the coefficient is smaller: a standard deviation increase in the pedagogical grade is associated with a 1.8 percent of a standard

[14] To reach this estimate, we use the standard value-added estimates from the literature: a standard deviation increase in the teacher value-added distribution decreases student achievement by ten percent of a standard deviation (Kane et al., 2008; Chetty et al., 2014)

deviation increase in student achievement. The coefficient is statistically significant but only at the 10 percent level. The fact that the coefficients are smaller in 12th grade than in 9th grade is consistent with the result from the literature according to which teacher effects are smaller in high school than in middle school or primary school (Jackson, 2012).

4.2.2 Robustness Checks

Table 17 reports robustness checks for both 9th grade and 12th grade. All regressions include student fixed effects (additionaly to year fixed effects, topic fixed effects, the interaction between year fixed effects and topic fixed effects). Each line corresponds to a single regression, where all the evaluations are included jointly. The first robustness check consists in not standardizing the pedagogical grade and the administrative grade. This is because the standardization implies that evaluators (inspectors or principals) are actually taking other teachers in the same rank in the wage scale, with the same level of certification as the reference group. A limitation of this standardization is that it does not allow comparison between different ranks in the wage scale, and a fortiori between different levels of experience and different levels of certification. The first line of each panel reports regression estimates without the standardization of the pedagogical and the administrative grade. Overall, the sign and the statistical significance of the results are robust. In 9th grade for example, a one point increase in the pedagogical grade is associated with a 0.6 percent of a standard deviation increase in student achievement.

The second robustness check consists in adding teachers' characteristics as control variables. Student fixed effects control for all students' fixed characteristics but do not control for any of the teachers' individual characteristics that might bias the results. For example, teacher experience can be both correlated with her evaluation grade and her ability to raise student achievement. The second line of each panel reports estimates teachers' control variables: number of years of experience, number of years of experience squared, gender, year of the certification examination. For 9th grade, the sign, statistical significance and magnitude of the pedagogical grade coefficient remains the same. A standard deviation increase in the pedagogical grade is associated with a 1.5 percent of a standard deviation increase in student achievement gain. The coefficient is statistically significant at the one percent level. For 12th grade, in the humanities track, coefficients are also very similar: a standard deviation increase in the pedagogical grade is associated with a 2.6 percent of a standard deviation increase in student achievement gain. In the baseline estimation, this coefficient was equal to

3.1 percent.

4.2.3 Subgroup Analysis

Table 18 reports regression estimates by Student socioeconomic status. Student socioeconomic status is measured by student financial aid status. Each line corresponds to a single regression, where all the evaluations are included jointly. In 9th grade, for financial aid recipient, a standard deviation increase in the pedagogical grade is associated with a 2 percent increase in student achievement. This coefficient is equal to 1.5 percent for non financial aid recipient. The difference is small but statistically significant at the one percent level. [15] In the science track of 12th grade, the coefficient associated to the pedagogical grade is equal to 1.6 percent for non financial aid recipients and is not statistically significant. For financial aid recipient, this coefficient is statistically significant at the five percent level and is equal to 2.9 percent. The difference is therefore larger than for 9th grade and is statistically significant at the one percent level. Finally, for the humanities track, the pedagogical grade coefficient is equal to, for non financial aid recipients, 2.9 percent and is statistically significant at the five percent level. This coefficient is equal to 4.3 percent for financial aid recipients. The difference between the two coefficients is equal to 1.4 percent and is statistically significant at the 1 percent level. Overall, the conclusion is that, whatever the grade or track, low income students are more sensitive to the pedagogical grade than other students.

5 Impact of the Classroom Observation Evaluation on Teacher Performance

5.1 Empirical Strategy

The objective is to estimate the impact of the classroom observation evaluation on teacher performance. Teacher performance is measured through their students' 9th grade test scores and their number of absence days. To overcome the empirical chal-

[15] To determine the significance of the difference between two coefficients $\widehat{\beta_1}$ and $\widehat{\beta_2}$, we use the following test statistic, distributed according to a t-distribution:

$$\frac{\widehat{\beta_1} - \widehat{\beta_2}}{\sqrt{\sigma^2_{\widehat{\beta_1}} + \sigma^2_{\widehat{\beta_2}}}}$$

lenges associated with the non-random teacher-student matching, I follow Taylor and Tyler (2012) and perform an event study within teacher:

$$A_{j,s,c,t} = \sum_{j,t} \delta_{j,t} 1\{t = \tau_j\}_{j,t} + Experience_{j,t} + \theta_j + \theta_s\theta_t + \theta_{c,t} + X_{j,s,c,t} + \epsilon_{j,s,c,t} \quad (4)$$

where $A_{j,s,c,t}$ is the teacher j outcome variable (average student test scores per year or number of absence days per month) in topic s, classroom c and period t, τ_j is the period (school year or month) during which teacher j is evaluated. This specification includes teacher fixed effects which account for time-invariant, non random differences in teacher - student matching within teacher. However, these teacher fixed effects do not account for time-varying counfounders. The most straightforward confounder is teacher experience, which has been shown by many studies (Rockoff, 2004) to be an important determinant of teacher quality. This is why I include teacher experience dummies as control. Other confounders are linked to student unobservable ability. For example, teachers may be assigned to more difficult (easy) students the year of evaluation. This is why I include classroom-year fixed effects, and a vector $X_{j,s,c,t}$ of students socioeconomic background (parental occupation and financial aid status) characteristics. Because teachers are topic-specific and are assigned to the same classroom with the same students for the whole school year, classroom-year fixed effects allows me to exploit within student, across teacher variations in teacher outcomes. This two way fixed effect specification provides unbiaised estimates of the impact of evaluation if, for a given teacher, the timing of her evaluation is not correlated to her students *topic-specific* ability.

The period just before the evaluation is the omitted category. The coefficients of interest are $\delta_{j,t}$. They capture variations in teacher outcome compared to the period just before evaluation. Robust standard errors are clustered by school, which is the most conservative level of clustering.

5.2 Results

Impact on student test scores. Table 19 reports regression estimates of the impact of the classroom observation on student test scores gains in 9th grade. Column 1 reports estimates of the naive specification, without teacher-school nor classroom-year fixed effects. According to this naive specification, the year of the evaluation, student test scores increase by 1.7 % of a standard deviation compared to the year

just before the evaluation. Student test scores increase by 3 % of a standard deviation in the subsequent years. The coefficients are statistically significant at the 1 % level. Estimates remain similar when teacher-school fixed effects are added. However most of the effect disappears when classroom-year fixed effects are added (column 3). This suggests that the effect from the naive specification comes from the positive assortative mating between teacher teachers and students after the evaluation. Thus, with the preferred specification which includes both teacher-school and classroom-year fixed effects, the classroom observation has no statistically significant impact on student test scores gains.

This result diverges from Taylor and Tyler(2012) who find that teachers are more productive during the school year when they are being evaluated, and even more productive in the years after evaluation. This divergence may be explained by the major difference in intensity and thoroughness between the Cincinnati Teacher Evaluation described above. The teacher classroom observation in France is a one shot evaluation of two hours, and has minor impact on teacher training and career.

Impact on teacher absence within year. Graph 8 reports estimates of the impact of the classroom observation on the number of teacher absence days. The specification includes teacher-school, topic, year and month fixed effects. The reference month is the month just before the evaluation. I observe that teacher are less absent in the months preceding the evaluation than in the months following it. Compared to the month just before the evaluation, the number of absence days in the months following the evaluation increases by 0.35-0.5 days. The number of days of absence in the month of the evaluation is not statistically significantly different from the number of days of absence in the month just before the evaluation. Finally, I observe a decreasing trend in the number of days of absence in the month leading to the evaluation. Thus, overall, this graph suggests that teacher classroom observation triggers a behavioral response from teachers by increasing their effort in the month leading and during the evaluation. This behavioral response is only temporary because the number of teacher absence days increases significantly in the months following the evaluation.

6 Conclusion

This paper uses a unique and very rich French administrative dataset matching teachers to their individual students in order to provide new evidence on the relationship between teacher evaluation and teacher performance in secondary school. To identify this relationship, this paper takes advantage of the fact that, in France, teachers are

topic-specific and exploits within student (or classroom-year), across topics variations in teachers.

I start by analysing the relationship between the teacher evaluation grades and teacher impact on student test scores. I find that, both in 9th grade and 12th grade, the pedagogical grade is the only evaluation to have a statistically significant relationship with student achievement gain, even conditional on the other evaluations. Both in 9th grade and 12th grade, one standard deviation increase in the pedagogical grade is associated with around two percents of a standard deviation increase in student achievement gains. In other words, one standard deviation increase in the pedagogical grade is on par with replacing an average teacher with a teacher at the 40th percentile of the teacher value-added distribution. The subgroup analysis suggests that low income students are more sensitive to the pedagogical grade, especially in 12th grade.

