Good teaching and good grades. Can you buy pedagogy? *

Manon Garrouste[†], Ronan Le Saout[‡]

July 15, 2019

Abstract

This paper assesses students objectivity in their evaluations of teaching, by analyzing the relationship between their grades and evaluations, and the dynamics of evaluations over time. We exploit an original data set from almost 100 courses during 7 academic years in a French higher education institution. We use generalized additive model, teacher fixed effects, and instrumental variables estimations to rule out any simultaneity or endogeneity bias. We find that students take their exam grade into account when they evaluate teaching. A better grade is associated with a better evaluation of a teacher's pedagogy, although the size of the effect is relatively small. A one-point increase in by-course mean grade corresponds to a less than one percentage point decrease in the proportion of students giving bad evaluations. We also find that students give lower evaluations after the exam and higher evaluations after getting their grades.

Keywords : Student evaluation of teaching, Post-secondary education, Grades

JEL Codes : A23, I21, I23

*We would like to thank M. Bizien and X. Zhang for exploratory work. For useful remarks, we are grateful to R. Aeberhardt, A. Boring, V.-E. Brunel, X. D'Haultfoeuille, N. Jacquemet, F. Kramarz, M. Lafourcade, B. Marc, M. Tô, and participants to the JMA, AFSE, and Labex Osé conferences, and Ined and LEM seminars. Manon Garrouste acknowledges support from a Labex Phd grant, Labex iPOPs ANR-l0-LABX-0089, hosted by Ined, in partnership with heSam Université.

[†]*Corresponding author.* Univ. Lille, Cnrs, UMR 9221 - Lem - Lille Économie Management, F-59000 Lille, France. Email: manon.garrouste@univ-lille.fr. Postal address: Cité scientifique - Bâtiment SH2 - 59655 Villeneuve d'Ascq France

[‡]Ensai, Crest. Email: ronan.le-saout@ensai.fr

Introduction

A growing part of post-secondary education institutions use student evaluation of teaching (SET). According to a survey by Becker et al. (2012), 93% of American departments of Economics reported using SET to evaluate teaching in 2011. In France, this practice was introduced only recently and is much more heterogenous.[1] Many universities and institutions do not have such evaluations, while some have been organizing them for several years. Within universities, the use of SET varies a lot across departments; it is more prevalent in scientific departments and in vocational tracks than in general ones. Moreover SET seems to be more frequent in selective tracks.[2]

The primary objective of SET is for the administration to assess the quality of teaching. On average, it accounts for 50% of American instructors' teaching evaluation, according to Becker et al. (2012). Thus SET plays a significant role in teachers' hiring and promotion procedures. However, the validity of SET as a good measure of teaching quality is highly controversial. The prime question is to know whether students are able to correctly evaluate the quality of teaching. Student evaluations seem to be a reliable measure, in the sense that inter-rater reliability is high (i.e. evaluations of the same course by two different students are highly correlated). They are stable over time, and they are relatively highly correlated with other measures of teaching quality (self-evaluation, peer evaluation, etc. Marsh and Roche, 1997). Moreover, student evaluations perform better than objective characteristics such as teacher's salary and status in explaining students' achievement (Hoffmann and Oreopoulos, 2009). But they may not measure all the dimensions that constitute good teaching (d'Apollonia and Abrami, 1997; Greenwald, 1997), they are biased by characteristics that are not related to teaching quality (Boring et al., 2016; Boring, 2017; De Witte and Rogge, 2011), and they are only weakly related to learning (Beleche et al., 2012; Boring et al., 2016; Braga et al., 2014; Carrell and West, 2010; Ewing, 2012; Isely and Singh, 2005; Krautmann and Sander, 1999; Uttl et al., 2017; Weinberg et al., 2009).

From a theoretical point of vue, SET can be seen as a principal-agent-client framework in which teachers' efforts and pedagogical qualities cannot be fully observed by the administration (Klitgaard, 1990). Teachers can get good evaluations

[1] A 1997 decree stipulates that every French higher education institution delivering a university diploma should evaluate teaching, and that this evaluation should take students judgment into account.

[2] So far and to our knowledge, there is no quantitative evaluation of the use of SET in French higher education institutions. For a qualitative review, see for instance the report for the "Haut Conseil de l'évaluation de l'école", Dejean (2002).

either by actually improving teaching or by "corrupting" (i.e. giving good grades in order to get good evaluations). Braga et al. (2011) consider, for instance, a model in which teachers choose between two types of teaching activities: real teaching or teaching-to-the-test. Since the latter requires less effort from students than the former, good teachers may receive bad evaluations if teaching-to-the-test is effective. An alternative theory (Franz, 2010) is that teachers may give good grades to prevent students from pestering for better grades.

Thus, it is difficult to assess a causal link between grades and student evaluations. First, good teachers should cause good grades and get higher evaluations at the same time. In that case, grades and evaluations are simultaneously determined. Second, students may self-select into courses they like and they are good at, and consequently they may give higher evaluations (or, on the contrary, they may evaluate more severely if their learning expectations are high). This would result in a selection bias. Third, unobserved teachers' characteristics (such as charisma, clarity, enthusiasm, etc.) are undoubtedly related to students' achievement and evaluations at the same time, which may create endogeneity.

In this paper, we want to assess students' objectivity in evaluating teaching quality, by analyzing the effect of grades on SET. If student evaluation was objective, there should not be any significant effect of their exam grades on their evaluations, once any simultaneity or selection bias is ruled out. If this is the case, at least two reasons are invoked in the literature to explain such a link. Students may infer course quality from received grades (this is the "attribution" theory of Greenwald and Gillmore, 1997). Or students may give good evaluations as a reward for good grades or an easy exam ("leniency" theory). Using an original data set of individual grades and evaluations from a French higher education institution, we ask whether students do account for their grades when evaluating the quality of teaching. We contribute to the literature in two ways. First, to our knowledge, this is the first study to analyze the causal link between grades and SET in the French context.[3] Second, compared with other data sets, we know the exact date when students evaluate each course, so that we are able to analyze the dynamics of evaluations over time. More precisely, we are able to study how evaluations change over time, when students get more information on expected grades (when they take the exam) and when they learn of their grades. Again, if students were objective, revealing information on grades should not have any effect on their evaluations. We use the traditional methods proposed in the literature, i.e. teacher's fixed effects and instrumental

[3]We are aware of only one data set (Boring, 2017; Boring and Philippe, 2017), used to analyze gender biases in student evaluations of teaching.

variables, as well as a generalized additive model estimation. We find that students give teachers higher evaluations when they get higher grades, but the size of the effect is small: a one point increase in by-course mean exam grade corresponds to less than a one percentage point decrease in the proportion of students giving bad evaluations. We also find that students take contemporaneous information into account when evaluating teaching. They give lower evaluations after the exam, and higher evaluations after getting their grades.

The paper is organized as follows. We present the data in Section 1. Section 2 gives some descriptive statistics. Section 3 presents the empirical strategy and gives the results. We conclude in Section 4.

1 The data

We use data from a French "*grande école*", which provides a three-year graduate-level program in the fields of statistics, economics, finance and actuarial science. The first two years consists in the same basic training for all students. In the third year, students specialize into one particular track.

Since the 2004-2005 school year, each student has been asked to evaluate the courses they attend. Students fill out an online form, which is the same for every course. Evaluation is completely anonymous. The evaluation form consists of seven questions about the course (interest, difficulty, student attendance, teacher's pedagogy, formalization, progression and quality of course material), and five about tutorial classes when they exist (difficulty, student attendance, teaching assistant's pedagogy, number of tutorials, relevance to the course). Questions regarding the course material, the number of tutorials and the relevance to the course were removed from this analysis. For each question, students have to choose between three to four ranked categories. The detail of questions and categories is given in Table 4 of the appendix. Students can also give free comments.

In France, student evaluation of teaching has been introduced only recently and is not commonly used to evaluate teachers. For now, recruitment and promotion procedures do not depend on SET at all. Teachers are usually civil servants, and they are recruited on the basis of a national competitive examination that does not take their previous evaluations into account. Teacher pay is based on a national salary scale, and does not depend on SET. However, each higher education institution administration may use SET to organize teaching. In the school we study, the administration may change a teacher from a particular course if student evaluations are poor for several consecutive years. Note that the use of SET by the admninistra-

tion may differ for teachers and for teaching assitants. Because teaching assistants are not permanent staff of the school, the administration may more easily choose not to keep them if their evaluations are poor.

We use three anonymous data sets. First we observe every course given in the school from the 2004-2005 to the 2010-2011 academic year. We excluded from our analysis some types of courses: courses without grades, language courses, and collective projects, for which there is no specific teacher. Furthermore, the form is not suited for courses taught by several teachers, so we excluded them. We also excluded small classes (less than 10 students registered), or courses which were given for only one or two academic years. In the end, we observe 97 courses, from 2004-2005 to 2010-2011, for a total of 485 observations. This is an unbalanced panel: not all courses in the sample are given every year, but they are all given at least two years. Among them, 17 are given in the first year of the school program, 27 in second year and 53 in third year. This reflects increasing specialization over the program. 37 courses have tutorial classes, 60 do not. Overall, we observe 128 individual teachers and 291 teaching assistants. Every teacher gives between 1 and 5 courses per school year and they give 1.3 courses a year on average. Every teaching assistant gives tutorial classes for 1 to 4 courses each year, with 1.5 on average. Courses are split into themes: 21 in Economics, 12 in Social Sciences, 21 in Finance and Actuarial Science, 16 in Mathematics and Computer Science, and 27 in Statistics and Econometrics.

