Grade repeaters: Pygmalion in the French educational system?

Applied Econometrics Project

Violetta van Veen*, Mahdi El Amin*, Catherine Berleur*†

April 2022

Abstract: In the context of a disadvantaged French district, we measure teachers' bias against class repeaters using differences in scores between school exams graded by teachers and anonymous exams graded blindly by external examiners. Relying on the "as good as random" assignment of students to teachers with different levels of stereotypes and bias in secondary school in France, we identify such bias using a differences-in-differences method. Our results suggest that there is a bias in the grading of French tests against students who have repeated a year, in particular when they come from disadvantaged backgrounds. There is also a significant negative effect of teachers' bias towards repeaters on their progress relative to non-repeaters.

1 Introduction

Repeating a year/grade retention means that a pupil has to to repeat a whole school year, by repeating the same lessons, when his or her level is deemed insufficient to move up to the next grade. France remains one of the countries with the highest rate of grade retention among students despite a steady decline over the past 30 years and an almost unanimous finding of its ineffectiveness by researchers (Marcoux and Crahay 2008; Seibel 1984). In 2012, 28% of the French 15-year-old students repeated a year during their schooling years. This counter-performance places France in 5^{th} place among the countries with the highest number of repeaters (OECD, PISA, 2012). France is an exception, since one third of OECD countries have a repetition rate at the same age of less than 5%. This tendency to repeat years is a particularity of French-speaking countries and education systems, with Luxembourg and Belgium (Walloon) also among the five countries with the highest percentage of repeaters. Some countries, such as Iceland and Norway, which are often seen as models in terms of education, have even banned repetition. The issue of repetition is essential in the social sciences, both from a public economics point of view in terms of funds spent for keeping students in school one year more and from the point of view of the econometrician who studies its effectiveness (Gary-Bobo and Robin 2012). The French peculiarity then raises questions, as a paradox easily emerges: why does the French education system continue to make pupils repeat a year, while knowing thanks to researchers that this is inefficient? From this paradox, the question of teacher bias towards pupils naturally emerges, since it is they who mainly decide on repetition. The dataset at our disposal allows an assessment of teacher bias through a natural experiment. Given the repetition rates, it is expected that teachers will have a positive bias towards repeaters. We will see later that the opposite happens, which reinforces the paradox of the issue of repetition in France and inspired us the title "Grade repeaters: Pygmalion in the French educational system?" in reference to *Pygmalion in the classroom* from Rosenthal and Jacobson 1968.

The difficulties of empirical evaluation are mainly linked to the endogenous nature of repetition decisions. Repetition seems to have harmful consequences on pupils' trajectory (Alexander, Entwisle, and Kabbani 2001). Although repeaters make slight progress in the year following their repetition, they make less progress than pupils who have been promoted (Seibel 1984), making this strategy counterproductive. Indeed, the beneficial effect of repetition disappears as soon as new skills or subjects have to be learned: there is therefore no overall positive cognitive effect (Goos 2013). However, some public authorities, teachers and even parents continue to have a positive image of repetition (Crahay 2007). Some econometric results lead to a more nuanced assessment: repetition woulds have certain favourable effects in the short term, but would be to some extent harmful in the long term. Dong 2010 found positive effects of repetition in the short term - in the first two years after the decision - but negative effects in the longer term. Among the negative consequences in the long-term, repetition has an impact on student self-confidence and, most importantly, on the increased risk of dropping out of school (Burkam et al. 2007; Reynolds, Magnuson, and Ou 2010) The issue of teacher bias thus naturally emerges from this phenomenon, which Dong 2010 explains primarily in terms of the greater support generally offered to repeaters than in terms of the repetition itself.

With respect to the literature on repetition, we place ourselves in a new perspective of studying the grade retention phenomenon: we extend the methods used by Lavy 2008 and Terrier 2020 to study teachers' biases

*Paris School of Economics

†The authors would like to thank Camille Terrier for the helpful comments.

on pupils according to their gender to the case of pupils in a repetition situation. Implementing a difference-in-difference strategy with fixed effects allows us to identify the mean difference in score gaps between repeaters and non-repeaters thanks to blind and non-blind scores. We also compute a difference-in-difference-in-difference estimator to add an additional comparison group and estimate treatment effects.

We use a data set of blind and non-blind test scores collected by Avvisati et al. 2014. The potential grade retention bias is therefore the difference between repeaters and non-repeaters gaps between the blind and non-blind test scores. Studying the biases of teachers assessing repeaters sheds light on the effects of repetition, adding to the literature on natural experiments about the effects of the beliefs and expectations of teachers on students (Rosenthal and Jacobson 1968). Our identification strategy is based on previous work on experiments comparing blind and non-blind examinations (Blank 1991; Goldin and Rouse 2000) and then applied to educational sciences (Lavy 2008, Breda and Ly 2015, Breda and Hillion 2016). We use the blind scores as counterfactuals to identify teachers' bias against repeaters. We implement a differences-in-differences in a fixed effect framework following the works of Lavy and Sand 2018, Terrier 2020 and Breda and Ly 2015. Moreover, we follow Terrier 2020 in developing a framework to assess the impact of teachers' bias towards repeaters on their relative progress.