I then investigate whether the classroom observation has an impact on teacher performance and behavior during the year of evaluation and in subsequent years. An event study shows that the classroom observation has no statistically significant impact on student achievement. I find that teachers are more absent during the months following the evaluation, suggesting that this evaluation provokes a temporary change in teacher behavior.

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.

Attali, A., & Bressoux, P. (2002). L'évaluation des pratiques éducatives dans les premier et second degrés. *Rapport pour le Haut conseil de l'évaluation de l'école.*

Bacher-Hicks A., Chin M., Kane T., Staiger D. (2017). An Evaluation of Bias in Three Measures of Teacher Quality: Value-Added, Classroom Observations, and Student Surveys, *NBER Working Paper.*

Boyd, D., Grossman P., Lankford H., Loeb S., & Wyckoff J. (2006). How Changes in Entry Requirements Alter the Teacher Workforce and Affect Student Achievement. Education Finance and Policy, 1 (2), 176–216.

Bressoux, P. (2008). L'évaluation des enseignants : recommandations pour une réforme de l'inspection en France. In J. Weiss (Ed.), *Quelle évaluation des enseignants au service de l'école ?* (pp. 19-28).

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the Impact of Teachers I: Evaluating Bias in Teacher Value-Added Estimates. *American Economic Review*.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2010). Teacher Credentials and Student Achievement in High School: A Cross-Subject Analysis with Student Fixed Effects. *Journal of Human Resources*, 45(3), 655-681.

Cour des comptes (2012), *La formation initiale et le recrutement des enseignants*, rapport public annuel.

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

DEPP (2011), *Les concours de recrutement de personnels enseignants du second degré dans l'enseignement public et privé*, Note d'information n°11-24, ministère de l'Éducation nationale, Direction de l'évaluation, de la prospective et de la performance.

DEPP (2014), *Géographie de l'école*, ministère de l'Éducation nationale, Direction de l'évaluation, de la prospective et la performance.

DGRH (2009), *Devenir personnel d'encadrement*, ministère de l'Éducation nationale, Direction générale des ressources humaines.

DGRH (2010), *Devenir personnel de direction*, ministère de l'Éducation nationale, Direction générale des ressources humaines.

Greene, W. H. (2003). *Econometric Analysis*. Pearson Education.

Grosperrin J. (2011), *Rapport d'information sur la formation initiale et les modalités de recrutement des enseignants*, Assemblée nationale.

Hanushek, E. A. (1979). Conceptual and Empirical Issues in the Estimation of Educational Production Functions. *Journal of Human Resources*, 351-388.

Hanushek, E. A., Kain, J. F., O'Brien, D. M., & Rivkin, S. G. (2005). *The Market for Teacher Quality* (No. w11154). 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), 267-271.

Henry, G. T., Bastian K., Fortner C.K., Kershaw D., Purtell K., Thompson C., & Zulli R. (2004). Teacher Preparation Policies and their Effects on Student Achievement. *Education Finance and Policy*, 9 (3), 1–40.

Herrmann, M. A., & Rockoff, J. E. (2012). Worker Absence and Productivity: Evidence from Teaching, *Journal of Labor Economics*, 30(4), 749-782.

IGEN (2013a), *Les difficultés de recrutement d'enseignants dans certaines disciplines*, rapport au ministre de l'Éducation nationale, Inspection générale de l'Éducation nationale.

IGEN (2013b), *L'évaluation des enseignants*, rapport au ministre de l'Éducation nationale, Inspection générale de l'Éducation nationale.

Isoré, M. (2009), Teacher Evaluation: Current Practices in OECD Countries and a Literature Review, *OECD Education Working Papers*, No. 23, OECD Publishing.

Jackson, C. K. (2014). Teacher Quality at the High-School Level: The Importance of Accounting for Tracks. *Journal of Labor Economics*, Vol. 32, No. 4

Jackson, C. K., Rockoff, J. E., & Staiger, D. O. (2014). Teacher Effects and Teacher-Related Policies. *Annual Review of Economics*.

Jacob, B. A., & Lefgren, L. (2005). *Principals as Agents: Subjective Performance Measurement in Education* (No. w11463). National Bureau of Economic Research.

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

Jacob B., Rockoff J., Taylor E., Lindy B., & Rosen R. (2016). Teacher Applicant Hiring and Teacher Performance: Evidence from DC Public Schools, NBER Working Paper No. 22054

Kane, T. J., & Staiger, D. O. (2008). *Estimating Teacher Impacts on Student Achievement: An experimental Evaluation* (No. w14607). National Bureau of Economic Research.

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

Kane, T. J., Taylor, E. S., Tyler, J. H., & Wooten, A. L. (2011). Identifying Effective Classroom Practices using Student Achievement Data. *Journal of Human Resources*, 46(3), 587-613.

Koedel, C. (2009). An Empirical Analysis of Teacher Spillover Effects in Secondary School. *Economics of Education Review*, 28(6), 682-692.

Kramarz, F., Machin, S. & Ouazad, A. (2014). Using Compulsory Mobility to Identify School Quality and Peer Effects. *Oxford Bulletin of Economics and Statistics*, doi: 10.1111/obes.12076.

Lavy, V. (2010). *Do Differences in School's Instruction Time Explain International Achievement Gaps in Math, Science, and Reading?: Evidence from Developed*

and Developing Countries. National Bureau of Economic Research.

Prost, C. (2013). Teacher Mobility: Can Financial Incentives Help Disadvantaged Schools to Retain their Teachers?. *Annales d'Economie et de Statistique*.

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*, 247-252.

Staiger D., and Rockoff J. (2010). Searching for Effective Teachers with Imperfect Information. *Journal of Economic Perspectives*, 24(3): 97-118.

Suchaut, B. (2012, February). L'évaluation des enseignants: contexte, analyse et perspectives d'évolution. In *Conférence sur l'évaluation, Grenoble (France)*.

Terrier, C. (2014), *Matching Practices for Secondary Public School Teachers – France* [URL: http://www.matching-in-practice.eu/matching-practices-of-teachers-to-schools-france/].

Taylor, E.S. and Tyler, J.H.(2012). The effect of evaluation on teacher performance. *American Economic Review*, 102(7), pp.3628-51.

Data Appendix

Table 1 – Description of the data

Name	Observation level	Period covered
OCEAN (CAG)	candidate x year	2002-2012 without the name variables; 2006-2012 with the name variables
ANNUAIRES (EPP)	teacher x year	2002-2012
RELAIS	teacher x class x year	2002-2012
FAERE	student x year	2006-2012

Description of the Merging Procedures

i) Merge between OCEAN (data on certification exam candidates) and EPP (data on teachers). Name of the matched database: CAGEPP

   (a) Matching variables: family name, first name, date of birth, sex.
   For the family name variable and the first name variable, we allow the Levenshtein distance to be equal to 1 or 2 [16]. More precisely, we conclude it is a match if two observations have the same date of birth and sex and if (a) the distance between the family names is equal to 0 or 1 and the distance between the first names is equal to 0,1 or 2. If two observations have the same date of birth and sex but the distance between surnames is equal to 1 and the distance between the first names is greater than 2, we look at the middle name (if there is one). Indeed, it happens that the first name in OCEAN (or in EPP) corresponds to the middle name in EPP (or in OCEAN). Therefore, if two observations have the same date of birth and sex but the distance between surnames is equal to 1, the distance between the first names is greater than 2 and the distance between the first name and the middle name is equal to 0 or 1, we conclude it is a match.

   (b) Proportion of teachers for whom we observe a certification grade by school year:

[16] We use a SAS function called COMPLEV.

- 2006-2007: 9.2 %
- 2007-2008: 12.4 %
- 2008-2009: 15 %
- 2009-2010: 17.6 %
- 2010-2011: 19.9 %
- 2011-2012: 21.6 %

ii) We clean CAGEPP mainly by suppressing duplicate observations.
These duplicates are mainly due to (a) teachers who took different exams the same year or (b) teachers who took different exams in different years. We only keep the observation corresponding to the teacher's actual certification level. For example, if (a) in a given year, a teacher passed both the CAPES and the *Agrégation* but is registered in EPP as an *Agrégé*, we only keep the observation corresponding to her certification grade at the *Agrégation*; if (b) in 2007, a teacher passed the CAPES but, in 2008, passed the *Agrégation*, we keep, in 2007, the observation corresponding to her certification grade at the CAPES but, in 2008, we only keep the observation corresponding to her certification grade at the *Agrégation*; if (c) in 2007, a teacher passed the CAPES but, in 2008, took the *Agrégation* and failed, we only keep, both in 2007 and 2008, the observation corresponding to her certification grade at the CAPES; if (d) in 2007, a CAPES recipient took the *Agrégation* but failed, we suppress this observation; if (e) a teacher passed both the CAPES of mathematics and the CAPES of physics but is registered in EPP as Math teacher, we only keep the observation corresponding to her certification grade at the CAPES of mathematics.
We also suppress observations corresponding to teachers without any certification status but teaching under a fixed-term contract (*enseignants contractuels*) who took and failed a certification examination.

iii) Merge between CAGEPP and RELAIS (data on teachers with the identification number of their class(es))

(a) Matching variable: teacher identification number

(b) Proportion of teachers in CAGEPP for whom we observe the identification number of their class(es) by school year:
- 2006-2007: 93.9 %

- 2007-2008: 85.2 %
- 2008-2009: 81.2 %
- 2009-2010: 82.3 %
- 2010-2011: 79.6 %
- 2011-2012: 80.7 %

iv) The identification number of the class variable is missing in the FAERE database before the 2009-2010 school year. Therefore, from the 2006-2007 school year to the 2008-2009 school year, we merge the FAERE database with the *Scolarité* database, in which the identification number of the class variable is not missing.