Second, we observe individual evaluations of each course. This corresponds to 17,560 individual observations. Note that the data are individual but anonymous; we cannot link different evaluations made by the same student. We observe, though, some individual characteristics, namely the way students entered the school (competitive examination or direct admission)[4], and the date when they evaluate courses. We also observe course-specific characteristics: whether it takes place in year 1, 2, or 3, the major, one identifier for the course, and one for the teaching assistant. The number of bad evaluations (category 1) is small (see Table 5 in the appendix). On average, 8% of students find that the teacher's pedagogy is bad. 19% rate it as fair, 49% as good and 24% as very good.

Third, we have a data set of individual grades for each course, containing 24,198 individual observations. Again, this data set is anonymous; we cannot link different grades obtained by the same student for different courses. Students may be graded through two different types of evaluations, depending on the course. First, they may have to take a written exam at the end of the semester. The exam is usually prepared

[4]Students are admitted through different processes and at different stages: students are admitted to either the first year through a competitive examination (Mathematics or Economics major), or directly to the second or third year on the basis of academic qualification.

and corrected by the teacher, sometimes with the help of teaching assistants. In the course panel, 359 observations out of 485 are evaluated through a written exam. Second, students may get a continuous assessment grade, by handing in homework, or a project, or by attending tutorial classes. If there is no written exam, then the final grade is given by the teacher, on the basis of a continuous assessment evaluation. This is the case for 126 observations out of 485. If there is a written exam and continuous assessment (112 observations out of 485), then the final grade is a weighted average of both grades. In this case, the continuous assessment is usually done by the teaching assistant, whereas the exam is graded by the teacher. For each individual observation in the grades data set, we know the individual written exam grade and continuous assessment grade, as well as the grade after resit exams, if any. We observe a student's way of admission, one identifier for the teacher and one for the teaching assistant. Grades are different across courses; mean grades in Social Sciences courses are higher than in courses with more technical content (Table 6 in the appendix).

The degree delivered at the end of the studies is the same regardless of the grades students get, and is a positive signal on the French job market. In that sense, grades may matter less than in other higher education institutions, where students may usually get distinctions depending on their grades. However, obtaining the school's degree is subject to some rules regarding students' grades. The degree is delivered at the end of Year 3, but in order to pass each year, students have to obtain an average of 12 over 20 during the year and to get at least a 6 in some core subjects. Students may repeat one year if they fail to meet these rules, but they cannot repeat more than one year during their studies. Furthermore, students are allowed to, and often do, attend another Master during Year 3 of the school. These students usually try and attend highly selective Master's programs, and thus need to get very good grades during the first two years of the school. Similarly, students who wish to continue with a Ph.D. may need good grades to get funding.

Because data sets are anonymous, we cannot link individual grades and evaluations. Thus our analysis is made at the aggregate teacher-year-subject level, which is in line with what is done in the literature. It would be possible to work at a more detailed level, by using the teaching assistant identifier. The number of observations would be larger. Moreover, as tutorial classes are formed almost randomly[5], there would be less worry concerning a potential selection of students within classes. However, contrary to the teacher's name, the teaching assistant's name is self-declared

[5]Depending on the school year, students are allocated to tutorial classes using alphabetical order, or a more random allocation.

by students and data is of poor quality. More precisely, students have to select the name of the teaching assistant from a drop-down menu, and the first name appearing in the menu systematically has many more evaluations than the others. For this reason, we chose not to work at the teaching assistant-year-subject level.

At the end of each semester, students can evaluate the courses they attended whenever they want by filling out the online form. The exact date when each evaluation is made is observed. Whatever the school year, there is a clear mode in the number of evaluations over time (see Figure 3 in the appendix). We assume that this mode corresponds to the date when grades are released. For each course, we are thus able to define the (unobserved) date when grades were released as the mode of the dates of individual evaluations. Because we know the date of the final written exam for each course, we are then able to study the dynamics of evaluations, according to the level of information students have on their expected grade.

Student evaluation of teaching was optional up to 2007-2008 and could be done after the final grades were released. From the 2008-2009 school year, it became compulsory and grades are now obtained only after having filled out the form. Thus, students who completed the evaluation before 2008 are likely to have specific characteristics. The student response rate is not observed, but, linking evaluations and grades at the teacher-year-subject level, we can observe the number of evaluations over the number of grades for each course. This ratio is 45% on average before 2008. After 2008, it is 100% on average. Moreover, although we do not observe any large differences in mean grades before and after 2008 (Table 6 in the appendix), mean evaluations significantly differ before and after 2008 (see Table 7 in the appendix). Before 2008, students report being more present on average. They rate the teacher's pedagogy to be lower, and they rate the speed of progression higher. They also write comments more often. Students evaluating before 2008 are more often in Year 2 or 3 of the program, and they entered the school more often through direct admission. In the following estimations, we will thus be careful about controlling for before and after 2008.

2 Descriptive statistics

Correlations between the different dimensions of teaching evaluated by students (Table 8 of the appendix) highlight two groups of variables: teacher's pedagogy, interest for the subject, and presence are positively correlated with one another and negatively correlated with the difficulty of the course, formalization, and speed of progression. Mean grades are positively correlated with the dimensions of the

first group and negatively correlated with the variables of the second group. The same results emerge from a principal component analysis (Figure 5 of the appendix), where the first two axes explain around 70 % of the total variance. The two groups appear clearly on either side of the first axis in the aggregate level analysis.

Figure 4 of the appendix confirms the direction of correlations between mean grades and evaluations. One more point in mean grade corresponds to 0.04 point more in the mean evaluation of teacher's pedagogy (and to 0.03 point more in the evaluation of the interest for the subject and in attendance). When the course progresses too fast, when it is too formalized, or when it is too difficult, mean grade is smaller. Those effects are as expected, but they are quantitatively small, about one tenth of a standard error. This is partly due to the fact that evaluations are categorical ordered variables.

Unsurprisingly, variance analysis in Table 9 of the appendix shows that most part of the variance in evaluations and grades comes from between students variation and that inter-course variance is small.

In the following sections, we will consider teacher's pedagogy as the dependent variable. Potential determinants of by-course mean evaluation of teacher's pedagogy are presented in Table 1. The first column gives the ordinary least squares estimates of the regression of mean teacher's pedagogy on course characteristics (observations are clustered at the course level). Students' mean evaluation of teacher's pedagogy is negatively correlated with the number of students in the class. This is in line with the literature considering that the teacher has less time to devote to each student when the class is bigger. The fact that the course is evaluated through a written exam is positively associated with the mean evaluation of teacher's pedagogy. The domain of the course (Social Sciences, Finance and Actuarial Sciences, Mathematics and Computer Sciences, or Statistics and Econometrics; Economics is the reference) is not significantly correlated with the mean evaluation of teacher's pedagogy. As already mentioned, mean evaluation is higher after 2008 than before. The second column of Table 1 adds observable characteristics of the teacher as regressors. As expected, no teaching experience, as measured by a dummy for the first year of teaching the course, is negatively correlated with the mean evaluation of the level of pedagogy. Whether the teacher is a man or a woman is not significantly correlated to the mean evaluation he or she gets (contrary to the results of Boring, 2017). Columns 3 and 4 add mean students' grades as covariates. The mean final grade is positively associated with the mean evaluation of teacher's pedagogy, which is actually driven by a positive correlation with the mean exam grade. Mean continuous assessment grade, when there is also a written exam, is positively but not significantly associated

with the mean evaluation. When there is no written exam, i.e. when the continuous assessment grade is the only way in which students are graded, then the correlation is negative, although not significant. Column 5 gives mean students' evaluations of their interest for the subject as an additional regressor. As expected, the correlation is significantly positive. Column 6 shows that the mean evaluation of teacher's pedagogy the preceding year is positively correlated with the current year mean evaluation.

In the following sections, we will further analyze the relationship between grades and evaluation of teachers' pedagogy. The teacher's pedagogy item seems to us as the best proxy of how students evaluate teaching quality. And, in practice, this is the main dimension that the school administration uses to assess the quality of teaching.