Our results suggest that there is a bias in the grading of French tests against students who have repeated a year. This goes against the positive image that most teachers have of repeating a year, and may therefore seem counter-intuitive (Goos 2013). When using a DDD estimation method, we find that this bias is evident only when repetition sums up with other disadvantages, such as unemployed parents or low socioeconomic status. Regarding the effect of bias on students' progress, we find a large and significant effect of bias on the relative progress of repeaters for both Maths and French. Spillovers effects between teacher biases within the same class are negligible.

Our data include not only students who repeated the year under study, but any year prior to the experiment. This could be a limitation of our results, as the effect of repeating a class is largely heterogeneous depending on the period in which it occurs (Cooley Fruedwirth, Navarro, and Takahashi 2011). However, repeating a grade in high school can happen at the request of the of the family, e.g. as part of an orientation strategy. In elementary school, on the other hand, repetition is most often a decision of the teacher. Therefore, in our sample we only have students who repeated a year because the teachers in elementary school deemed it necessary.

Section 2 describes the data and the context, section 3 sets out the model, section 4 gives the main results, section 5 checks for robustness, section 6 concludes.

2 Data and context

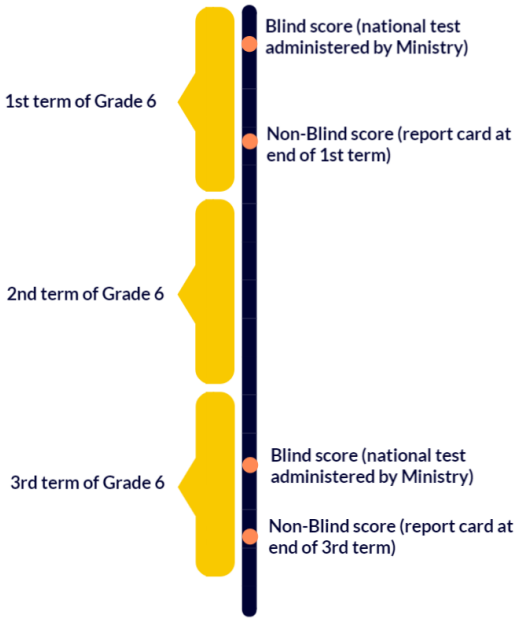
We use the dataset for the 2008–2009 school year collected by Avvisati et al. 2014, which covers 35 middle schools, 191 classes and 4490 students in the French district of Créteil.

At the French elementary school level, repeating a year is is proposed by the teachers' council, after the opinion of the National Education Inspector, and must be the subject of a dialogue phase conducted with the student's legal representatives. It must be be accompanied by a support system. French regulations on repeating a year are not very restrictive and not very precise: the decision to repeat a year was left to the discretion of the teaching staff before 2014. Parents can appeal against schools' decisions, therefore the weight of parents is still central compared to other European regulations.

We use externally marked tests administered by the French Education Ministry as blind scores and the average grades given by the teacher on end-of-term report cards as non-blind scores. People who grade the Ministry tests, in fact, have no access to names or any personal characteristics of the student. Teachers, instead, have an everyday contact with the student for the whole year, so their grades may be biased by their prior ideas on the student. For each student, we have blind and non-blind scores collected during the first and last term of grade 6. We consider separately French and maths scores.

Our sample is collected in the first year of middle school (grade 6). In the French system, children who did not repeat a year enter sixth grade at the age of 11. There is no streaming by ability across schools and pupils stay in the same class with the same teacher (one for French, one for maths) throughout the school year.

Figure 1: Timeline of data collection.



2.1 Definition of the variables

2.1.1 Repetition

Most of the grades retention happen in primary school, i.e. before the collection of our data. The only data available is the date of birth of pupils: in particular, if they are born before 1/1/1997, we code them as repeaters. Thus, not only we are not able to know which school year was repeated, but we cannot know if for example they repeated a grade or began school one year later for any reason. However, in France it is not customary to make migrant children who came to France repeat a class in order to learn French: until 2012 (so in all years relevant for our sample, collected in 2008), middle schools had to organize one class called CLA ("Classe d'accueil"). Children who do not speak French are enrolled in an ordinary class, which generally corresponds to their age, and do some courses (for example, Physical Education) with their ordinary class. Thus, even if they are more likely to (Lorcerie 1998; Payet, Zanten, et al. 1996), migrant children do not by default repeat a year. If by default all migrant children repeated a year, we would not be able to disentangle the bias against repeaters from the bias against migrants.

2.1.2 Punishment dummy

We have administrative data on whether the pupil received a disciplinary sanction or was temporarily excluded in each of the three terms of grade 6. We computed a punishment dummy equal to 1 if the pupil has been punished at least once in one of the 3 terms. In this way, we want to study heterogeneous effects for repeaters who also have disciplinary problems.

One might wonder if there is some degree of simultaneity between the bias of the teacher and the teacher's decision of punishing the child. However, both types of punishments are not given by a single teacher, but by the class head teacher, after agreeing with all the teachers of the class council. Thus, punishment is not determined solely by the teacher handing out grades.

2.1.3 Teachers' questionnaires

At the end of grade 6, teachers were asked to fill in a questionnaire on each student with 5 questions. Each question had a yes or no answer (so 1 or 0). The questions were as follows:

- Is he/she pleasant in class?
- Did he/she work diligently?
- Did he/she progress over the year?
- Was dialogue with the child's parents satisfactory?
- Did his/her parents provide him/her with support with school work?