   (a) Matching variables: date of birth, place of birth, school identification number, gender, socioeconomic background of her mother, socioeconomic background of her father, options and lunch status.

   (b) Proportion of students in FAERE before 2009-2010 for whom we observe the identification number of their class: 90.8 %

v) Match between Junior high school teachers and Junior high school students

   (a) Matching variables: class identification number, grade identification number, school identification number

   (b) Proportion of distinct Junior high school Math or French teachers in CAGEPP matched with their Junior high school students in FAERE by school year:

- 2006-2007: 97.2 %
- 2007-2008: 92.4 %
- 2008-2009: 91.1 %
- 2009-2010: 78.2 %
- 2010-2011: 83.2 %
- 2011-2012: 99.7 %

vi) Match between Senior high school Math or French teachers and Senior high school students

   (a) Matching variables: class identification number, grade identification number, school identification number

(b) Proportion of distinct Senior high school Math or French teachers in CAGEPP matched with their Senior high school students in FAERE by school year:

- 2006-2007: 60.7 %
- 2007-2008: 90.6 %
- 2008-2009: 94.5 %
- 2009-2010: 93.7 %
- 2010-2011: 92.8 %
- 2011-2012: 94.7 %

Construction of the Estimation Samples

Our final samples cover teachers who have passed their certification examination between school years 2005-2006 and 2010-2011. In particular, they do not include teachers who passed their certification examination before 2005-2006 because the name variable– essential to our merging procedure– is not available for this period. Our samples cover students who have taken the DNB or the *Baccalauréat* between school years 2006-2007 and 2011-2012. More precisely, the two samples we analyze in this study are the following:

(i) ninth grade students (*élèves de troisième*) matched to their Math and French teachers. The sample is composed of students fulfilling the following conditions: we observe both their Math and French teachers, both their Math teacher and their French teacher passed the certification exam the same year (to control for differences in teachers cohort composition–which the "masterisation" reform is likely to make even more significant – and to make teachers' certification grades as comparable as possible), we observe both their Math and French teachers certification grade, pedagogical grade and administrative grade.

(ii) 12th grade Senior high school students (*élèves de terminale*) –hereafter Senior high school students – matched to their Math and French teachers [17]. The sample is composed of student fulfilling the same conditions as those required for Junior high students plus an additional one. This supplementary condition is that we observe not only the student's *Baccalauréat* test scores but also her

[17]The French examination is actually taken in 11th grade (*classe de première*). Therefore, we match students to their 11th grade French teacher.

DNB test scores. This condition is actually only strictly required for the value-added analysis we perform in section 5 but we also apply it for the sample on which is based the within student, across topics analysis in order to guarantee the comparability of the two approaches. The sample counts 8,295 students and 821 distinct teachers.

We focus on Math and French topics for three main reasons. The first reason is that Math and French are the only topics (with History-Geography) for which externally graded test scores are available and relatively comparable both for Junior and Senior high school. The second reason is that it enables us to improve the comparability of our results with those of the literature – as most of the literature on teacher quality focuses on Math and English. The third reason is that Math and French are the two topics for which the threat of teacher spillover effects across topics seems the less plausible. Koedel (2009) for example suggests that teacher spillover effects between Math and English high school teachers are not statistically significant. The threat of teacher spillover effects seems however, *a priori*, more plausible for History-Geography for example, because students' History-Geography test scores also measure students' reading and writing skills taught by their French teacher. Students' Math test scores (French test scores) seem less likely to be contaminated by the effect of teachers teaching another topic than Math (French) to these students.

To facilitate the interpretation and the comparability of our results, we adopt several normalizations. First, we normalize students test scores by subject and by year. Second, we normalize the teacher certification grade by certification level, subject and year. Finally, we normalize teacher pedagogical and administrative grades by year, certification level and ranking on the wage scale, according the national grading tables presented in section 1 (Table 8 and Table 5). These normalizations imply that the estimated coefficients can be interpreted as fractions of a standard deviation of the distribution of individual scores.

Tables and Figures

Table 2 – Individual Characteristics of the Selection Board of the CAPES Certification Exam (2012)

	Math	**French**
Share of Agrégé*	0.48	0.42
Share of Male	0.59	0.46
Age	43.27	43.61
Pedagogical grade (percentile rank)	68.06	55.00
Administrative grade (percentile rank)	32.97	40.31
Experience (in years)	18.86	18.38
Number of observations	128	105

* Those who are not Agrégés are either professional inspectors or university professors. **Source**: French Ministry of Education website (http://www.devenirenseignant.gouv.fr/) and the author's computations. This table is constructed by matching the names of the members selection board of the Capes certification exam in 2012 with the names of the teachers and inspectors in the administrative data on secondary school teachers and inspectors. Members of the selection board who are university professors are not included.

Table 3 – Regression Estimates of Certification Grades on Candidates' Individual Characteristics

Dependent variable:	**All**		**Admitted**	
	Written Grade	Oral Grade	Written Grade	Oral Grade
	(1)	(2)	(3)	(4)
Previous occupation (Ref.: Certified teacher)				
Student	0.182***	0.234***	-0.515***	-0.075***
	(0.007)	(0.017)	(0.016)	(0.020)
Contract teacher	- 0.141***	0.100***	-0.458***	-0.017***
	(0.008)	(0.020)	(0.019)	(0.023)
Male	0.033***	-0.134***	0.112***	0.009***
	(0.005)	(0.009)	(0.008)	(0.009)
Age	-0.016***	-0.018***	-0.008***	-0.009***
	(0.000)	(0.000)	(0.000)	(0.001)
Degree (Ref.: Bachelor's degree)				
Master's degree	-0.141***	-0.050***	0.115***	-0.043***
	(0.007)	(0.013)	(0.012)	(0.014)
Grande école	0.045	-0.042***	0.200***	0.110**
	(0.029)	(0.054)	(0.048)	(0.052)
Written exam standardized grade	–	0.490***	–	0.089***
		(0.008)		(0.009)

Note: * p < 0.1; ** p < 0.05; *** p < 0.01. Robust standard errors in parenthesis. This table reports estimates of regressions of the certification grades (written exam and oral exam) on candidates' individual characteristics. Each column corresponds to a single regression. Columns (1) and (2) reports regression estimates on all candidates. Columns (3) and (4) reports regression estimates on admitted candidates. The sample is all candidates and all admitted candidates, in all teaching topics, from 2002 to 2012.

Table 4 – Professional Inspectors' Individual Characteristics (2012)

Male	0.54
	(0.50)
Age (in years)	52.43
	(6.60)
Experience as a teacher (in years)	20.22
	(8.31)
Experience as an inspector (in years)	7.70
	(6.03)
Nb of observations	3,295

Note: This table reports professional inspectors' individual characteristics in 2012 as reported in the administrative data. The number of observations corresponds to the number of inspectors. All inspectors covering secondary teachers, whatever their topic, are included. Standard deviations in parenthesis.

Table 5 – National Grading Table for the Pedagogical Grade by Certification Level

	Capésiens		*Agrégés*	
Ranking on the wage scale	Min. grade	Max. grade	Min. grade	Max. grade
1,2,3,4	32	47	37	48
5	33	48	39	50
6	34	49	41	51
7	35	50	43	54
8	36	51	45	56
9	38	53	47	58
10	40	55	49	60
11	42	57	51	60

Source: French Ministry of Education website (http://www.education.gouv.fr/cid58632/notations-des-personnels-enseignants.html). This table reports the official national grading table given to inspectors. For example, inspectors are instructed to give teachers who have the Capes and are on the fifth rank on the wage scale (*échelon*) a pedagogical grade comprised between 33 and 48.