	Teacher's pedagogy					
	(1)	(2)	(3)	(4)	(5)	(6)
Nbr students	-0.003**	-0.003**	-0.003**	-0.003**	-0.001	-0.001
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Written exam	0.100*	0.090	0.254***	-1.198	-1.607**	-1.105
	(0.06)	(0.06)	(0.08)	(0.75)	(0.67)	(0.67)
Social sciences	0.041	0.024	0.009	-0.008	-0.098	-0.166
	(0.13)	(0.14)	(0.13)	(0.12)	(0.13)	(0.11)
Finance	-0.044	-0.036	-0.024	-0.025	-0.114	-0.105
	(0.13)	(0.13)	(0.13)	(0.13)	(0.12)	(0.10)
Maths, Computer	0.012	0.011	-0.004	0.003	0.041	-0.022
	(0.12)	(0.12)	(0.12)	(0.12)	(0.11)	(0.09)
Stats, Econometrics	0.006	0.015	-0.002	0.025	0.024	0.030
	(0.13)	(0.13)	(0.13)	(0.13)	(0.11)	(0.08)
After 2008	0.124*	0.130	0.121	0.117	0.103	0.135**
	(0.06)	(0.08)	(0.08)	(0.08)	(0.06)	(0.05)
Intercept	2.828***	2.834***	1.762***	3.124***	1.492**	0.636
	(0.17)	(0.19)	(0.32)	(0.71)	(0.67)	(0.72)
First year teaching		-0.188***	-0.184***	-0.174***	-0.133**	-0.072
		(0.07)	(0.06)	(0.06)	(0.07)	(0.08)
Male prof		0.047	0.047	0.041	-0.011	0.001
		(0.09)	(0.09)	(0.09)	(0.11)	(0.11)
Mean final grade			0.073***			
			(0.02)			
Mean exam grade				0.081***	0.056***	0.050***
				(0.02)	(0.02)	(0.01)
Mean cont. assess. with exam				0.007	-0.005	-0.004
				(0.01)	(0.01)	(0.01)
Mean cont. assess. without exam				-0.020	-0.065	-0.034
				(0.05)	(0.04)	(0.04)
Mean interest					0.774***	0.654***
					(0.08)	(0.09)
L.Mean pedagogy						0.270***
						(0.06)
Dummies school grade	Yes	Yes	Yes	Yes	Yes	Yes
Dummy direct admission	Yes	Yes	Yes	Yes	Yes	Yes
R2	0.07	0.10	0.13	0.14	0.39	0.48
Nb obs	485	485	485	485	485	373
Nb clusters	97	97	97	97	97	97

Table 1 – OLS estimations of potential determinants of teacher's pedagogy mean evaluation

Note: * ($p < 0.10$), ** ($p < 0.05$), *** ($p < 0.01$). The unit of observation is a triplet teacher-course-year. Observations are weighted by the number of students in the unit. Standard errors in brackets are clustered at the course level.

Reading note: On average in the sample, the teacher's pedagogy is rated 2.828 out of 4 before 2008 and 2.952 (=2.828+0.124) after 2008. This difference is significant at the 10% significance level.

3 Empirical strategy and results

3.1 Empirical strategy

When analyzing the effect of grades on the evaluation of teaching, we first need to address the simultaneity issue. Better teaching leads to better grades and better evaluations at the same time. In order to better understand the link between evaluations and grades, let us consider the following model:

$$Eval_{icjt} = \alpha_0 + \alpha_1 Grade^e_{icjt} + \alpha_2 X^1_{it} + \alpha_3 Y^1_{ct} + q_{cj} + u_{icjt} \quad (1)$$

where subscript i corresponds to the student, c corresponds to the course, j corresponds to the teacher, and t corresponds to the school year. $Eval_{icjt}$ is thus student i's evaluation of teacher j's pedagogy for course c in school year t. $Grade^e_{icjt}$ is the grade student i expects to receive from teacher j for course c in school year t. X^1_{it} and Y^1_{ct} are student and course specific observed characteristics respectively. q_{cj} is the unobserved pedagogical qualities of teacher j teaching course c. Error terms u_{icjt} are self-correlated within a course. If student evaluations were purely objective (and if they had complete information on teaching quality), α_1 would be 0, i.e. students' evaluations would depend only on observed and unobserved course characteristics, and not on their expected grades. In other words, students would not react to their grades when evaluating the course. We want to know whether this is true or not. α_1 is thus our parameter of interest.

The difficulty comes from the fact that student i's expected grade is also explained by observed and unobserved individual and course characteristics:

$$Grade^e_{icjt} = \gamma_0 + \gamma_1 X^2_{it} + \gamma_2 Y^2_{ct} + q_{cj} + \gamma_i + v_{icjt}$$

with γ_i an unobserved student fixed effect corresponding to student i's individual ability. We assume that u_{ij} and v_{ij} are not correlated.

This equation shows that we also need to take possible endogeneity issues into account, due to the fact that teacher's quality and student's ability are unobserved and potentially correlated with observed variables. More formally, there is a correlation between $Grade^e_{icjt}$ and the error term $q_{cj} + u_{icjt}$ in equation (1), due to student's fixed effect, teacher's fixed effect, and possible student's selection into courses.[6]

Again, note that, because data are anonymous, we cannot link individual grades

[6] $E[Grade^e_{icjt}(q_{cj} + u_{icjt})] = E[(\gamma_0 + \gamma_1 X^2_{it} + \gamma_2 Y^2_{ct} + q_{cj} + \gamma_i + v_{icjt})(q_{cj} + u_{icjt})] = E(q^2_{cj}) + E(\gamma_i u_{icjt})$. The correlation is the sum of a teacher's quality effect and a students' selection into course effect.

and individual evaluations. In the following, we will then consider mean evaluations and mean grades by teacher-course-year. We will note $\overline{x}_{cjt} = \frac{1}{N_{cjt}} \sum_{i=1}^{N_{cjt}} x_{icjt}$, where N_{cjt} is the number of students enrolled in course c, with teacher j, in school year t. We then consider the pseudo-panel model aggregated at the teacher-course-year level:

$$\overline{Eval}_{cjt} = \alpha_0 + \alpha_1 \overline{Grade^e}_{cjt} + \alpha_2 \overline{X^1}_{cjt} + \alpha_3 Y^1_{ct} + q_{cj} + \overline{u}_{cjt} \quad (2)$$

Note that the errors are inherently heteroscedastic in this model. In order to get efficient estimates, regressions are weighted by the number of students per course.

Let us consider that students know their grade or at least have a good expectation of it, so that their expectations are correct on average, i.e. $\overline{Grade^e}_{cjt}$ is replaced by $\overline{Grade}_{cjt}$ in equation (2).

Our first identification strategy consists in estimating equation (2) with teacher-course fixed effects. Identification is then based on variations from one year to another in a given course taught by a given teacher. This is manageable only for teacher-course pairs that we observe several times. Moreover, for grades not to be correlated with the error term, we have to assume that i)$\overline{\gamma}_{cjt}$ is independent from $\overline{u}_{cjt}$, i.e. course average students' ability is independent from evaluations unobserved characteristics, and that ii)$\overline{v}_{cjt}$ is independent from $\overline{u}_{cjt}$, i.e. there is no common idiosyncratic shock affecting both grades and evaluations. Assumption i) is valid if course average students' ability is constant over years. This seems realistic, at least for compulsory courses, or if course characteristics are properly controlled for. Assumption ii) seems realistic in general. It would not be valid if, for instance, the course took place at some very inconvenient time of the day, which would affect both students' learning and evaluations.

Our second identification strategy consists in estimating equation (2) with two stage least squares, by instrumenting mean grade $\overline{Grade}_{cjt}$. We thus need to choose instruments that would explain mean grade but not evaluations directly. A classical choice in the literature is to use mean grade the preceding year, that we will note $\overline{Grade}_{cj,t-1}$ (provided that the teacher is the same in t and $t-1$). However, in our setting, the exclusion restriction is not very credible because lagged mean grade contains the unobserved quality of teacher. Thus, we propose to use the lagged mean grade only when the teacher changes between $t-1$ and t. In this case, mean grade the preceding year is more likely to be uncorrelated with $\overline{u}_{cjt}$. We will also use the proportion of resit exams as an instrumental variable. This is a good predictor of mean grade, that should not be correlated to unobserved characteristics of mean evaluation.

3.2 Results

Table 2 presents the estimation results using ordinary least squares (OLS), teacher's fixed effects, and two-stage least squares (2SLS). The first outcome we consider is the mean evaluation of teacher's pedagogy by students enrolled in the course. In order to further analyze the distribution of evaluations, two other outcomes are considered, namely the proportion of students giving very good evaluations, and the proportion of bad evaluations. The explanatory variables of interest are the course mean written exam grade, and the course mean continuous assessment grade, interacted with a dummy which equals one if there is a written exam and zero otherwise.

The null hypothesis we want to test is that students do take their grades into account when evaluating teacher's pedagogy. If this is true, then the coefficient of mean exam grade should be significant, even after correcting for endogeneity. The corollary is that the mean evaluation of teacher's pedagogy should not depend on the mean continuous assessment grade when there is a written exam, since, in this case, continuous assessment is made by teaching assistants. On the contrary, when there is no written exam, the mean evaluation of teacher's pedagogy should depend on the mean continuous assessment grade.

Regressions are controlled for course characteristics and teacher characteristics. Course characteristics are the domain (Economics, Social Sciences, etc.), whether the course is given in year 1, 2 or 3 of the school program, and whether this is a remedial course for students entering directly in year 2 or 3. Teacher characteristics are a dummy for men, and a dummy for the first year of teaching. Because we are concerned by a potential selection issue in evaluations before 2008, a dummy for observations after 2008 is also added as a covariate.