We used a sum of the results to measure a teacher's total questionnaire score for each student. The scores range from 0 to 5 and can only take the form of a digit. In other words, the lowest possible value this score can have is 0, all answers are no, and the highest possible value is 5, all answers are yes. We use this value as an indicator of the teacher's perception of the student, i.e. the higher the score is, the more the teacher thinks of the student as a "good" or "hard-working" student.

2.1.4 Blind tests

The anonymous scores we have are taken from the "Evaluations nationales en classe de sixième" (National Evaluations in the sixth grade). These are conducted at the national level by the DEPP ("Direction de l'évaluation, de la prospective et de la performance"), a statistical service of the Ministry of Education in France. The tools for this evaluation were designed with expert groups composed of educational advisors, master trainers, school teachers and certified teachers, set up by the DEPP in collaboration with the Inspectorate General. Each student is assessed in French and mathematics.

2.2 Descriptive statistics

28% of students in our dataset have repeated a class in the past. This is determined by comparing the repeater's age to the age they're supposed to be at their grade level. The average number of repeaters per class is 6.48, only in one of the 191 classes there are no repeaters. Results are robust to the exclusion of this class and of the one with 24 repeaters. Table 1 provide summary statistics divided for repeaters and not. There is a similar shift in the average blind and non-blind scores between repeaters and not repeaters (See figure 4 in the Appendix). As pointed out by Davaillon and Nauze-Fichet 2004, repeaters tend to come from poorer backgrounds. We have data on one parent, the self-declared "responsible legal 1": repeaters have significantly less employed parents, more mono-parental families and more need-based scholarships. Based on social categories from INSEE[1], repeaters have a significantly lower percentage of parents working as managers and executives, the two professions with higher salaries. The percentage of girls and boys among repeaters is not completely balanced, with 41% of girls and 59% of boys. This difference is milder than the one found at the national level by CNESCO 2014: a boy has a 48% higher relative probability of having repeated a year than a girl. Coherently with CNESCO 2014, however, more repeaters than non repeaters have a monoparental family in our sample, and at a national level living in a single parent family increases the probability of repeating a grade by 37%. Despite lower grades by teachers and a higher percentage of disciplinary sanctions, repeaters tend significantly less to repeat grade 6 too.

2.3 Identification strategy

In terms of identification, it is relevant that parents are not free to choose the public school that their children will attend (otherwise than through residential location, or the choice to attend a private school). Similarly, teachers cannot choose how many repeaters they have in class. This is relevant for the identification of the effect of being assigned a teacher who is 1 SD more biased against repeaters on repeaters' average progress relative, after controlling for the initial achievement gap.

Our diff-in-diff approach requires that:

1. The difference between blind and non-blind test is not correlated with other factors that have an impact on the scores
2. Repeating the year is not systematically affected by other variables that are in the error term

Concerning the first Zero Conditional Mean assumption, a threat to our identification is whether blind and non-blind[2] tests are actually comparable. Do they measure the same skills? The Ministry test for French has items evaluating the comprehension of literary texts, documents and oral information. The one for Mathematics focuses on geometry, arithmetics and the ability to use quantities and units of measurement. Standardized tests at the beginning and at the end of grade 6 are very similar and equal for all schools in our dataset. On the one hand, teachers may evaluate their students using oral presentations, homeworks and essay writing, which involve different competences (i.e. public speaking). On the other hand, both tests in class and the national evaluation rely more on written questions than on multiple choice[3]. Moreover, almost half French and math teachers report that they design their evaluations based on the national evaluation (Braxmeyer, Guillaume, and Levy 2004). This is particularly noteworthy as if repeaters or non-repeaters were more endowed in an ability measured more by the class exam, the double difference would not capture only teachers' bias. It would also potentially capture repeaters' or non-repeaters' particular skills measured by one of the scores (including homework, for instance).

Secondly, are blind and non-blind tests administered in the same way? On the one hand, both are administered by the teacher of the subject and, in 2008, were taken on paper (the national evaluation is now computer-based).

1. Farmers, Self-Employed, Managers, Executives, Salaried Employees, Manual Workers, Retirees, Other Inactive, Unknown according to the INSEE
2. Documentation and examples of the externally graded tests are available for Math and French.
3. In externally graded tests only 18% of items in French and 5% in Maths ask to choose an answer from a list.

Table 1: Descriptive statistics for repeaters and non-repeaters.