Table 6 – Regression Estimates of the Standardized Pedagogical Grade on Teacher Characteristics

Dependent variable: Standardized pedagogical grade	(1)	(2)	(3)
Male	-0.059***	-0.079***	–
	(0.020)	(0.024)	
Experience (in years)	0.034***	0.032***	0.064**
	(0.009)	(0.010)	(0.028)
$Experience^2$	-0.001	-0.000	-0.002
	(0.000)	(0.000)	(0.003)
Topic (Ref.: History)			
French	-0.125***	-0.127***	–
	(0.025)	(0.028)	
Math	-0.030	-0.018	–
	(0.024)	(0.027)	
Agrégé	-0.071**	-0.120***	–
	(0.034)	(0.042)	
Nb of absence spells	0.000	-0.001	-0.003
	(0.000)	(0.003)	(0.003)
Nb of days of absence	0.002	0.000	0.002
	(0.002)	(0.000)	(0.000)
Adjusted R^2	0.006	0.147	0.644
School Fixed Effect	No	Yes	No
Teacher Fixed Effect	No	No	Yes

Note: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$. Robust standard errors clustered by teacher in parenthesis. This table reports estimates of regressions of the pedagogical on secondary school teachers (middle and high school) individual characteristics. Each column corresponds to a single regression. The level of observation is teacher x year. The dependent variable is the standardized (according to the national grading table, cf. Table 5).

Table 7 – School Principals' Individual Characteristics (2012)

Male	0.55
	(0.50)
Age (in years)	50.43
	(7.50)
Experience as a teacher (in years)	18.61
	(7.27)
Experience as a principal (in years)	6.49
	(3.56)
Pedagogical grade (percentile rank)	51.66
	(24.55)
Administrative grade (percentile rank)	61.25
	(38.77)
Nb of observations	13,714

Note: This table reports school principals' individual characteristics in 2012 as reported in the administrative data. The number of observations corresponds to the number of school principals in 2012. Standard deviations in parenthesis.

Table 8 – Grading Table for the Administrative Grade by Certification Level

	Capésiens		*Agrégés*	
Ranking on the wage scale	Min. grade	Max. grade	Min. grade	Max. grade
1,2	30	35	32	35
3	30	35	32.2	36
4	31	36	32.5	37
5	33.5	37.5	33.5	38
6	34.5	38.5	34.5	39
7	36	39	36	40
8	36.5	39.5	37	40
9	37	40	37.5	40
10	38	40	38.5	40
11	39	40	38.5	40

Source: French Ministry of Education website (http://www.education.gouv.fr/cid58632/notations-des-personnels-enseignants.html). This table reports the official national grading table given to school principals. For example, school principals are instructed to give teachers who have the Capes and are on the third rank on the wage scale (*échelon*) an administrative grade comprised between 30 and 35.

Table 9 – Regression Estimates of the Standardized Administrative Grade on Individual Teacher Characteristics

Dependent variable: Standardized administrative grade	(1)	(2)	(3)
Male	-0.053***	-0.031*	–
	(0.015)	(0.020)	
Experience	0.073***	0.079***	0.090***
	(0.007)	(0.007)	(0.019)
$Experience^2$	-0.002***	-0.002***	-0.003
	(0.000)	(0.000)	(0.005)
Topic (Ref.: History)			
French	0.005	0.009	–
	(0.019)	(0.019)	
Math	0.000	-0.016	–
	(0.019)	(0.019)	
Agrégé	-0.241***	-0.234***	–
	(0.038)	(0.054)	
Nb of absence spells	-0.004	-0.002	0.000
	(0.002)	(0.003)	(0.000)
Nb of days of absence	0.000	0.000	0.002*
	(0.000)	(0.000)	(0.003)
Adjusted R^2	0.018	0.22	0.53
School Fixed Effect	No	Yes	No
Teacher Fixed Effect	No	No	Yes

Note: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$. Robust standard errors clustered by teacher in parenthesis. This table reports estimates of regressions of the administrative on secondary school teachers (middle and high school) individual characteristics. Each column corresponds to a single regression. The level of observation is teacher x year. The dependent variable is the standardized (according to the national grading table, cf. Table 8).

Table 10 – Average Teacher Characteristics by Grade (All Teachers and Sampled Teachers)

	All (1)	Sample (2)	Difference (3) = (1) - (2)
A. Demographics			
Female	0.66	0.64	0.02
	(0.47)	(0.48)	(0.02)
Age (in years)	41.40	30.20	11.20***
	(10.10)	(4.90)	(0.22)
B. Qualifications			
Experience (in years)	15.70	2.90	12.70***
	(10.2)	(1.30)	(0.07)
Agrégés	0.06	0.09	-0.03**
	(0.25)	(0.28)	(0.01)
Certifiés	0.84	0.85	-.01
	(0.36)	(0.35)	(0.01)
Other certification status	0.09	0.06	0.03***
	(0.29)	(0.24)	(0.01)
C. School			
Average school size	471.80	544.28	-72.40***
	(213.90)	(201.90)	(8.80)
Teaching in the Parisian suburbs	0.16	0.38	-0.22***
	(0.37)	(0.49)	(0.02)
Number of teachers	106,892	22,519	

Notes: The t-statistic for the comparison of means (columns 3 and 6) is equal to the ratio of the mean of the difference to the standard error of the difference. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$. Standard errors in parenthesis. The statistics are reported for all secondary school teachers (column 1) and for teachers in the estimation sample(column 2), as defined in the data appendix.

Table 11 – Average Student Characteristics (All Students and Sampled Students)

	All (1)	**Sample** (2)	**Difference** (3) = (1) - (2)
A. Demographics			
Female	0.50	0.51	-.01**
	(0.50)	(0.50)	(0.00)
Financial aid recipient	0.21	0.31	-0.10***
	(0.41)	(0.46)	(0.00)
B. Achievement			
Average test scores (/20)	10.40	9.20	1.30***
	(3.90)	(4.0)	(0.04)
Repeated at least once since kindergarten	0.28	0.38	-.11***
	(0.44)	(0.46)	(0.00)
Number of students	1,288,858	502,302	

Notes: The t-statistic for the comparison of means (columns 3 and 6) is equal to the ratio of the mean of the difference to the standard error of the difference. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$. Standard errors in parenthesis. The statistics are reported for all student in 9th or 12th grade (column 1) and for all students in the estimation sample(column 2), as defined in the data appendix.

Table 12 – Pearson Pairwise Correlation Coefficient between the Certification Grades, the Pedagogical Grade and the Administrative Grade

	Certif. (written)	**Certif. (oral)**	**Pedag.**	**Admin.**
A. 9th grade teachers (N= 13,815)				
Certif. (written part)	1.00	0.07***	0.21***	0.02**
Certif.(oral part)	0.07***	1.00	0.22***	-0.02**
Pedag.			1.00	0.39***
B. 12th grade teachers (N = 8,704)				
Certif. (written part)	1.00	0.07***	0.19***	0.05***
Certif.(oral part)	0.07***	1.00	0.13***	-0.07***
Pedag.			1.00	0.49***

Notes: *** $p < 0.01$; Pedagogical and administrative grades are averaged over years. The statistics are computed on the sampled teachers (see data appendix for the definition of the sample).

Table 13 – Regression Estimates of Student Test Scores on Teacher Evaluations in 9th Grade – Naive Estimation

	Agrégé	Certif. (written)	Certif. (oral)	Pedag. grade	Admin. grade
	(1)	(2)	(3)	(4)	(5)
Eval. Separately	.0918***	.0005	.0063**	.0097***	.0099***
	(.0099)	(.0031)	(.0031)	(.0031)	(.0032)
Eval. Jointly	.0892***	-.0002	.0032	.0078**	.0138***
	(.0105)	(.0033)	(.0034)	(.0035)	(.0040)
Controls	No	No	No	No	No
Student fixed effects	No	No	No	No	No
Nb of observations	1,206,907	1,206,907	1,206,907	1,206,907	1,206,907

Notes: * p < 0.1; ** p < 0.05; *** p < 0.01. Robust standard errors clustered by teacher in parenthesis. The dependent variable is the teacher's student standardized test scores at the 9th grade national exam (Diplôme national du brevet). Student test scores are standardized by topic and year. In the first column, Agrégé is a dummy variable equal to one if the teacher has the Agrégation. For column 2 to 5, the evaluation grades are standardized. The certification grades are standardized by year, topic and level of certification (Agrégation vs. Capes). The pedagogical grade and the administrative grades are standardized according to their respective national grading table (cf. Table 5 and Table 8). For the first line (teacher evaluations included separately), each column corresponds to a different regression. For the second line (evaluations included jointly in the same regression) corresponds to a single regression. The level of observation is teacher (topic) x student, from 2006 to 2012. The regressions are run on the sample as defined in the data appendix. All regressions include year fixed effects, topic fixed effects and the interaction between year fixed effects and topics fixed effects.

