There's Always Room for Improvement: The Persistent Benefits of a Large-scale Teacher Evaluation System - Online Appendix

Simon Briole & Éric Maurin

Paris School of Economics, France

Online Appendix A - Additional Tables and Figures

Descriptive statistics

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

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

Note: The table shows teachers' gross annual wage in euros in 2008 for each possible position on the wage scale as well as the total number of years of teaching experience needed to reach each level by promotion track. The 30% of teachers who get the best evaluation ratings can access the fast track (Grand Choix). The next 50% best evaluated teachers are promoted through the regular track (Choix). The 20% of teachers with the lowest ratings are promoted through the slow track, which corresponds to the minimal promotion speed based on experience. Source: *Décret n°72-581 du 4 juillet 1972 relatif au statut particulier des professeurs certifiés.*

Table A2: *Inspecteurs*' characteristics

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

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

Table A3: Distribution of between-evaluation spacing, by education region

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

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

Table A4: Student characteristics - difference between priority and non priority schools

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

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

Table A5: Teachers' characteristics

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

Note: The table refers to our working sample of teachers who teach 9th grade students between 2008-2009 and 2011-2012. It shows teachers' average number of years of teaching experience in 2008, proportion of female, type of school in 2008 and average number of external evaluations undertaken over the 4-year period under consideration. The first column refers to the subsample of math teachers whereas the second column refers to the subsample of French language teachers.

Table A6: Math teachers' evaluations and 9th grade teaching

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

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

Balancing tests - Tables and Figures

Table A7: Balancing test - 9th grade math teacher evaluation and student characteristics

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

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

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

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

Note: The table shows the results of regressing 9th grade classes' average characteristics (average age of students, proportion of girls, proportion from low-income families, proportion studying German and proportion studying Latin or ancient Greek) on a dummy indicating that their math teacher underwent an external evaluation in t and on a dummy indicating that they underwent an evaluation between 2008-2009 and $t-1$. Standard errors (in parentheses) are clustered at the teacher level. * p<0.10, ** p<0.05.

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

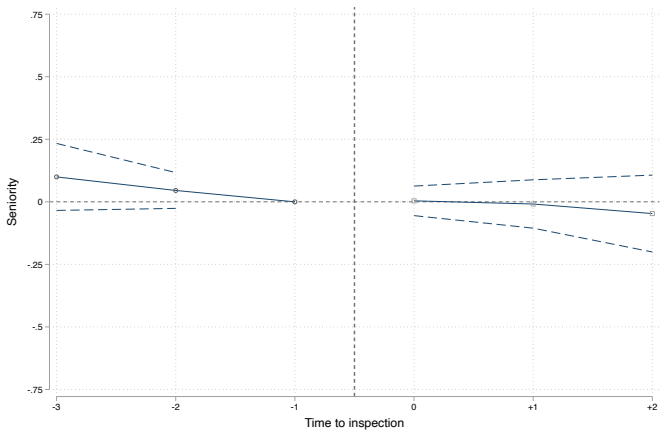
	(1) Teacher seniority	(2) Priority schools	(3) School performance	(4) Colleagues' experience	(5) Colleagues' seniority
All teachers (N=29156)	0.033	0.004	-0.002	-0.027	0.008
	(0.033)	(0.003)	(0.005)	(0.096)	(0.088)
Female teachers (N=15318)	0.060	0.002	-0.003	-0.120	-0.129
	(0.044)	(0.004)	(0.006)	(0.132)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

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

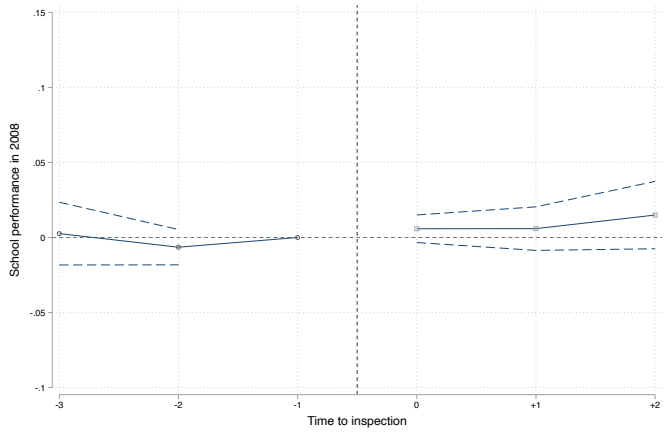
Table A10: Balancing test - 9th grade math teacher evaluation, teacher mobility and colleagues' characteristics II