The first column of Table 2 shows a significantly positive correlation between students' mean evaluation of teacher's pedagogy and the mean grades which are given by the teacher. When there is a written exam, mean continuous assessment grade is negatively, though not significantly, correlated with the mean evaluation of teacher's pedagogy. These naive estimates are thus in line with the hypothesis that students do take their grades into account when evaluating teaching. More precisely, a one point increase in mean exam grade would correspond to a 0.077 point increase in the mean evaluation of teacher's pedagogy, going from 1.803 on average to 1.880. However, this estimate does not take the endogeneity of mean grades into account. When endogeneity is controlled for, using teacher-course fixed effects in column 2, and using instrumental variables in column 3, the coefficient of mean exam grade remains positive and significant. The size of the effect is a bit larger when using instrumental variables. A one point increase in mean exam grade corresponds to

a 0.086 point increase in students' mean evaluation of teacher's pedagogy. The coefficient associated to mean continuous assessment grade without a written exam is larger (0.094) and significant in the fixed effect specification, but it is not significant in the instrumental variables one. When there is a final written exam, the effect of mean continuous assessment grade on teacher's pedagogy mean evaluation can be seen as a Placebo test, as in this case the continuous assessment grade is not given by the teacher. The fact that the coefficient remains not significantly different from zero is thus reassuring.

The positive effect of grade on mean evaluation seems to be driven by a smaller proportion of bad evaluations. A one point increase in mean continuous assessment grade, when given by the teacher, corresponds to a 0.95 percentage point decrease in the proportion of bad evaluations that the teacher gets. The effects of grades on the proportion of very good evaluations of teacher's pedagogy are not significantly different from zero, in either specifications (except for continuous assessment grade with exam in the OLS specification).

In the 2SLS specification, mean grade the preceding year, and percentage of resit exams seem to be valid instruments for current mean grade. According to Table 3, the coefficients of lagged mean grades are highly correlated to corresponding current mean grades in the first stage estimations. When there is a written exam, the proportion of resit exams is also a significant determinant of mean grade. Furthermore, the test for joint significance ("F weak" statistic) rejects the weak instruments hypothesis.

Again, the exclusion restriction is unlikely to hold for lagged mean grade. Tables 10 and 11 in the appendix thus replicate the preceding results, but the two stage least squares are estimated only on observations for which the teacher changed between $t-1$ and t. First stage estimates are of course not as significant as before, and the F statistics are much lower, although higher than 10 (Table 11). In the second stage (Table 10), the number of observations drop from 314 to 59, and the intercept is out of the support. However, the sign of the coefficients is as expected, and an increase in the mean grades given by the teacher significantly decreases the proportion of bad evaluations.

To sum up, our results suggest a positive relationship between student evaluations of teacher's pedagogy and grades which are given by the teacher. This is in line with the literature, casting doubt on students' objectivity when they evaluate teaching quality. We cannot however distinguish the channels of such an effect. Do students reward (or respectively punish) teachers for lenient (or respectively severe) grading? Or do they attribute a good grade to good teaching? To try and learn

more, we propose to study the dynamics of evaluations over time.

Teacher's pedagogy	mean			%very good			%bad		
	OLS	FE	IV	OLS	FE	IV	OLS	FE	IV
Mean exam grade	0.077***	0.072*	0.086*	0.124	0.447	-0.576	-0.762**	-0.552	-0.645
	(0.022)	(0.037)	(0.047)	(0.236)	(0.326)	(0.463)	(0.314)	(0.424)	(0.628)
Mean cont. assess. without exam	0.059***	0.094**	0.061	0.093	0.293	-0.521	-0.767***	-0.951**	-0.657
	(0.018)	(0.036)	(0.039)	(0.204)	(0.339)	(0.384)	(0.286)	(0.438)	(0.531)
Mean cont. assess. with exam	-0.001	0.005	-0.002	0.219**	-0.075	0.147	-0.086	0.157	-0.022
	(0.008)	(0.012)	(0.010)	(0.094)	(0.161)	(0.106)	(0.085)	(0.115)	(0.114)
Intercept	1.803***	1.856***	1.815***	15.350***	15.749***	23.506***	23.670***	18.406***	21.577***
	(0.276)	(0.464)	(0.561)	(3.551)	(3.896)	(5.737)	(4.229)	(5.148)	(7.913)
Course characteristics	Yes	No	Yes	Yes	No	Yes	Yes	No	Yes
Teacher characteristics	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
After 2008 dummy	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R2	0.11	0.77	0.08	0.20	0.73	0.21	0.12	0.72	0.13
Nb obs	485	485	314	485	485	314	485	485	314
Nb clusters	97	97	96	97	97	96	97	97	96

Table 2 – Estimation of the effect of mean grade on teacher's pedagogy mean evaluation

Note: * ($p < 0.10$), ** ($p < 0.05$), *** ($p < 0.01$). The unit of observation is a triplet teacher-course-year. Observations are weighted by the number of students in the unit. Standard errors in brackets are clustered at the course level. Course characteristics = dummies for type of subject, dummies for year 1, 2 or 3 of the school program, dummy for direct admission catch-up course. Teacher characteristics = dummy for male, dummy for first year of teaching. The FE specification contains teacher-course fixed effects. In the IV specification, instrumental variables = lagged mean exam grade, lagged mean continuous assessment grade without exam, lagged mean continuous assessment grade with exam, and proportion of students retaking the exam.

Reading note: On average in the sample, 23.670% of students in a class rate the teacher's pedagogy as bad. A one point increase in mean exam grade is associated with a 0.762 decrease in this proportion. This coefficient is significant at the 5% significance level.

First stage	Mean exam	Mean cont. assess. without exam	Mean cont. assess. with exam
Lag mean exam	0.377***	0.115*	0.007
	(0.087)	(0.068)	(0.107)
Lag mean cont. assess. without exam	-0.479***	0.995***	0.037
	(0.060)	(0.036)	(0.097)
Lag mean cont. assess. with exam	-0.014	0.008	0.867***
	(0.016)	(0.010)	(0.058)
% resit exams	-0.049***	-0.007	0.060*
	(0.013)	(0.010)	(0.032)
Intercept	6.755***	-0.793	1.064
	(1.013)	(0.727)	(1.367)
Course characteristics	Yes	Yes	Yes
Teacher characteristics	Yes	Yes	Yes
After 2008 dummy	Yes	Yes	Yes
R2	0.86	0.92	0.92
Fstat 1st stage	281	507	242
Nb obs	314	314	314
Nb clusters	96	96	96

Table 3 – First stages

Note: * ($p < 0.10$), ** ($p < 0.05$), *** ($p < 0.01$). The unit of observation is a triplet teacher-course-year. Observations are weighted by the number of students in the unit. Standard errors in brackets are clustered at the course level. Course characteristics = dummies for type of subject, dummies for year 1, 2 or 3 of the school program, dummy for direct admission course. Teacher characteristics = dummy for male, dummy for first year of teaching.

Reading note: A one point increase in mean exam grade the preceding year is associated with a 0.377 increase in current mean exam grade. This coefficient is significant at the 1% significance level.

3.3 Evaluation dynamics over time

Students have a different set of information, both on the quality of the course and on their own achievement, depending on when they evaluate. Let us have a closer look at the dynamics of students evaluations with respect to the date when they take the written exam, and the date when they get their grades.

Figure 1 plots the mean evaluation of teachers' pedagogy over time. The graph first shows a slightly positive trend in student evaluations of teaching before the exam, followed by a sudden and significant decline at the exam date. Evaluations then linearly increase between the exam and the date when grades are released. After grades are released, the trend clearly reverses, although the variance of evaluations considerably rises.