Table 14 – Regression Estimates of Student Test Scores on Teacher Evaluations in 9th Grade – With Student Fixed Effects

	Agrégé	Certif. (written)	Certif. (oral)	Pedag. grade	Admin. grade
	(1)	(2)	(3)	(4)	(5)
Eval. Separately	-.0011	-.0020	.0059	.0144***	.0054
	(.0163)	(.0047)	(.0043)	(.0042)	(.0045)
Eval. jointly	-.0108	-.0059	-.0001	.0160***	.0071
	(.0175)	(.0049)	(.0049)	(.0049)	(.0056)
Controls	No	No	No	No	No
Student fixed effects	Yes	Yes	Yes	Yes	Yes
Nb of observations	1,206,907	1,206,907	1,206,907	1,206,907	1,206,907

Notes: * p < 0.1; ** p < 0.05; *** p < 0.01. Robust standard errors clustered by teacher in parenthesis. The dependent variable is the teacher's student standardized test scores at the 9th grade national exam (Diplôme national du brevet). In the first column, the variable Agrégé is a dummy variable equal to one if the teacher has the Agrégation. For column 2 to 5, the evaluation grades are standardized. The certification grades are standardized by year, topic and level of certification (Agrégation vs. Capes). The pedagogical grade and the administrative grades are standardized according to their respective national grading table (cf. Table 5 and Table 8). For the first line (teacher evaluations included separately), each column corresponds to a different regression. For the second line (evaluations included jointly in the same regression) corresponds to a single regression. The level of observation is teacher (topic) x student, from 2006 to 2012. The regressions are run on the sample as defined in the data appendix. All regressions include year fixed effects, topic fixed effects and the interaction between year fixed effects and topics fixed effects.

Table 15 – Regression Estimates of Student Test Scores on Teacher Evaluations in 12th Grade – Naive Estimation

	Agreg. (1)	Certif. (written) (2)	Certif. (oral) (3)	Pedag. grade (4)	Admin. grade (5)
A. Science Track (N =255,128)					
Eval. Separately	.0525***	-.0010	.0251***	.0150**	.0047
	(.0126)	(.0064)	(.0061)	(.0062)	(.0049)
Eval. jointly	.0324**	-.0009	.0190***	.0033	.0076
	(.0145)	(.0069)	(.0066)	(.0070)	(.0056)
B. Humanities Track (N= 149,981)					
Eval. Separately	.0060	-.0063	.0027	.0167**	-.0048
	(.0136)	(.0068)	(.0068)	(.0070)	(.0044)
Eval. jointly	-.0163	-.0088	-.0066	.0205***	-.0033
	(.0157)	(.0074)	(.0072)	(.0080)	(.0053)
Controls	No	No	No	No	No
Student fixed effects	No	No	No	No	No

Notes: * p < 0.1; ** p < 0.05; *** p < 0.01. Robust standard errors clustered by teacher in parenthesis. The dependent variable is the teacher's student standardized test scores at the 12th grade national exam (Baccalauréat). In the first column, Agrégé is a dummy variable equal to one if the teacher has the Agrégation. For column 2 to 5, the evaluation grades are standardized. The certification grades are standardized by year, topic and level of certification (Agrégation vs. Capes). The pedagogical grade and the administrative grades are standardized according to their respective national grading table (cf. Table 5 and Table 8). For the first line (teacher evaluations included separately), each column corresponds to a different regression. For the second line (evaluations included jointly in the same regression) corresponds to a single regression. The level of observation is teacher (topic) x student, from 2006 to 2012. The regressions are run on the sample as defined in the data appendix. All regressions include year fixed effects, topic fixed effects and the interaction between year fixed effects and topics fixed effects.

Table 16 – Regression Estimates of Student Test Scores on Teacher Evaluations in 12th Grade – With Student Fixed Effects

	Agreg. (1)	Certif. (written) (2)	Certif. (oral) (3)	Pedag. grade (4)	Admin. grade (5)
A. Science Track (N =255,128)					
Eval. Separately	.0300	.0039	.0095	.0230***	.0025
	(.0174)	(.0081)	(.0080)	(.0089)	(.0084)
Eval. jointly	.0150	.0024	.0056	.0177*	-.0002
	(.0209)	(.0093)	(.0092)	(.0108)	(.0094)
B. Humanities Track (N= 149,981)					
Eval. Separately	-.0089	-.007	.0027	.0202**	-.008
	(.0202)	(.0027)	(.0099)	(.0095)	(.006)
Eval. jointly	-.0370	-.0129	-.0113	.0311***	-.0064
	(.0242)	(.0108)	(.0115)	(.0114)	(.0103)
Controls	No	No	No	No	No
Student fixed effects	Yes	Yes	Yes	Yes	Yes

Notes: * p < 0.1; ** p < 0.05; *** p < 0.01. Robust standard errors clustered by teacher in parenthesis. The dependent variable is the teacher's student standardized test scores at the 12th grade national exam (Baccalauréat). In the first column, Agrégé is a dummy variable equal to one if the teacher has the Agrégation. For column 2 to 5, the evaluation grades are standardized. The certification grades are standardized by year, topic and level of certification (Agrégation vs. Capes). The pedagogical grade and the administrative grades are standardized according to their respective national grading table (cf. Table 5 and Table 8). For the first line (teacher evaluations included separately), each column corresponds to a different regression. For the second line (evaluations included jointly in the same regression) corresponds to a single regression. The level of observation is teacher (topic) x student, from 2006 to 2012. The regressions are run on the sample as defined in the data appendix. All regressions include year fixed effects, topic fixed effects and the interaction between year fixed effects and topics fixed effects.

Table 17 – Regression Estimates of Student Test Scores on Teacher Evaluations in 9th Grade and 12th Grade– Robustness Checks

	Agrégé	Certif. (written)	Certif. (oral)	Pedag. grade	Admin. grade
	(1)	(2)	(3)	(4)	(5)
A. 9th grade (N = 1,206,907)					
Without standardisation	-0.001	-0.002	-0.000	0.006***	0.002
	(0.016)	(0.005)	(0.005)	(0.002)	(0.004)
With control variables	-0.012	-0.004	-0.000	0.015***	0.004
	(0.023)	(0.005)	(0.005)	(0.005)	(0.005)
B. 12th grade – Science Track (N = 255,128)					
Without standardisation	0.040	0.004	0.011	0.002	-0.005
	(.025)	(.009)	(.009)	(.003)	(.003)
With control variables	0.033	0.004	0.008	0.013	-0.003
	(0.026)	(0.009)	(0.009)	(0.011)	(0.009)
C. 12th grade – Humanities Track (N = 149,981)					
Without standardisation	-0.028	-0.010	-0.007	0.005**	-0.005
	(.029)	(.011)	(.011)	(.002)	(.004)
With control variables	-0.024	-0.013	-0.009	0.026**	-0.009
	(0.030)	(0.011)	(0.011)	(0.011)	(0.010)
Student fixed effects	Yes	Yes	Yes	Yes	Yes

Notes: * p < 0.1; ** p < 0.05; *** p < 0.01. Robust standard errors clustered by teacher in parenthesis. The dependent variable is, for 9th grade, the teacher's student standardized test scores at the 9th grade national exam (Diplôme national du brevet) and for 12th grade, the teacher's student standardized test scores at the 12th grade national exam (Baccalauréat). In the first column, Agrégé is a dummy variable equal to one if the teacher has the Agrégation. For column 2 to 5, the evaluation grades are standardized. The certification grades are standardized by year, topic and level of certification (Agrégation vs. Capes). The pedagogical grade and the administrative grades are standardized according to their respective national grading table (cf. Table 5 and Table 8). Each line corresponds to a single regression, where all five evaluation grades are included jointly. The level of observation is teacher (topic) x student, from 2006 to 2012. The regressions are run on the sample as defined in the data appendix. All regressions include year fixed effects, topic fixed effects and the interaction between year fixed effects and topics fixed effects.

Table 18 – Regression Estimates of Student Test Scores on Teacher Evaluations in 9th Grade and 12th Grade – Subgroup Analysis by Student Socio-economic Status

	Agreg.	Certif. (written)	Certif. (oral)	Pedag. grade	Admin. grade
	(1)	(2)	(3)	(4)	(5)
A. 9th Grade					
Non Financial Aid(N=856,905)	.001	-.004	.001	.015***	.005
	(.018)	(.005)	(.005)	(.005)	(.006)
Financial Aid (N=349,994)	.008	-.009	-.005	.020***	.010*
	(.022)	(.006)	(.006)	(.006)	(.005)
B. 12th Grade – Science track					
Non Financial Aid(N=214,858)	.017	.002	.009	.016	-.000
	(.021)	(.009)	(.009)	(.011)	(.009)
Financial Aid (N=40,270)	.009	.002	-.011	.029**	.002
	(.032)	(.015)	(.014)	(.014)	(.014)
C. 12th Grade – Humanities track					
Non Financial Aid(N =121,773)	-.027	-.015	-.009	.029**	-.003
	(.026)	(.011)	(.012)	(.012)	(.011)
Financial Aid (N=28,208)	-.080**	-.003	-.019	.043**	-.020
	(.037)	(.016)	(.016)	(.017)	(.016)
Student fixed effects	Yes	Yes	Yes	Yes	Yes