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

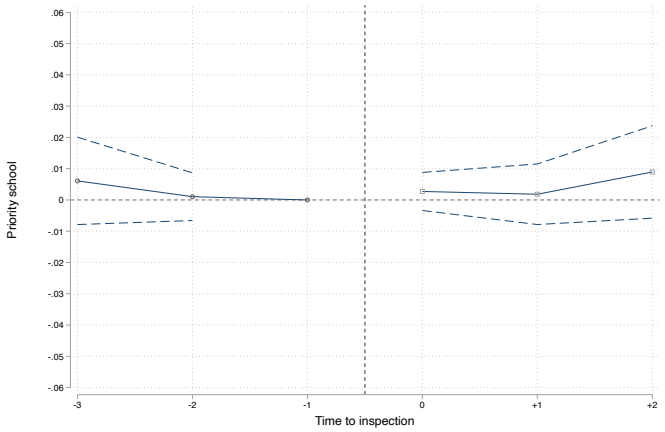
Note: The table shows the results of regressing teacher seniority, school characteristics (priority school, school performance) and colleagues' characteristics (experience, seniority) on a dummy indicating that the math teacher underwent an external evaluation in t and on a dummy indicating that she underwent an evaluation between 2008-2009 and $t-1$. School performance in column (3) is the average math test score in 2008 of the school in which the math teacher teaches in year t. Finally, colleagues' experience and seniority in columns (4) and (5) refer to the average characteristics of the 9th grade French language and history teachers who teach the same 9th grade students as the math teacher in year t. Standard errors (in parentheses) are clustered at the teacher level. * p<0.10, ** p<0.05.



(a) Seniority



(b) School performance



(c) Priority school

Figure A1: Math teacher evaluation and teacher mobility

Note: the solid lines in Figures A1 (a) to A1 (c) show the estimated difference between evaluated and non-evaluated math teachers before and after evaluations in terms of teacher seniority (a), school performance as measured by the school average math test scores in 2008 (b) and teacher probability to teach in a priority school (c). The dotted lines show 95% confidence intervals.

Robustness check - Goodman-Bacon Decomposition

Table A11: Robustness check - Goodman-Bacon Decomposition

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

Note: This table shows the average effects and weights for the two basic types of diff-in-diff (DD) variations used in this paper, namely those that compare treated and never treated teachers and those that compare groups of teachers treated at different point in time, using a Goodman-Bacon (**?**) decomposition. The table refers to the subsample of teachers who are observed at all periods between 2008-2009 and 2011-2012. Models include a full set of teachers, region and year fixed effects as well as controls for teachers' experience and seniority and for students' average age, gender, family social background, German language study and Ancient language study. Standard errors (in parentheses) are clustered at the teacher level. * p<0.10, ** p<0.05.

French language teacher evaluation and student performance

Table A12: 9th grade French language teacher evaluation and student performance by French language subtopic test scores and by subgroups

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

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

Math and French language teachers' external evaluations and student performance

Table A13: 9th grade math and French language teacher evaluation and student performance

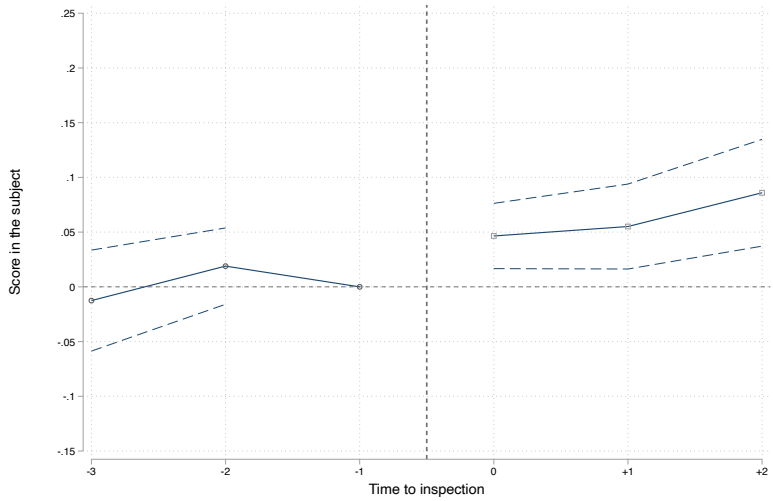
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A : Score in the subject						
Evaluation	0.038**		0.038**		0.040**	
	(0.012)		(0.012)		(0.010)	
Evaluation in t		0.029**		0.029**		0.032**
		(0.012)		(0.012)		(0.011)
Evaluation before t		0.057**		0.055**		0.055**
		(0.015)		(0.015)		(0.013)
Panel B : Score in other subjects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