	Not-late (1)		Late (2)		Difference (3) = (2)-(1)	
	Mean	Std. Dev.	Mean	Std. Dev.	Diff. in Means	Std. Error
Students' characteristics						
Girls (%)	0.51	0.50	0.41	0.49	-0.11	0.02
First child	0.54	0.50	0.54	0.50	-0.01	0.02
Academic results						
Non-Blind French (1st Term)	12.59	3.28	9.86	3.08	-2.73	0.11
Non-Blind French (3rd Term)	11.82	3.50	8.94	3.39	-2.89	0.12
Non-Blind Math (1st Term)	13.32	3.69	10.33	3.69	-2.99	0.13
Non-Blind Math (3rd Term)	11.96	4.08	8.86	3.86	-3.10	0.14
Blind French (1st Term)	0.23	0.95	-0.60	0.87	-0.83	0.03
Blind French (3rd Term)	0.17	0.97	-0.56	0.87	-0.73	0.04
Blind Math (1st Term)	0.22	0.96	-0.61	0.87	-0.83	0.03
Blind Math (3rd Term)	0.18	0.97	-0.61	0.86	-0.79	0.04
N	3314		1283			
Behaviour (3rd Term)						
Disciplinary warning (%)	0.08	0.27	0.15	0.36	0.07	0.01
Grade 6 retention (%)	0.03	0.17	0.01	0.10	-0.02	0.00
Honours ("Mention")	0.42	0.49	0.22	0.41	-0.21	0.02
Half-day absences in 3rd term	2.81	5.39	7.08	11.51	4.27	0.40
N	3005		1167			
Socio-economic characteristics						
At least one parent employed	0.89	0.32	0.74	0.44	-0.14	0.01
High SES (%)	0.22	0.42	0.09	0.29	-0.13	0.01
2 parents in the household	0.76	0.43	0.64	0.48	-0.12	0.02
Need-based scholarship	0.29	0.46	0.41	0.49	0.12	0.02
N	3314		1283			
Teacher's questionnaire (%)						
Behaviour in class	0.64	0.48	0.46	0.50	-0.18	0.02
Diligence in class work	0.65	0.48	0.32	0.47	-0.33	0.02
Progress over the year	0.69	0.46	0.35	0.48	-0.34	0.02
Dialogue with parents	0.85	0.36	0.65	0.48	-0.19	0.02
Parents' support to homework	0.28	0.45	0.08	0.27	-0.20	0.01
N	2516		808			

Blind scores are standardized to a normal $N(0,1)$. High SES (Socio-Economic Status) for the parents professions takes value 1 if the parents belong to the French administrative category "manager" or "executive". Teacher questionnaire were administered to the reference teacher at the end of grade 6. Honours obtained indicates a judgement on the general attitude. In this table, there is the percentage of students who obtained either the lower level (encouragements), the second level (compliments) *or* the highest honours (félicitations).

In both tests students are in their usual classroom. On the other hand, they may create different incentives. The results of standardized tests are not sent to parents. Conversely, teachers' end-of-term report cards are an important part of the reward/sanction system in schools.

In terms of the Zero Conditional Mean assumption for the repetition dummy, it is likely that there are factors that have an impact on both the scores and the fact of having repeated a year in the past. It's important to note that the decision of retaking the year and the grades in our sample are not simultaneous (as students have repeated a year in the past). Apart from the direct effect of retaking a year on scores, as we have seen in the descriptive statistics and from the literature (CNESCO 2014), the probability of repeating a year is higher for kids who were born later in the year. The birth date may have an effect on the scores too, but this effect tends to fade as the child gets older. In our sample, children are 11-12 years old and are in their 6th year of school. For example, data from PISA 2018 in France show mixed results, with children born later in the year outperforming those born before in some subjects (Givord 2020). Thus, we argue that even if the month of birth may have an impact on the probability of having retaken a year, its impact on the scores is not clear enough to say it is biasing the coefficient of the impact of repetition on grades.

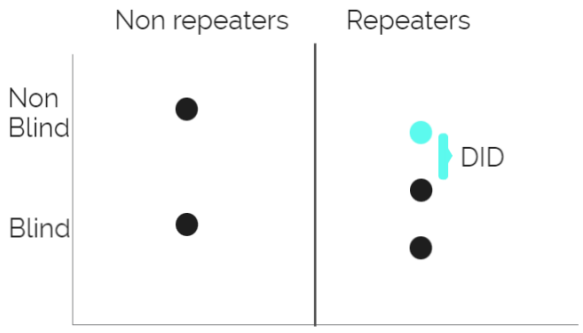
Children from monoparental families tend to repeat more often (CNESCO 2014), as we can also see from the descriptive statistics in our sample. However, we are able to control for this confounding factor, taking the monoparental family dummy out of the error term (see Table 10). Similarly, repetition is higher among low-educated and unemployed parents: we run a DDD regression adding heterogeneity for unemployed and employed parents (CNESCO 2014). Even if we are not able to control directly for the parents' education, we can proxy it using their social category.

Some characteristics had an important effect on the probability of retaking the year in the past, but seem no longer significant in the considered period (2008-2009): in particular, sex of the child, family characteristics and the language spoken at home (CNESCO 2014).

3 Method

3.1 Difference in difference

Figure 2: Difference in difference framework and parallel trend assumption.



We follow Lavy 2008, Lavy and Sand 2018 and Terrier 2020 in using a difference-in-difference estimation strategy. Our base model is:

$$S_{itj} = \beta_0 + \beta_1 NB_{itj} + \beta_2 R_{itj} + \beta_3 (NB_{itj} \times R_{itj}) + \psi_j + u_{itj} \quad (1)$$

where j is the class, i is the student and t is the period (at the beginning or at the end of grade 6). S_{itj} is the blind or non-blind grade a pupil receives. NB_{itj} is a dummy equal to 1 when the score is non-blind. Its coefficient β_1 identifies the "grade inflation" for non-repeating students in the teacher grading. R_{itj} is a dummy equal to 1 when the student has repeated a year in the past. Its coefficient β_2 identifies the repeaters achievement gap. ψ_j are class fixed effects. u_{itj} is the error term, which for now we assume to be independent. The zero conditional mean assumption is required for the unbiasedness of the diff-in-diff estimator: we require that both the dummy repeater (R) and the dummy non-blind (NB) are not systematically related to other factors that affect the scores S_{itj} and are hidden in u_{itj} (Wooldridge 2010). This assumption is discussed in the identification section 2.3.