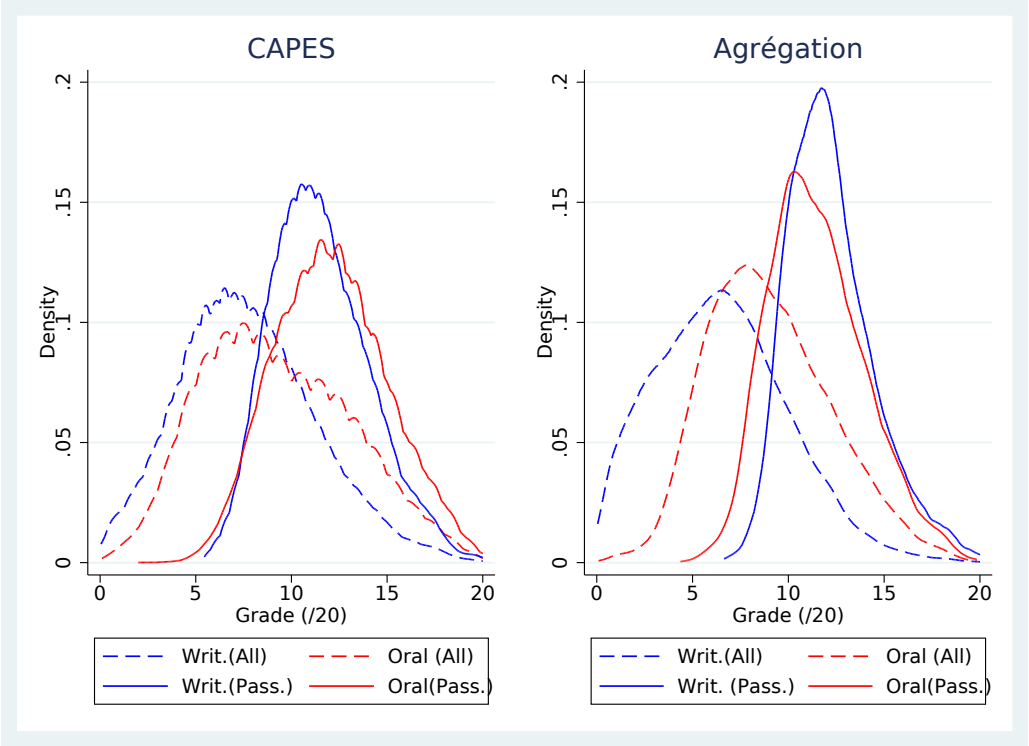
Notes: * p < 0.1; ** p < 0.05; *** p < 0.01. Robust standard errors clustered by teacher in parenthesis. The dependent variable is, for 9th grade, the teacher's student standardized test scores at the 9th grade national exam (Diplôme national du brevet) and for 12th grade, the teacher's student standardized test scores at the 12th grade national exam (Baccalauréat). In the first column, Agrégé is a dummy variable equal to one if the teacher has the Agrégation. For column 2 to 5, the evaluation grades are standardized. The certification grades are standardized by year, topic and level of certification (Agrégation vs. Capes). The pedagogical grade and the administrative grades are standardized according to their respective national grading table (cf. Table 5 and Table 8). Each line corresponds to a single regression, where all five evaluation grades are included jointly. The level of observation is teacher (topic) x student, from 2006 to 2012. The regressions are run on the sample as defined in the data appendix. All regressions include year fixed effects, topic fixed effects and the interaction between year fixed effects and topics fixed effects.

Table 19 – Impact of Classroom Observation on Teacher Performance

	(1)	(2)	(3)	(4)
Year relative to the year of inspection				
(Year prior inspection omitted)				
Year -3	-0.016	-0.027*	-0.002	0.006
	(0.007)	(0.010)	(0.012)	(0.012)
-2	-0.007	-0.007	0.007	0.016
	(0.007)	(0.010)	(0.010)	(0.008)
0	0.017***	0.013*	0.006	0.003
	(0.006)	(0.007)	(0.007)	(0.006)
1	0.028***	0.026***	0.013*	0.007
	(0.005)	(0.009)	(0.007)	(0.005)
2	0.031***	0.038***	0.011	0.008
	(0.006)	(0.010)	(0.007)	(0.004)
3	0.027***	0.041***	0.009	0.006
	(0.006)	(0.012)	(0.007)	(0.003)
Teacher-school fixed effect	No	Yes	No	Yes
Classroom-year fixed effect	No	No	Yes	Yes
Year x topic fixed effect	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes
Nb of observations	240,299	240,299	240,299	240,299

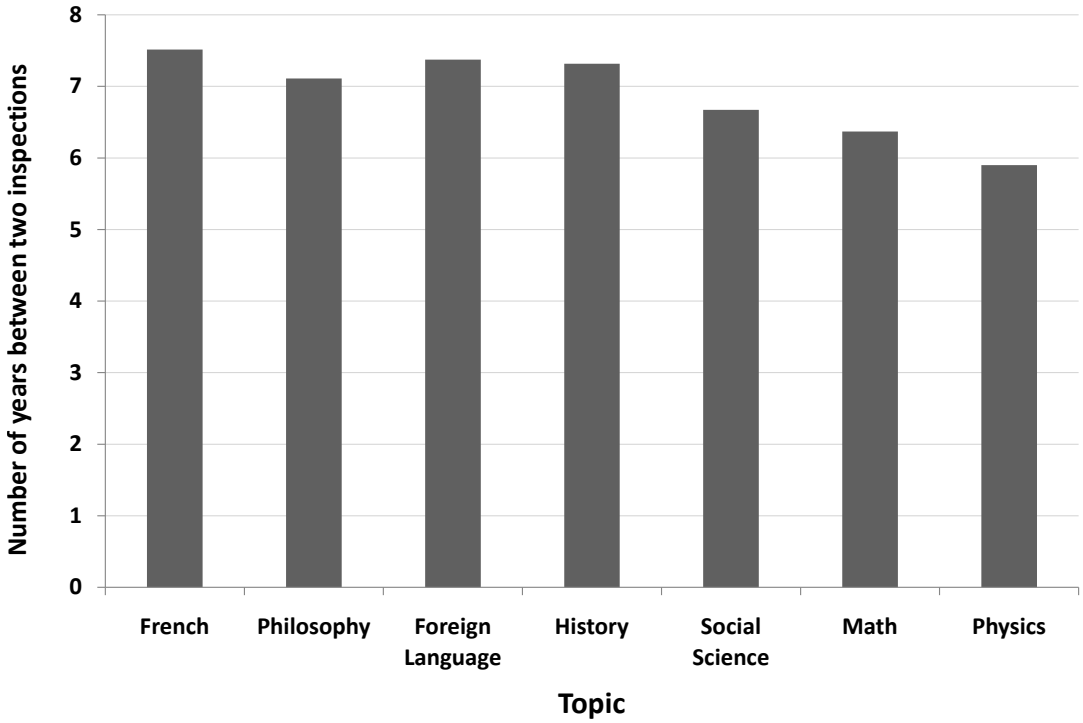
Notes: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.0$. Robust standard errors clustered by school in parenthesis. Each column corresponds to a single regression. Teacher performance is measured here through the average of her students' test scores at the 9th grade exam, by topic. The level of observation is teacher(topic) x classroom-year, from 2006 to 2012 and 9th grade teachers in French, Math and History. Controls includes teacher experinence dummies, share of students per parental occupation category, share of students receiving financial aid.

Figure 1 – Kernel Density of the Certification Grades (Written and Oral) for all Candidates and for Passing Candidates



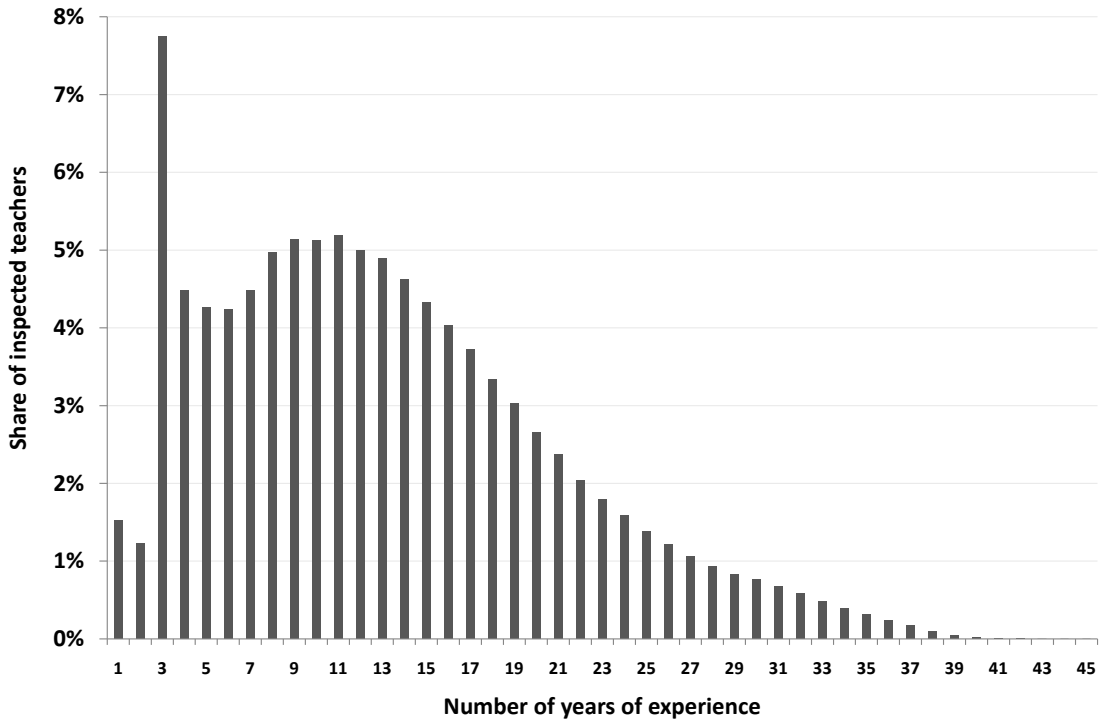
Notes: This figure plots the kernel density of the written exam grade (blue line) and the oral exam grade (red line) for all candidates (solid line) and for candidates who passed the exam (dotted line). The left graph plots the density for the Capes and the right graph plots the density for the Agrégation. The sample includes all the candidates who are present to the exam (see notes to Figure **??**) from 2002 to 2012, in Math, French and History.