In order to further understand these dynamics, Figure 2 separately presents the

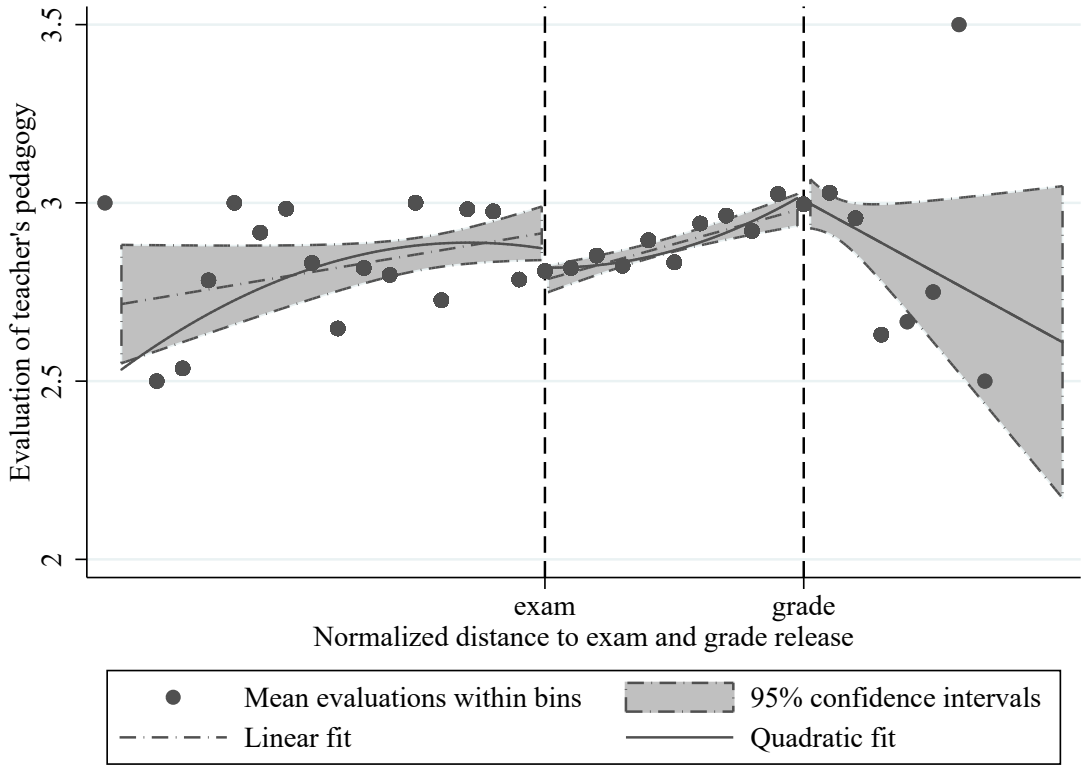


Figure 1 – Students evaluations of teaching over time

Note: The x-axis represents the distance to exam date and grade release date, centered in the exam date and divided by the difference between grade release date and exam date. Thus the unit of the x-axis has no interpretation.

proportion of bad, fair, good, and very good evaluations over time. It highlights that the negative jump in mean evaluation after the exam is due to a significantly higher proportion of fair evaluations, as well as a significantly but slightly lower proportion of very good evaluations.

To confirm these results, we propose to use a very flexible model of the relationship between students' evaluations of a teacher's pedagogy and the date of evaluation. More precisely, we use a generalized additive model, assuming that the expected value of the evaluation variable is an unknown function of the date of evaluation, in an additive relationship. The model is of the form:

$$g[\mathbb{E}\left(Eval_{icjt}|dist, u\right)] = \alpha + f\left(dist_{icjt}\right) + u_{ct} \tag{3}$$

As before, $Eval_{icjt}$ measures student i's evaluation of teacher j's pedagogy for course c in school year t.[7] Individual dates of evaluations are centered in the exam

[7]We used a dummy variable for very good (respectively bad) evaluations as an alternative

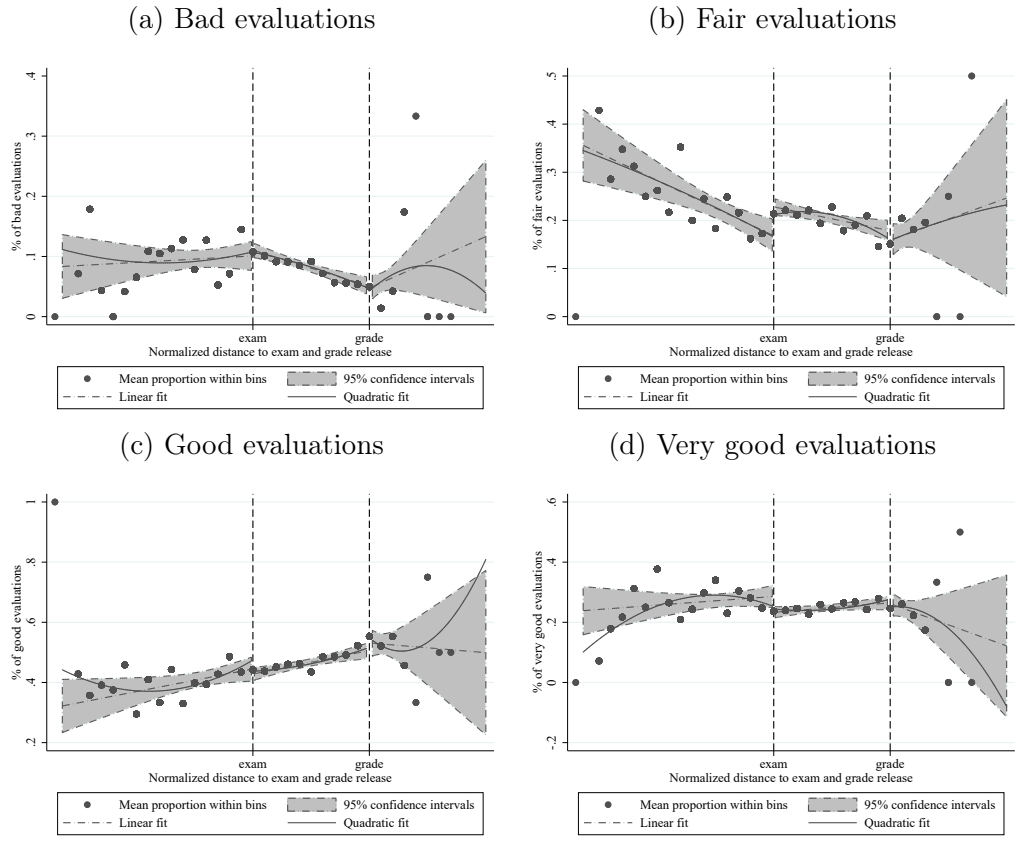


Figure 2 – Student evaluations of teaching over time (proportion of bad, fair, good, or very good)

Note: The x-axes represent the distance to exam date and grade release date, centered in the exam date and divided by the difference between grade release date and exam date. Thus the unit of the x-axis has no interpretation.

date, so that $dist_{icjt}$ is the individual distance to exam date, in days. In a second model, we also consider the distance to grade release.[8] The link function g is assumed to be a normal distribution and f is estimated using splines (of degree that is chosen using a generalised cross-validation method). Course-year fixed effects u_{ct} are added to control for differences in evaluations across courses and year.

Figure 6 in the appendix gives a graphical representation of the estimation of function f, separately for distance to exam, distance to grade release, for courses with and without exams. A significantly positive value of $\hat{f}$ means a significantly positive relationship between SET and distance to exam date (respectively grade release date). First, note that there is no significant relationship between evaluations and distance to grade release, for courses which are not graded through a written

outcome. The results are in line with this model's results.

[8] A model using both variables (distance to exam, and distance to grade release) together was also estimated, with similar results.

exam (Graph (c)). This result implies that students do not use (or no not react to) the information given by their grades to evaluate teacher's pedagogy, in the case when the course is graded through continuous assessment. Moreover, this suggests no selection effect with respect to the date of evaluation. In other words, students evaluating sooner do not rate teachers higher or lower. On the contrary, teachers giving a written exam are evaluated differently according to the date of evaluation (Graphs (a) and (b)). More precisely, Graph (a) shows a U-shape function, with a small but significant decrease in $\hat{f}$ after the exam, meaning that students rate teachers lower after taking the exam (for about a week). This could be due to an effect of disappointment, or a way of punishing teachers. Note, however, that the decrease begins before the exam, and could thus hide selection effects, i.e. satisfied-students may evaluate sooner. After grades are released (Graph (b)), $\hat{f}$ significantly increases, meaning that students rate a teacher's pedagogy higher on average after getting their grades. Again, this could suggest either a reward, or an attribution of good grade to good teaching.

To sum up, our results show that students give teachers lower evaluations after the exam but better evaluations after getting their grades. These results highlight that the way students evaluate teaching depends highly on the contemporaneous information they get. This suggests that they are not able to correctly evaluate teaching quality with the information they have from attending the course and that they use any additional information they get to derive a measure of teaching quality.

4 Conclusion

This paper evaluates students' objectivity in their evaluation of teaching. We use a unique data base in the context of a French higher education institution. First, we analyze the relationship between grades and evaluations. Using teacher-course fixed effects and instrumental variables, we confirm that students do take their grades into account when they evaluate the pedagogy of a teacher. The relationship is positive, suggesting either that students reward teachers for good grades, or that they attribute a good grade to a good teaching. The size of the effect is relatively small and driven by a decrease in the proportion of unhappy students; a one point increase in by-course mean exam grade corresponds to less than a 1 percentage point decrease in the proportion of students giving bad evaluations. Second, we study the timing of evaluations, using information on the exact date when students evaluate teaching. We find that students use contemporaneous information when they evaluate the teacher's pedagogy. They give lower evaluations after taking the

exam, and higher evaluations after getting their grades. When students are graded through continuous assessment, however, they do not seem to react to the release of their grades.

Our results are based on data from one particular French higher education institution. They may not be representative of every higher education institution, nor even of every French university. The institution studied here is a very selective school, and delivers a degree that is a positive signal on the French labor market. Once they have been selected to enter the school, students may not care what grades they get as much as other students. This may partly explain why we find relatively small effects. Although we cannot compare our results with other French institutions, we can speculate that student grades could bias their evaluation of teaching even more in other institutions. Moreover, teacher recruitment procedures in this school are different than in other institutions, so that the use of SET by the administration is presumably not the same. It may be that SET matters more for the organization of teaching in this particular school than in other French higher education institutions.