We run separate regressions for math and French.

Our parameter of interest is β_3, which identifies the mean difference in score gaps. A negative β_3 indicates that repeaters, conditional on the blind scores, receive lower grades from their teachers than non-repeaters.

The assumption that both tests measure the same abilities corresponds to the common trend identification hypothesis. Without teachers' bias, the difference between the non-blind score and the blind score should be the same for repeaters and non-repeaters.

We include class fixed effects in order for our coefficients not to be biased by all elements which affects grades in a specific class (e.g. teachers' tightness with grades or peer effects). Errors are clustered at school level, since we expect unobserved components for children within a school to be correlated. The experimental design by Avvisati et al. 2014, in fact, involves sampling a selection of middle schools in the district, and in our case data was collected for all 6th grade classes in the sampled schools.

3.2 DDD

We compute a difference-in-difference-in-difference (DDD) estimator (Wooldridge 2010). Equation 1 becomes:

$$\begin{aligned} S_{itj} =& \beta_0 + \beta_1 NB_{itj} + \beta_2 R_{itj} + \beta_3 (NB_{itj} \times R_{itj}) + \beta_4 X_{ij} \\ &+ \beta_5 (X_{ij} \times NB_{itj}) + \beta_5 (X_{ij} \times R_{itj}) + \beta_6 (X_{ij} \times NB_{itj} \times R_{itj}) + \psi_j + u_{itj} \end{aligned}$$

Each DDD estimate may be expressed as a triple difference. For example, the interaction of the DDD with gender is:

$$\hat{\beta_6} = (\bar{S}_{R,F,NB} - \bar{S}_{R,F,B}) - (\bar{S}_{NR,F,NB} - \bar{S}_{NR,F,B}) - (\bar{S}_{R,M,NB} - \bar{S}_{R,M,B})$$

Where R stands for repeaters, NR for not repeaters, F for female, M for male, B for blind and NB for non-blind.

Having in mind the DID framework (figure 2) is useful to understand to what extent a difference in difference in difference (DDD) approach is informative in our context.

First, researchers may decide to add an additional comparison group and estimate treatment effects using a DDD design when the parallel trend assumption does not seem to hold. Wing, Simon, and Bello-Gomez 2018, in the set up of a policy treatment randomized across different states, take the example of a time-varying confounder that is not state-invariant. This raises a major problem in our DID method by invalidating the parallel trend assumption. In this case, it is necessary to find a comparison group that is not exposed to treatment but is exposed to the problematic time-varying confounder. This violation of the parallel trend assumption translates in our setting into a violation of the fact that blind and non blind scores do not move equally, *i.e.* do not measure the same skills. Let's take the example of kids whose main parent is a white-collar worker: they are likely to perform better in homework than kids whose parent is not a white-collar worker (for example, because the parent can help them). Thus, since non-blind scores include homework, the parallel trend might not hold.

The advantage of this identification is that it allows repeaters and non-repeaters to have different unobserved characteristics: β_6 is identified as long as these differences do not vary across the additional dimension. For instance, in our case, including a dummy for girls allow for different unobserved characteristics of repeaters vs. non repeaters which affect the scores, as long as these unobserved characteristics do not vary across genders.

The main disadvantage is that the DDD estimate requires multiple parallel trends assumptions. In the case of boys and girls, we need to assume that:

1. Without repetition, scores for girls and boys would have been parallel in blind vs. non blind tests,
2. Without repetition, scores for girls vs. boys would have been parallel in blind tests.

The second reason why we include DDD estimates is that it allows treated and control groups in a DID design to differ along a third dimension. In other terms, we are able to study the heterogeneous distribution of the bias against repeaters along different dimensions. Taking the example of the socio-economic dimension, the DDD parameter β_6 captures the comparison between:

1. The difference in scores among repeaters and non-repeaters in blind vs. non blind scores for children whose parent is a white collar, i.e. the bias against repeaters *and* kids of whitecollars,
2. The difference in scores among repeaters and non-repeaters in blind vs. non blind scores for children whose parent is a blue collar, i.e. the bias against repeaters *and* kids of bluecollars.

All covariates we use in the DDD are dummy variables except for teacher questionnaire scores which is an ordered discrete variable. As the values of this covariate are not normally distributed, we run a DDD regression for each unique possible score of the questionnaire, total of 5 different covariates since the variable can only be an integer between 0 and 5.

3.3 Effect of bias on progress

Most of the literature on bias compare a blind and a non-blind test (Goldin and Rouse 2000; Blank 1991; Lavy 2008; Falch and Naper 2013; Breda and Ly 2015). Instead, our dataset with two blind and non-blind measures allows us to study how the bias affects students' progress at a class level (see timeline in Figure 1).