Figure 2 – Average Number of Years between Two Inspections by Teaching Topic



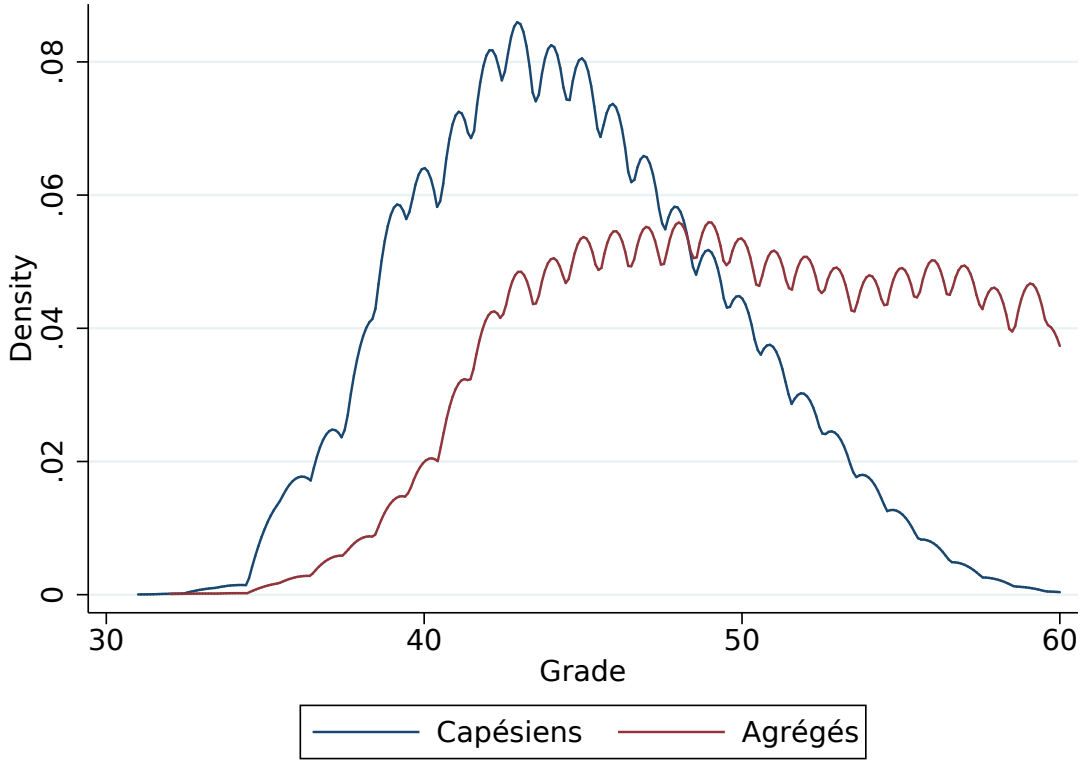
Notes: This figure plots the average number of years between two inspections, by teaching topic. The sample includes all active secondary school teachers, from 2004 to 2012, for which at least two inspections are observed over the 2004-2012 period. Both middle school and high school teachers are included in the sample, except for Philosophy and Social Sciences, because these topics are only taught in high school.

Figure 3 – Distribution of Inspections by Teachers' Number of Years of Experience



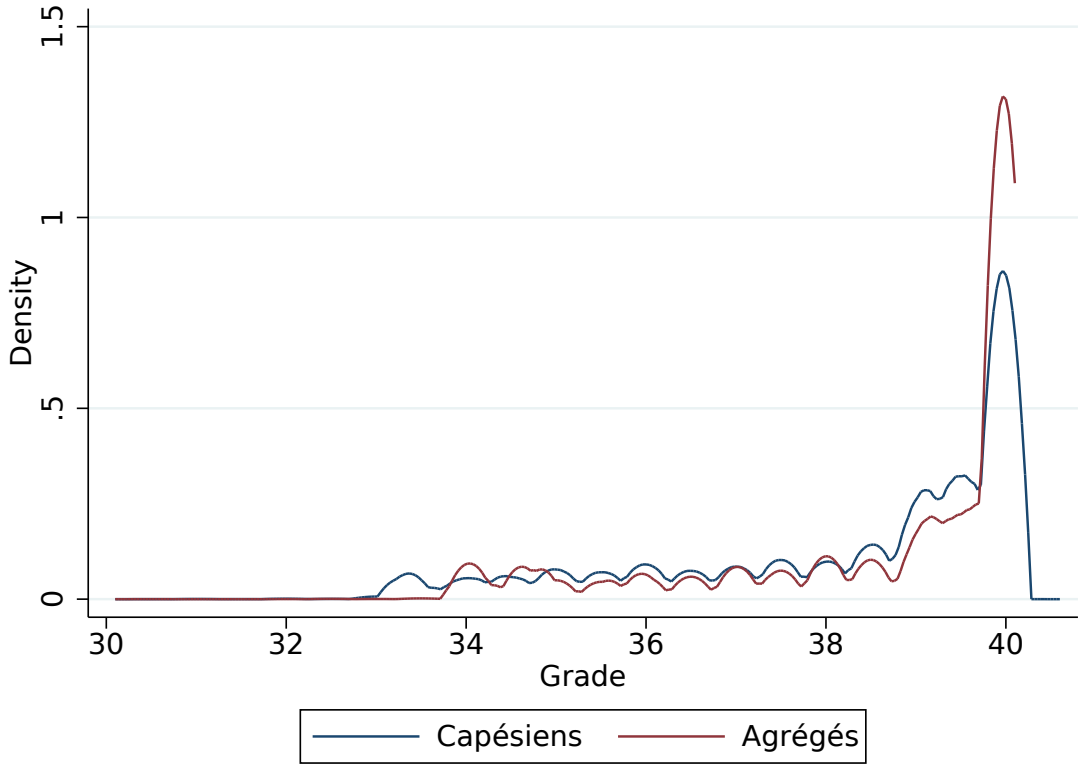
Notes: This figure plots the distribution of inspections by inspected teachers' year of experience. The sample includes all secondary school teachers (middle and high school), from 2004 to 2012 who are inspected at least over in the observed periode (2004 - 2012). The number of years of experience is defined as the number of year since entry in the teaching profession.