Note: the table refers to the joint sample of math teachers and French language teachers who teach 9th grade students between 2008-2009 and 2011-2012. The first row of the upper (lower) panel shows the result of regressing their students' average standardized score in the subject they teach (subjects they don't teach) at the end of year t on a dummy indicating that they underwent an external evaluation between 2008-2009 and t. Rows 2 and 3 respectively show the result of regressing the same dependent variable on a dummy indicating that they underwent an external evaluation in t and on a dummy indicating that they underwent an evaluation between 2008-2009 and $t-1$. All regressions include the full set of teacher, region and year fixed effects. Columns (3) and (4) further include dummies for teachers' number of years of experience and seniority level. Columns (5) and (6) further controls for students' average age, gender, family social background, German language study and Ancient language study. Standard errors (in parentheses) are clustered at the teacher level. * p<0.10, ** p<0.05.

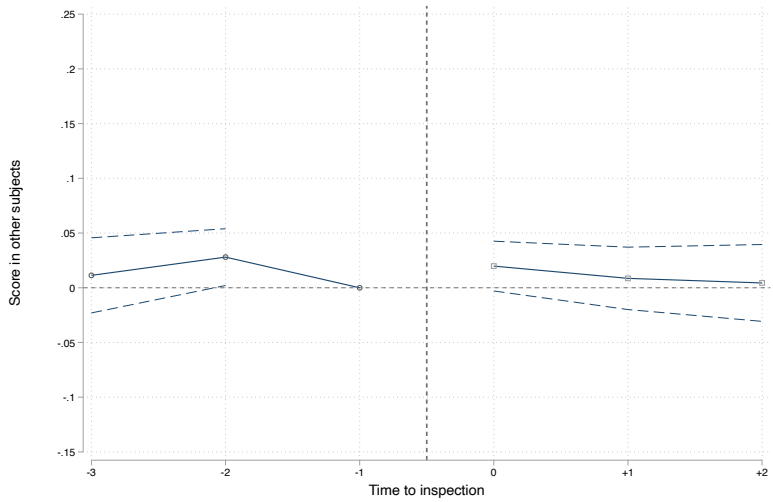
Table A14: Math and French language teachers' evaluations and student performance - by subgroups

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

Note: The table refers to the joint sample of math and French language teachers who teach 9th grade students between 2008-2009 and 2011-2012. The first (second) row shows the results of regressing their students' average score in the subject they teach (subjects they don't teach) at the end of year t on a dummy indicating that they underwent an external evaluation between 2008-2009 and t. The first column refers to the full sample, whereas columns (2) to (7) refer to subsamples defined by teachers' gender, number of years of teaching experience in 2008-2009 (above/below 11 years), and type of school attended in 2008-2009 (priority/non priority). Models include a full set of teachers, region and year fixed effects as well as controls for teachers' experience and seniority and for students' average age, gender, family social background, German language study and Ancient language study. Standard errors (in parentheses) are clustered at the teacher level. * p<0.10, ** p<0.05.



(a)



(b)

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

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

School-level analysis

Table A15: Balancing test - 9th grade math teacher evaluation and student characteristics, school level

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

Note: The table shows the results of regressing 9th grade students' school average characteristics (average age of students, proportion of girls, proportion from low-income families, proportion studying German and proportion studying Latin or ancient Greek) on the school proportion of math teachers who underwent an evaluation between 2008-2009 and t. The first row refers to the full working sample, whereas rows 2 and 3 refer to subsamples defined by type of school (priority vs non-priority). Standard errors are in parentheses. * p<0.10, ** p<0.05.

Between-evaluation spacing and the effect of evaluation

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

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

Note: The table shows the result of regressing 9th grade math teachers students' average standardized score in math at the end of year t on a dummy indicating that they underwent an external evaluation between 2008-2009 and t. Rows 2 and 3 respectively show the result of regressing the same dependent variable on a dummy indicating that they underwent an external evaluation in t and on a dummy indicating that they underwent an evaluation between 2008-2009 and $t-1$. The upper panel of the table refers to the 3 educational regions where the variance in between-evaluation spacing is minimal, and the lower panel refers to all other regions. All regressions include the full set of teacher, region and year fixed effects, controls for teachers' experience and seniority level, as well as controls for students' average age, gender, family social background, German language study and Ancient language study. Standard errors (in parentheses) are clustered at the teacher level. * p<0.10, ** p<0.05.