In particular, we follow Terrier 2020 in modelling the student's progress. The blind score is a noisy measure of the true student ability. We assume that the student ability does not depend on the class j where she is, but the error does depend on it (for example, errors in the administration of the test may happen for the whole class).

$$B_{itj} = \theta_{it} + \epsilon_{itj} \tag{2}$$

We define the bias as the difference between a student's true ability θ_{it} and the non-blind grade NB_{it} given by the teacher.

$$Bias_{it} = NB_{it} - \theta_{it} \tag{3}$$

The progress of the student is the difference between the true abilities at the beginning and at the end of grade 6, $\theta_{i2} - \theta_{i1}$. It depends from the teacher's bias in the first period, the fact that the student is a repeater, a teacher value added T_i and the true ability in the first period. This allows for more room for improvement for low achieving students (as most repeaters are).

$$\theta_{i2} - \theta_{i1} = \beta Bias_{i1} + \eta R_i + \mu T_i + \gamma \theta_{i1} + \omega_i \tag{4}$$

We only have a noisy measure of this progress, i.e. the evolution of the blind score $B_{i2j} - B_{i1j}$

$$\begin{aligned} B_{i2j} - B_{i1j} &= \theta_{i2} + \epsilon_{i2j} - (\theta_{i1} + \epsilon_{i1j}) \\ &= (\theta_{i2} - \theta_{i1}) + \epsilon_{i2j} - \epsilon_{i1j} \\ &= \beta Bias_{i1} + \eta R_i + \mu T_i + \gamma \theta_{i1} + \epsilon_{i2j} - \epsilon_{i1j} \\ &= \beta NB_{i1} - \beta \theta_{i1} + \eta R_i + \mu T_i + \gamma \theta_{i1} + \epsilon_{i2j} - \epsilon_{i1j} \\ &= \beta NB_{i1} - \beta B_{i1} + \beta \epsilon_{itj} + \eta R_i + \mu T_i + \gamma B_{i1} - \gamma \epsilon_{itj} + \epsilon_{i2j} - \epsilon_{i1j} \\ &= \beta(NB_{i1} - B_{i1}) + \eta R_i + \mu T_i + \gamma B_{i1} + \epsilon_{i2j} + (\beta - 1 - \gamma)\epsilon_{i1j} \end{aligned} \tag{5}$$

Where we used the definition of bias (eq. 3) in the second line, the definition of progress (eq. 4) in the third line and the definition of blind score (eq. 2) in the fourth line.

We can define a first-difference specification of this progress between repeaters (R) and not repeaters (NR) by aggregating this progress at a class (j) level. We always run separate regressions for French and math.

$$[(B_{2R} - B_{1R}) - (B_{2NR} - B_{1NR})]_j = \eta + \beta[(NB_{R1} - B_{R1}) - (NB_{NR1} - B_{NR1})]_j + (B_{1R} - B_{1NR}) + (\omega_R - \omega_{NR})_j \tag{6}$$

$[(B_{2R} - B_{1R}) - (B_{2NR} - B_{1NR})]_j$ represents the repeaters' progress relative to the non-repeaters. $NB_{i1} - B_{i1}$ is the difference between the non-blind and the blind score for a student. $[(NB_{R1} - B_{R1}) - (NB_{NR1} - B_{NR1})]_j$ is the difference between the non-blind and the blind score for repeaters versus non repeaters. We define it as the bias of the teacher against repeaters. β is thus our coefficient of interest, as it captures the effect of the bias against repeaters on the relative progress of repeaters. We are able to identify it, i.e explain differences in the average progress among repeaters and non repeater with differences in average exposure to bias only if:

1. There is a quasi-random assignment of students to teachers with different degrees of bias
2. There are differences in average exposure to bias, i.e. there are more or less biased teachers.

The first can safely be assumed to be true, since, even if there is not a proper lottery, students are assigned in classes by the headmaster. Teachers cannot choose some students for their class or decide to avoid others. Moreover, we are studying the first year of middle school, so most students were enrolled in a elementary school the year before, except the probably small number of students who are repeating exactly grade 6 for the second time. Thus, headmasters have little information on students to predict students' progress and - consciously or unconsciously - assigning more promising students to less biased teachers or vice versa. We cannot test for the random assignment of students, but we can see if the assignment to a biased teacher is independent from observed characteristics.

The second is evident by plotting the bias against repeaters separately for French and Math teachers (See Figure 3, where each dot represent a teacher). Even if the average is centered close to zero, we see that there is a great deal of variation in teachers' bias.

This specification allows us to get rid of the class invariant teacher's effect. Moreover, we avoid reverse causality effects at the individual level. For example, the progress of a student may determine the non-blind score given by the teacher, as teachers may want to reward the effort. At class level, this reverse causality is attenuated. In fact, it seems unlikely that the progress of all repeaters compared to non repeaters may increase the non-blind scores given by the teacher to all repeaters.

The measure of blind scores is now averaged at the repeater-class level: we use in the regression the average of the blind score for all repeaters in a class (B_{Rt}) and for all non-repeaters in a class (B_{NRt}). In this way, the measurement error ϵ_{itj} of equation 1 is attenuated compared to an individual measure B_{it}.

4 Results

4.1 Average bias with difference-in-difference specification

4.1.1 Average bias in French

Tables 2 and 3 display the results of equation 1 for Maths and French test scores respectively. The estimate associated with the coefficient of the average bias of teachers against or towards students who have repeated the

grade, represented by Repetition × Non-Blind, is statistically significant at the 5% level without the inclusion of covariates for French scores but not for Maths scores, refer to column (1) in both tables.

From this statistically significant estimate of the coefficient we may infer a bias against repeaters for French scores.