Figure 4 – Kernel Density of the Pedagogical Grade, by Level of Certification



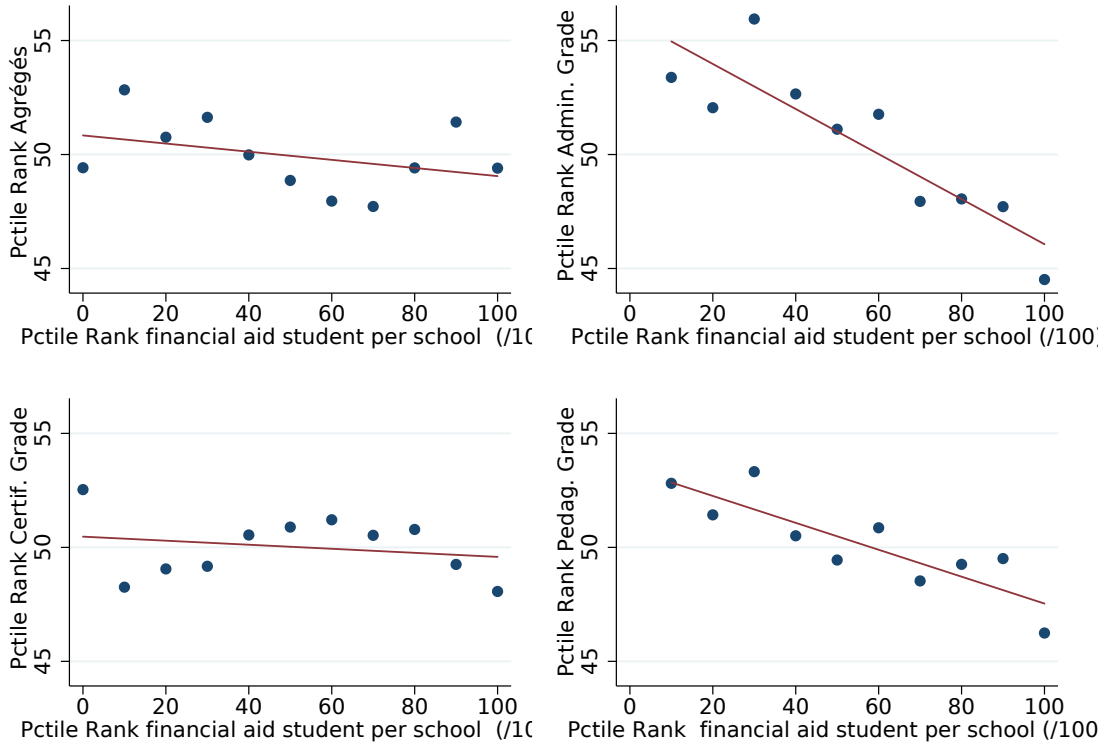
Notes: This figure plots the distribution of the pedagogical grade, without standardization. The blue line represents the distribution for teachers with the Capes (Capésiens) and the red line the distribution for teachers with the Agrégation (Agrégés). The sample includes all secondary teachers who are inspected at least once over the observed period (2004-2012).

Figure 5 – Kernel Density of the Administrative Grade, by Level of Certification



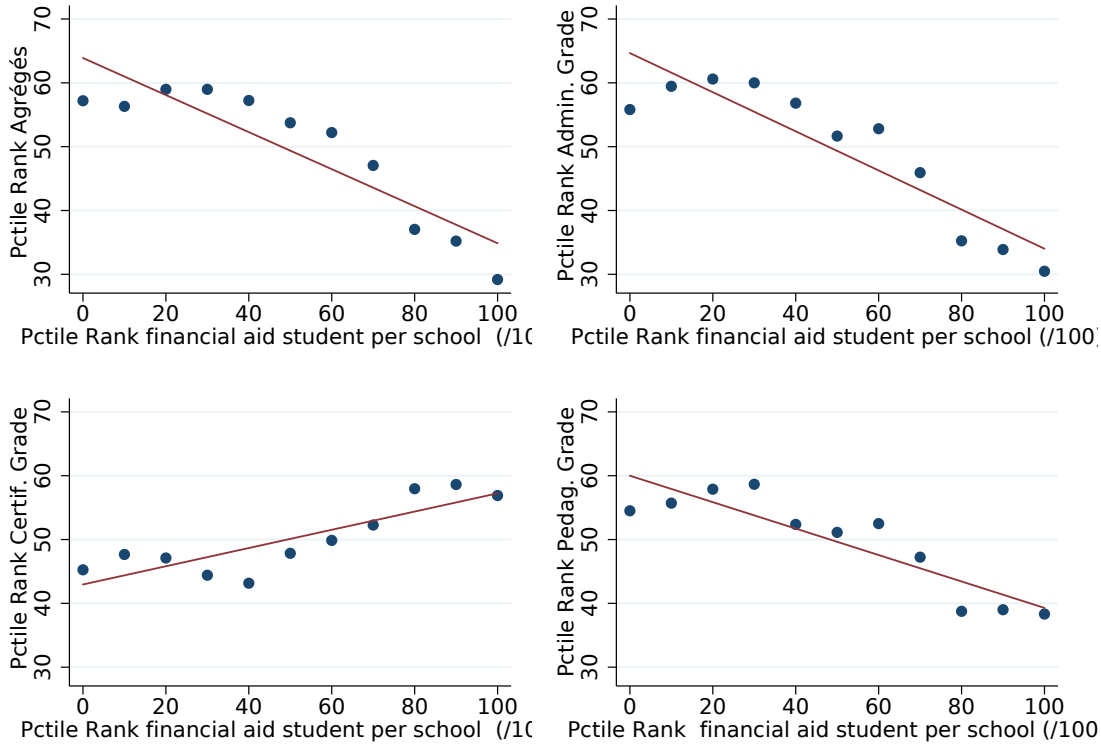
Notes: This figure plots the distribution of the administration grade, without standardization. The blue line represents the distribution for teachers with the Capes (Capésiens) and the red line the distribution for teachers with the Agrégation (Agrégés). The sample includes all secondary teachers over the observed period (2004-2012).

Figure 6 – Percentile Rank of the Evaluation Grades by Percentile Rank Share of Financial Aid Student per School – 9th Grade



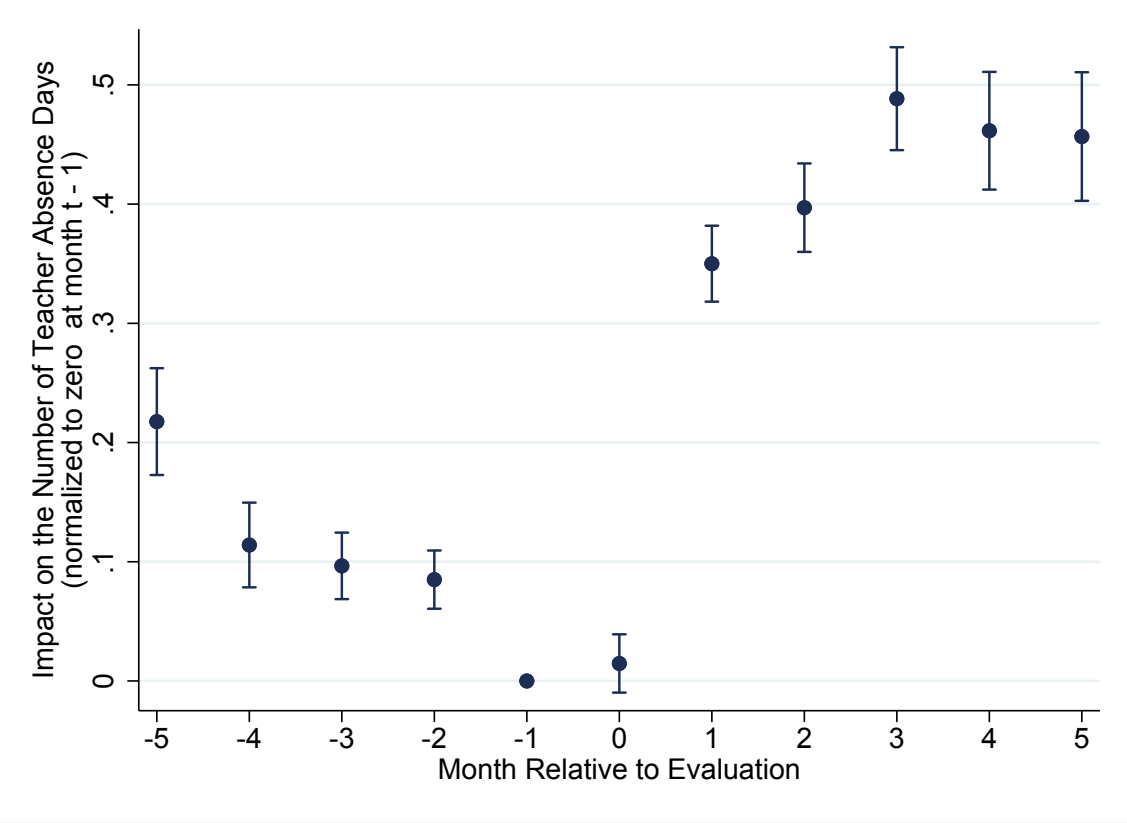
Notes: This figure plots the average share of *Agrégés*(ordered by percentile rank), the average percentile rank administrative, certification and pedagogical grades by the share of financial aid student per school (ordered by percentile rank). The sample includes all 9th grade teachers over the observed period (2004-2012).

Figure 7 – Percentile Rank of the Evaluation Grades by Percentile Rank Share of Financial Aid Student per School – 12th Grade



Notes: This figure plots the average share of *Agrégés*(ordered by percentile rank), the average percentile rank administrative, certification and pedagogical grades by the share of financial aid student per school (ordered by percentile rank). The sample includes all 12th grade teachers over the observed period (2004-2012).

Figure 8 – Impact of the Classroom Observation on Teacher Absence



Notes: This figure plots the impact of the classroom observation on the number of teacher absence days (zero included). This corresponds to a single regression. The specification includes teacher-school, topic, year and month fixed effects. The reference month is the month just before the evaluation. The level of observation is teacher x classroom x month x year. Robust standard errors are clustered by school.