These results lead to two conclusions. First, they confirm that evaluations may be distorted by teachers trying to buy good evaluations, or by students trying to extrapolate the quality of the course through the exam or through the grades they obtain. Second, our results highlight that students have difficulties evaluating teaching and use available information to do it. A solution may be to make all students evaluate together at a single date in time. This may homogenize the information they get, but this would not prevent students from using the contemporaneous information they have at that date.

Our purpose is not to recommend against the use of SET. On the contrary, it has proven to be a relevant measure (at least of some aspects of) teaching quality. After all, who better than the persons attending the course to best evaluate the quality of teaching? However, institutions should be aware of potential distortions both teachers and students are likely to create. In order to obtain an objective measure of teaching quality, the administration may also want to rely on other ways to evaluate teaching.

References

Becker, W. E., W. Bosshardt, and M. Watts (2012). Revisiting how departments of economics evaluate teaching. Working paper presented at the annual meetings of the American Economic Association.

Becker, W. E. and M. Watts (1999). How departments of economics evaluate teaching. *The American Economic Review 89*(2), 344–349.

Beleche, T., D. Fairris, and M. Marks (2012). Do course evaluations truly reflect student learning? evidence from an objectively graded post-test. *Economics of Education Review 31*(5), 709–719.

Boring, A. (2017). Gender biases in student evaluations of teaching. *Journal of Public Economics 145*, 27 – 41.

Boring, A., K. Ottoboni, and P. Stark (2016). Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research*.

Boring, A. and A. Philippe (2017). Reducing Discrimination through Norms or Information: Evidence from a Field Experiment on Student Evaluations of Teaching. TSE Working Papers 17-865, Toulouse School of Economics (TSE).

Braga, M., M. Paccagnella, and M. Pellizzari (2011). Evaluating students' evaluations of professors. *IZA Discussion Paper No. 5620*.

Braga, M., M. Paccagnella, and M. Pellizzari (2014). Evaluating students' evaluations of professors. *Economics of Education Review 41*, 71 – 88.

Carrell, S. E. and J. E. West (2010). Does professor quality matter? evidence from random assignment of students to professors. *Journal of Political Economy 118*(3), 409–432.

d'Apollonia, S. and P. C. Abrami (1997). Navigating student ratings of instruction. *American psychologist 52*(11), 1198.

De Witte, K. and N. Rogge (2011). Accounting for exogenous influences in performance evaluations of teachers. *Economics of Education Review 30*(4), 641 – 653.

Dejean, J. (2002). L'évaluation de l'enseignement dans les universités françaises. Technical report, Haut Conseil de l'évaluation de l'école.

Ewing, A. M. (2012). Estimating the impact of relative expected grade on student evaluations of teachers. *Economics of Education Review 31*(1), 141–154.

Franz, W.-J. I. (2010). Grade inflation under the threat of students' nuisance: Theory and evidence. *Economics of Education Review 29*(3), 411–422.

Greenwald, A. G. (1997). Validity concerns and usefulness of student ratings of instruction. *American Psychologist 52*(11), 1182.

Greenwald, A. G. and G. M. Gillmore (1997). Grading leniency is a removable contaminant of student ratings. *American psychologist 52*(11), 1209.

Hoffmann, F. and P. Oreopoulos (2009). Professor qualities and student achievement. *The Review of Economics and Statistics 91*(1), 83–92.

Isely, P. and H. Singh (2005). Do higher grades lead to favorable student evaluations? *The Journal of Economic Education 36*(1), 29–42.

Klitgaard, R. (1990). *Controlling corruption*. University of California Press.

Krautmann, A. C. and W. Sander (1999). Grades and student evaluations of teachers. *Economics of Education Review 18*(1), 59–63.

Marsh, H. W. and L. A. Roche (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist 52*(11), 1187.

Uttl, B., C. A. White, and D. W. Gonzalez (2017). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation 54*, 22 – 42. Evaluation of teaching: Challenges and promises.

Weinberg, B. A., M. Hashimoto, and B. M. Fleisher (2009). Evaluating teaching in higher education. *The Journal of Economic Education 40*(3), 227–261.

Appendix

	Answers			
Questions	1	2	3	4
Interest for the subject	☐ Not interesting	☐ Moderately interesting	☐ Interesting	☐ Very interesting
Difficulty of the course	☐ Easy	☐ Moderate	☐ Difficult	☐ Very difficult
Course attendance	☐ Less than half	☐ About half	☐ All courses or almost	
Teacher's pedagogy	☐ Bad	☐ Fair	☐ Good	☐ Very good
Level of formalization	☐ Too low	☐ Adequate	☐ Too high	
Speed of progression	☐ Too slow	☐ Adequate	☐ Too fast	
Difficulty of the tutorials	☐ Easy	☐ Moderate	☐ Difficult	☐ Very difficult
Attendance in tutorials	☐ Less than half	☐ About half	☐ All tutorials or almost	
Teaching assistant's pedagogy	☐ Bad	☐ Fair	☐ Good	☐ Very good

Table 4 – Description of the evaluation form

Reading note: To the question concerning their interest for the subject, students could answer that they find it not interesting, interesting, moderately interesting, or very interesting. The variable is coded 1, 2, 3 or 4 respectively.

	Interest	Difficulty	Attendance	Pedagogy	Formalization	Progression
	%	%	%	%	%	%
1	5	3	16	8	7	4
2	13	37	27	19	78	76
3	49	45	57	49	15	20
4	33	15		24		
Total	100	100	100	100	100	100
Observations	17337	17260	17295	17254	17082	17016

Table 5 – Distribution of evaluations in the sample

Reading note: On average, 5% of students find that the subject of the course is not interesting and 33% find it very interesting. See table 4 for the meaning of each category.

	Before 2008		After 2008		Test
	N	(mean/sd)	N	(mean/sd)	(pvalue)
Grades					
Economics	1,966	12.12	2,760	11.47	0.106
		0.31		0.47	
Social Sciences	817	13.48	1,343	13.10	0.293
		0.47		0.45	
Finance and Actuarial Sciences	3,387	12.08	2,441	12.58	0.155
		0.31		0.34	
Mathematics and Computer Sciences	1,948	11.81	2,473	11.84	0.974
		0.87		0.38	
Statistics and Econometrics	3,479	11.48	3,584	12.02	0.060
		0.42		0.38	

Table 6 – Main characteristics of exam grades in the sample

Note: Standard errors are clustered at the course level.

Reading note: On average, students get 12.12 out of 20 on the exam before 2008 and 11.47 after 2008, for courses of the Economics major. The difference is not significantly different from zero.

	Before 2008		After 2008		Test
	N	(mean/sd)	N	(mean/sd)	(pvalue)
Evaluations					
Interest for the subject	4,735	3.07	12,602	3.09	0.699
		0.03		0.03	
Difficulty	4,682	2.71	12,578	2.73	0.698
		0.05		0.04	
Attendance	4,722	2.49	12,573	2.39	0.001
		0.03		0.02	
Pedagogy	4,627	2.82	12,627	2.93	0.084
		0.07		0.04	
Formalization	4,553	2.10	12,529	2.07	0.174
		0.02		0.02	
Speed of progression	4,531	2.19	12,485	2.13	0.027
		0.03		0.02	
Comments=1	4,923	0.46	12,637	0.23	0.000
		0.01		0.01	
School year					
1st=1	4,027	0.14	12,626	0.27	0.002
		0.04		0.06	
2nd Eco=1	4,027	0.19	12,626	0.16	0.256
		0.04		0.03	
2nd Fin=1	4,027	0.13	12,626	0.20	0.000
		0.02		0.03	
2nd Stat=1	4,027	0.14	12,626	0.08	0.003
		0.02		0.01	
3rd=1	4,027	0.41	12,626	0.29	0.017
		0.07		0.05	
Admission					
Direct=1	3,900	0.38	12,617	0.32	0.003
		0.02		0.03	
Eco=1	3,900	0.25	12,617	0.27	0.150
		0.02		0.02	
Math=1	3,900	0.37	12,617	0.41	0.015
		0.02		0.02	

Table 7 – Main characteristics of evaluations in the sample

Note: Standard errors are clustered at the course level.

Reading note: On average, students rate 3.07 out of 4 the interest for the subject before 2008 and 3.09 after. The difference is not significantly different from zero.

	Interest	Difficulty	Attendance	Pedagogy	Formalization	Progression	Grade
Interest	1.00						
Difficulty	-0.01	1.00					
Attendance	0.44***	-0.05	1.00				
Pedagogy	0.54***	-0.27***	0.41***	1.00			
Formalization	-0.14*	0.70***	-0.16**	-0.23***	1.00		
Progression	0.02	0.73***	0.04	-0.31***	0.55***	1.00	
Grade	0.18**	-0.45***	0.18**	0.30***	-0.40***	-0.30***	1.00

Table 8 – Correlations between mean evaluations and mean grades

Note: * ($p < 0.10$), ** ($p < 0.05$), *** ($p < 0.01$). Correlations are calculated on teacher-subject-year means.