4.1.2 Average bias in Maths

There is no apparent bias in Maths scores towards or against repeaters with and without controlling for other covariates shown in our experiment. Only when controlling for just the teacher questionnaire scores, column (4) of Table 2, did a bias appear towards repeaters who scored exactly a 4 in the teacher questionnaire via a estimate that is statistically significant at the 10% level. Clearly, this result is not sufficient to show biasedness.

4.1.3 Heterogeneous effects for French scores

In this paragraph, we will comment on heterogeneous effects for French scores: using a DDD method for Maths results does not change the significancy of our coefficient of interest, the coefficient on Repetition × Non-Blind.

For French scores, once using a DDD method, in columns (2), (3), and (5), that estimate is no longer statistically significant at any level. The shift in the distribution of scores for repeaters and not repeaters is similar for blind and non-blind scores. This provides a graphical intuition of the absence of average bias (figure 4).

Gender

In column (2) of Table 3 we run a DDD including interactions with the Girl dummy. We see that the estimate of the bias against repeaters is no longer statistically significant and its magnitude is halved. Since boys have a higher probability to repeat a year, both across all France (Seibel 1984) and in our sample (59% of repeaters are boys), does the average bias against repeaters may indeed capture a bias against boys, as found by Lavy 2008; Robinson and Lubienski 2011; Falch and Naper 2013; Cornwell, Mustard, and Van Parys 2013? The coefficient of $Repetition \times NonBlind$ represent the bias for male repeaters, and its estimate is not statistically significant. The coefficient of $Girl \times NonBlind \times Repetition$ captures instead the difference between the bias for female repeaters and the bias for male repeaters. To retrieve the bias for female repeaters, we simply need to sum these 2 coefficients. Female repeaters are expected to have a bias of 16.2% of a SD on average. To see if this is statistically significant, we define a dummy "boys", equal to 1 if the dummy girls is 0. We run the same regression using the variable "boys" instead of girls, i.e. $S_{itj} = \beta_0 + \beta_1 NB_{itj} + \beta_2 R_{itj} + \beta_3(NB_{itj} \times R_{itj}) + \beta_4 Boys_{ij} + \beta_5(Boys_{ij} \times NB_{itj}) + \beta_5(Boys_{ij} \times R_{itj}) + \beta_6(Boys_{ij} \times NB_{itj} \times R_{itj}) + \psi_j + u_{itj}$. Now, we can interpret the coefficient β_3 as the bias against female repeaters. We get that this is indeed equal to 16.2% of a SD and its estimate is statistically significant at 10% level (Table reported in the Appendix). This means that there is a bias against female repeaters.

Disciplinary sanctions

We include the interaction of Non-Blind and Repetition with the dummy taking value 1 if the child received a disciplinary warning or was temporarily excluded from the school at least once during grade 6. Once such interaction is included, the coefficient of Repetition × Non-Blind estimates the bias against repeaters who never received a punishment and such estimate is no longer statistically significant. This suggests that the bias we observe against repeaters may actually capture repeaters' disruptive behavior. The difference between the bias against repeaters and punished kids and repeaters and non-punished kids (the estimate of the coefficient Punishment × Non-Blind × Repetition) is not significant either.

Teacher's questionnaire

In column (4) of Table 3, results of equation 1 where the covariates are only those related to the teacher questionnaire scores, the average bias of teachers against student who have repeated the grade remains statistically significant but only at a 10% level. There is also a bias towards students who have repeated the grade and scored highly (either 4 or 5) in the teacher questionnaire, shown via statistically significant coefficients at the 5% level.

In column (5), when controlling for gender, disciplinary sanctions (punishment), and teacher questionnaire scores, there is a difference in bias against repeaters between those who have and have not received disciplinary sanctions, shown via the statistically significant estimate (at the 1% level) of the interaction term between punishment and repetition for non-blind scores. Since the estimate of the coefficient of Repetition × Non-Blind is not statistically significant, the bias against non-punished repeaters is not relevant, while the difference of bias between punished and non-punished repeaters is.

Monoparental families (Table 10)

The repetition in France is designed such that repeaters come more often from monoparental families (see descriptive statistics Table 1).

Including a dummy that is equal to 1 if the family is made of two parents makes the estimate of the bias against repeaters decrease and become statistically insignificant, suggesting that the bias against repeaters measured in column (1) of table 3 may partly capture an effect of bias against kids coming from monoparental families. Even

if the difference in bias against repeaters coming from monoparental vs. biparental families (estimate associated with coefficient of Repetition × Biparental × Non-Blind) is not statistically , the bias against repeaters coming from monoparental families is actually statistical significant and relevant (13.6% of a SD). We can run a regression studying only the bias for children coming from a monoparental family, i.e. regressing Non-Blind, Monoparental and Non-Blind×Monoparental on the scores (not reported). We get that the estimate of the bias against kids from monoparental families *per se* is not statistically significant. Thus, a bias seems to exist only against children who are at the same time repeaters and coming from families with just one parent. This suggests that there is a certain degree of bias against repeaters only when they come from specific disadvantaged situations.

Socio-economic characteristics (Table 11)

This intuition seems to hold when we investigate heterogeneous effects for repeaters whose main parent is not a whitecollar worker. The bias against repeaters coming from disadvantaged background is captured by the coefficient of Repetition × Non-Blind (column (1) of Table 11), which is statistically significant. Again, the estimate of the difference between repeaters whose main parent is a whitecollar or not (i.e. the coefficient Repetition × Non-Blind × White-collar) is not statistically significant.