Online Appendix B - Data Construction

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

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

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

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

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

level database with information on math and French language teachers, and most notably with information on whether (and when) they underwent an external evaluation between 2008-2009 and 2011-2012[1].

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

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

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

Online Appendix C - Additional Robustness Checks

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

	(1) All	(2) Female	(3) Male	(4) Low-exp	(5) High-exp	(6) Priority	(7) Non Prior.
Math	0.042**	0.030	0.057**	0.053**	0.035*	0.079**	0.032**
	(0.014)	(0.019)	(0.020)	(0.020)	(0.019)	(0.030)	(0.015)
Humanities	0.009	-0.005	0.025	0.017	0.004	0.007	0.012
	(0.013)	(0.018)	(0.020)	(0.020)	(0.018)	(0.032)	(0.015)
Observations	31102	16492	14610	14319	16783	6475	24627

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

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

	(1) All	(2) Female	(3) Male	(4) Low-exp	(5) High-exp	(6) Priority	(7) Non Prior.
Science as	0.004**	0.000	0.008**	0.007**	0.001	0.008**	0.003
major field	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)	(0.004)	(0.002)
	[0.180]	[0.188]	[0.172]	[0.163]	[0.195]	[0.125]	[0.195]
Graduation	0.004**	0.001	0.008**	0.009**	0.001	0.008**	0.003
in science	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)	(0.004)	(0.002)
	[0.153]	[0.159]	[0.145]	[0.136]	[0.167]	[0.101]	[0.166]
Observations	31102	16492	14610	14319	16783	6475	24627

Note: The table refers to the same working sample of math teachers as Table 1, augmented by teachers with more than 25 years of teaching experience. The first row shows the result of regressing the proportion of their 9th grade students who will choose science as their major field of study at the end of 10th grade on a dummy indicating that they underwent an external evaluation between 2008-2009 and t. The second row shows the result of regressing the proportion of their 9th grade students who will graduate in science at the end of 12th grade on the same independent variable. The first column refers to the full sample, whereas columns (2) to (7) refer to subsamples defined by teachers' gender, number of years of teaching experience in 2008-2009 (above/below 11 years), and type of school attended in 2008-2009 (priority/non priority). Models include a full set of teachers, region and year fixed effects as well as controls for teachers' experience and seniority and for students' average age, gender, family social background, German language study and Ancient language study. Sample means of the dependent variables are within square brackets. * p<0.10, ** p<0.05.

Table C3: Robustness check - 9th grade math teacher evaluation and student performance by subgroups, without student controls

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

Note: The table refers to our working sample of math teachers who teach 9th grade students between 2008-2009 and 2011-2012. The first (second) row shows the results of regressing their students' average standardized score in math (humanities) at the end of year t on a dummy indicating that they underwent an external evaluation between 2008-2009 and t. Columns (1) and (2) refer to the subsamples of female and male teachers, columns (3) and (4) to the subsamples of teachers whose number of years of work experience is either above or below the median in 2008-2009 (i.e., above or below 11 years), columns (5) and (6) to the subsample of teachers who were in priority education schools in 2008-2009 and the subsample who were in non-priority schools. Models include a full set of teacher, region and year fixed effects as well as controls for teachers' experience and seniority. Standard errors (in parentheses) are clustered at the teacher level. * p<0.10, ** p<0.05.

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

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

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

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

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

Note: The table refers to our working sample of French language teachers who teach 9th grade students between 2008-2009 and 2011-2012. The first (second) row shows the results of regressing their students' average score in French language (mathematics) at the end of year t on a dummy indicating that they underwent an external evaluation between 2008-2009 and t. Columns (1) and (2) refer to the subsamples of female and male teachers, columns (3) and (4) to the subsamples of teachers whose number of years of work experience is either above or below the median in 2008-2009 (i.e., above or below 11 years), columns (5) and (6) to the subsample of teachers who were in priority education schools in 2008-2009 and the subsample who were in non-priority schools. Models include a full set of teacher, region and year fixed effects as well as controls for teachers' experience and seniority. Standard errors (in parentheses) are clustered at the teacher level. * p<0.10, ** p<0.05.