Reading note: The correlation between mean evaluation of a teacher's pedagogy and mean evaluation of interest for the subject is 54 %. The correlation between mean evaluation of a teacher's pedagogy and mean exam grade is 30 %.

	Inter-student variance	Inter-course variance
Pedagogy	86 %	14 %
Interest	92 %	8 %
Difficulty	74 %	26 %
Attendance	91 %	9 %
Formalization	91 %	9 %
Progression	89 %	11 %
Grades	84 %	16 %

Table 9 – Variance decomposition of evaluations and grades

Reading note: 86% of the total variance of evaluations of teacher's pedagogy is due to inter-student variance (intra-course variance). 14% is due to inter-course variance.

Teacher's pedagogy	mean			%very good			%bad		
	OLS	FE	IV	OLS	FE	IV	OLS	FE	IV
Mean exam grade	0.077***	0.072*	0.390	0.124	0.447	3.848	-0.762**	-0.552	-8.038*
	(0.022)	(0.037)	(0.315)	(0.236)	(0.326)	(2.822)	(0.314)	(0.424)	(4.642)
Mean cont. assess. without exam	0.059***	0.094**	0.323	0.093	0.293	2.806	-0.767***	-0.951**	-6.571*
	(0.018)	(0.036)	(0.241)	(0.204)	(0.339)	(2.129)	(0.286)	(0.438)	(3.488)
Mean cont. assess. with exam	-0.001	0.005	0.056	0.219**	-0.075	0.796	-0.086	0.157	-1.509
	(0.008)	(0.012)	(0.073)	(0.094)	(0.161)	(0.670)	(0.085)	(0.115)	(1.165)
Intercept	1.803***	1.856***	-3.059	15.350***	15.749***	-32.441	23.670***	18.406***	128.997*
	(0.276)	(0.464)	(4.496)	(3.551)	(3.896)	(40.373)	(4.229)	(5.148)	(67.035)
Course characteristics	Yes	No	Yes	Yes	No	Yes	Yes	No	Yes
Teacher characteristics	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
After 2008 dummy	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R2	0.11	0.77	.	0.20	0.73	.	0.12	0.72	.
Nb obs	485	485	59	485	485	59	485	485	59
Nb clusters	97	97	42	97	97	42	97	97	42

Table 10 – Estimation of the effect of mean grade on teacher's pedagogy mean evaluation - IV with a change of teacher

Note: * ($p < 0.10$), ** ($p < 0.05$), *** ($p < 0.01$). The unit of observation is a triplet teacher-course-year. Observations are weighted by the number of students in the unit. Standard errors in brackets are clustered at the course level. Course characteristics = dummies for type of subject, dummies for year 1, 2 or 3 of the school program, dummy for direct admission catch-up course. Teacher characteristics = dummy for male, dummy for first year of teaching. The FE specification contains teacher-course fixed effects. In the IV specification, instrumental variables = lagged mean exam grade, lagged mean continuous assessment grade without exam, lagged mean continuous assessment grade with exam, and proportion of students retaking the exam, in the case when the teachers changes betwwen $t-1$ and t.

Reading note: On average in the sample, 23.670% of students in a class rate the teacher's pedagogy as bad. A one point increase in mean exam grade is associated with a 0.762 decrease in this proportion. This coefficient is significant at the 5% significance level.

First stage	Mean exam	Mean cont. assess. without exam	Mean cont. assess. with exam
Lag mean exam	0.258	-0.022	-0.447
	(0.298)	(0.293)	(0.466)
Lag mean cont. assess. without exam	-0.214	0.495	-0.333
	(0.284)	(0.321)	(0.329)
Lag mean cont. assess. with exam	-0.073	-0.022	0.574***
	(0.119)	(0.125)	(0.175)
% retake	0.145*	-0.249**	0.135
	(0.085)	(0.100)	(0.080)
Intercept	6.214	5.778	8.197
	(4.868)	(5.466)	(6.777)
Course characteristics	Yes	Yes	Yes
Teacher characteristics	Yes	Yes	Yes
After 2008 dummy	Yes	Yes	Yes
R2	0.65	0.72	0.74
Fstat 1st stage	10	15	13
Nb obs	59	59	59
Nb clusters	42	42	42

Table 11 – First stages with a change of teacher

Note: * ($p < 0.10$), ** ($p < 0.05$), *** ($p < 0.01$). The unit of observation is a triplet teacher-course-year. Observations are weighted by the number of students in the unit. Standard errors in brackets are clustered at the course level. Course characteristics = dummies for type of subject, dummies for year 1, 2 or 3 of the school program, dummy for direct admission course. Teacher characteristics = dummy for male, dummy for first year of teaching.

Reading note: A one point increase in mean exam grade the preceding year is associated with a 0.258 increase in current mean exam grade. This coefficient is not significantly different from zero.

(a) in 2007-2008 (b) in 2009-2010

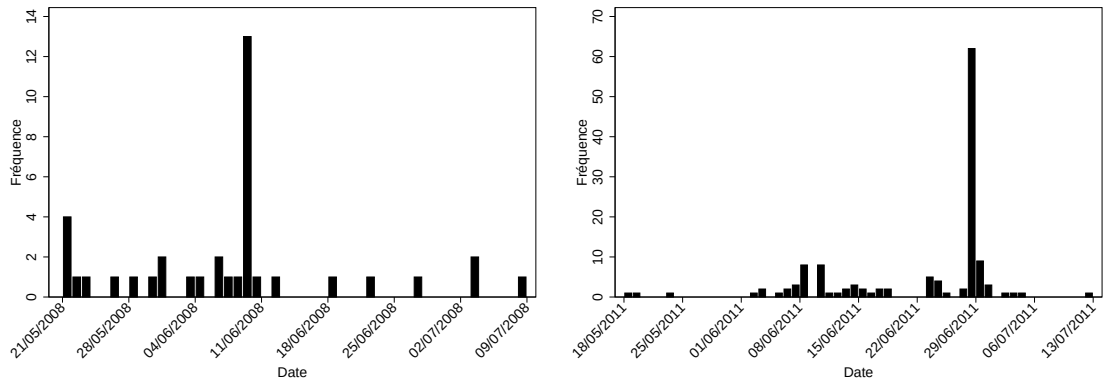


Figure 3 – Dates of evaluations

Reading note: These graphs present the histograms of the number of evaluations per day for one course in 2007-2008 (when evaluation was optional) and in 2009-2010 (when evaluation was compulsory).

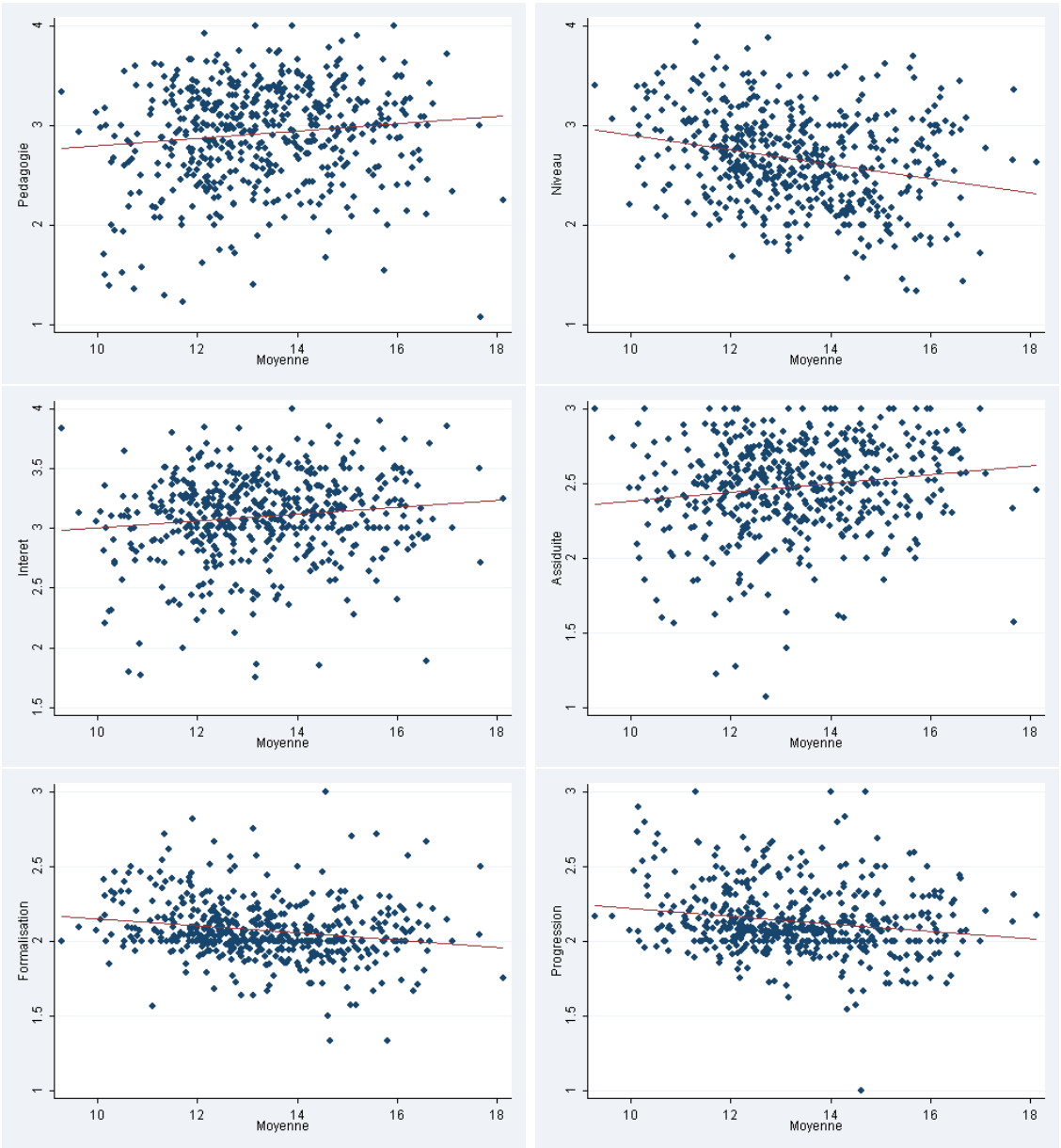
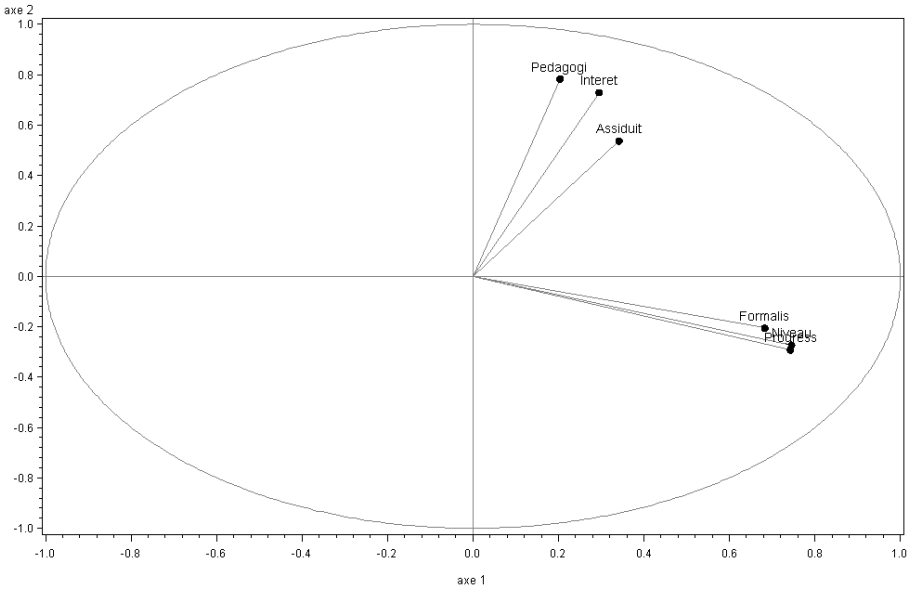


Figure 4 – Linear relationship between evaluations and exam grade

Note: These graphs plot by teacher-subject-year mean evaluation over mean exam grade. The lines represent the linear regression fit of mean evaluations over mean grade.

(a) At the individual level



(b) At the teacher-subject-year aggregate level

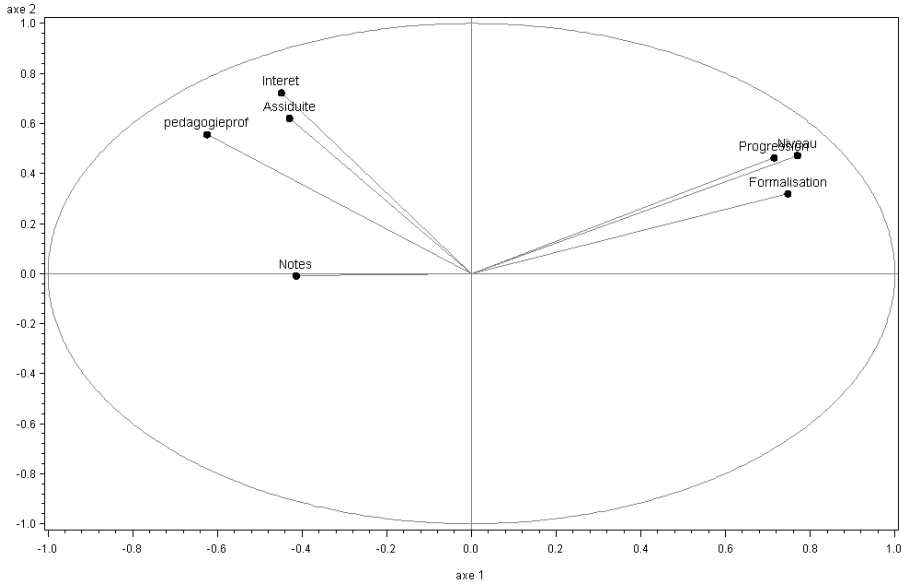
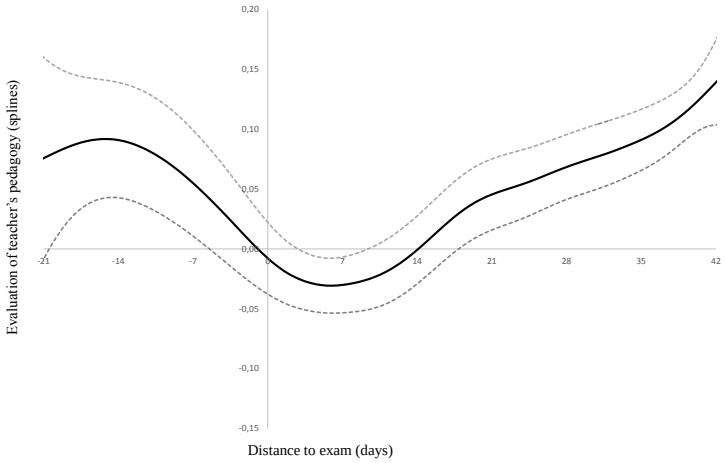


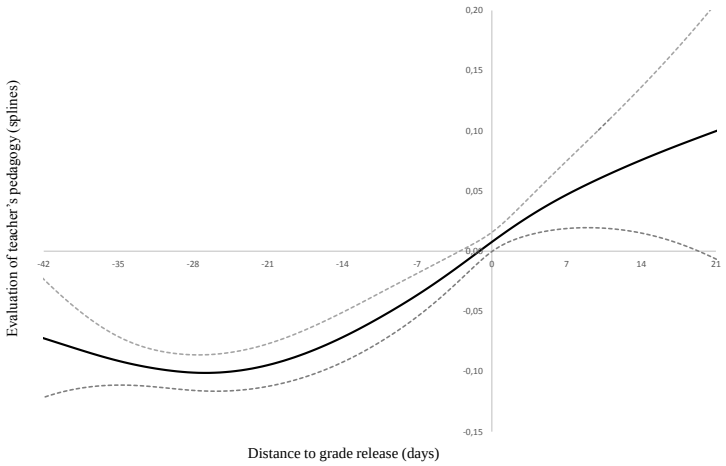
Figure 5 – Multivariate relationship between evaluations and exam grade

Note: These graphs represent the scatter plot of a principal component analysis, at the individual level (top), and at the teacher-subject-year aggregate level (bottom).

(a) Distance to exam (courses with exam)



(b) Distance to grade release (courses with exam)



(c) Distance to grade release (courses without exam)

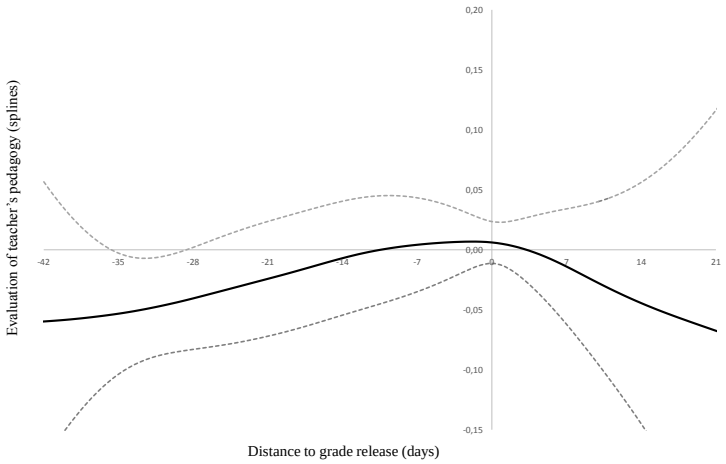


Figure 6 – Estimation of the relationship between evaluations and distance to exam (respectively grade release) date

Note: The x-axes represent the distance to exam date (respectively grade release date), centered in the exam date (respectively grade release date), in days. The y-axes represent the estimation of function f using a generalized additive model, with course-year fixed effects. Function f is estimated using splines of degree 6 for distance to exam, and 3 for distance to grade release. Dashed lines give 95% confidence intervals. A positive value of $\hat{f}$ means a positive relationship between SET and distance to exam date (respectively grade release date).