A similar effect is found when studying how the bias against repeaters changes for children whose main parent is unemployed. The estimated bias against kids who are at the same time repeaters *and* children of unemployed people is statistically significant, while the difference among those and repeaters whose main parent is employed (Repetition × Non-Blind × Unemployed) is not.

This is bias is still not found for Maths teachers. We might wonder if this is partly driven by the fact that blind test focus more on comprehension, while grades given by French teachers may take into account - consciously or unconsciously - the ability of children to express themselves (for example, having an extensive vocabulary). For example, parents who have a high socio-economic status may have bought more books for their children and thus their children's ability of speech is higher.

Heterogeneity for honors and good conduct grades (Table 9)

Honors ("mention") in the French school system do not only reflect marks in each topic, but represent a judgement on the general attitude. There are 3 levels of honors: in our case we construct a dummy = 1 if the student received one of the three levels of honors in at least one term.

Similarly, "good conduct mark" ("note de vie scolaire") is an evaluation of assiduity, contribution to the class environment and respectful behaviour (Avvisati et al. 2014). Since this grade is skewed to the right, we construct a dummy = 1 if the grade on the conduct is 19 or 20 in at least one term. Both are attributed by the class council, and not from the single teacher (thus, we avoid problems of endogeneity between those and the bias of a single teacher).

If before we saw how the teacher bias against repeaters is heterogeneous between repeaters who has been and has not been subject to a disciplinary sanction, now we look at the upper tail of the distribution, i.e. repeaters who received "good conduct" and "honours".

The estimate of the bias against repeaters who did not receive a good conduct grade is not statistically significant. The same goes for those who did not receive a honor. Similarly, the estimate of the difference of the bias between repeaters who received an honor and those who did not is not statistically significant, and the difference for good conduct is not significant either.

Quantile regression (Table 4 and 5)

Implementing a quantile regression allows us to study heterogeneity. Since there are no particular outliers in the standard normal distribution of scores, we don't need to check the robustness of our model regarding outliers. Instead, the quantile regression provides us a set of estimated linear quantiles to see how the bias differs in different parts of the distribution of the dependent variable. Conditional symmetry is likely to hold without the need for transformations in our case in which dependent variable is a normal (Wooldridge 2010).

We interpret the quantile coefficients following the framework proposed by Koenker 2005. In case of homogeneity across the quantiles, we would expect the coefficients at each of the five estimated quantiles to be the same as in the OLS, for example as the -0.118 mean effect for French.

A great deal of variation is observed for the distribution of the bias in French. Indeed, the estimates have a lower magnitude and are not statistically significant in the tails of the distribution and in particular for the $5^{th}$and 75^{th} percentile. The only statistically significant estimate is the one for the 25^{th} percentile, while the estimate of the bias at the 95^{th} percentile is even positive, although not statistically significant.

In Maths, there is an even larger degree of heterogeneity: the bias is positive in the tails of the distribution (5^{th} and 95^{th} percentile) and the estimate of the 95^{th} percentile is statistically significant at the 5% level.

This positive bias for better performing repeaters might indicate that teachers, in particular Maths teachers, may decide to be more lenient to repeaters who are more dedicated and achieve higher grades. This effect is not confounded when we simply run a mean regression model.

Table 2: Estimation of the Repetition Bias in French Scores

	Third-term marks in French				
	(1)	(2)	(3)	(4)	(5)
Repetition	−0.010	−0.172	0.003	0.230*	0.047
	(0.131)	(0.131)	(0.131)	(0.122)	(0.131)
Non-Blind	0.036	−0.006	0.003	−0.129***	−0.142**
	(0.024)	(0.034)	(0.031)	(0.046)	(0.058)
Repetition × Non-Blind	−0.118**	−0.063	−0.034	−0.133*	0.013
	(0.049)	(0.065)	(0.064)	(0.076)	(0.103)
Heterogeneous effects for gender:					
Girl		0.344***			0.211***
		(0.034)			(0.038)
Girl × Non-Blind		0.082*			0.019
		(0.047)			(0.048)
Girl × Repetition		0.034			0.030
		(0.071)			(0.078)
Girl × Non-Blind × Repetition		−0.099			−0.044
		(0.096)			(0.097)
Heterogeneous effects for punishments:					
Punishment			−0.580***		−0.260***
			(0.048)		(0.048)
Punishment × Non-Blind			−0.359***		−0.162**
			(0.077)		(0.073)
Punishment × Repetition			0.340***		0.261***
			(0.087)		(0.084)
Punishment × Non-Blind × Repetition			−0.201		−0.303***
			(0.129)		(0.122)
Heterogeneous effects for teacher questionnaire scores:					
TQS = 1				0.031	0.021
				(0.062)	(0.071)
TQS = 2				0.281***	0.224***
				(0.067)	(0.080)
TQS = 3				0.575***	0.482***
				(0.061)	(0.072)
TQS = 4				0.736***	0.549***
				(0.053)	(0.063)
TQS = 5				1.086***	0.883***
				(0.054)	(0.066)
TQS = 1 × Non-Blind × Repetition				0.087	0.157
				(0.129)	(0.140)
TQS = 2 × Non-Blind × Repetition				0.157	0.176
				(0.143)	(0.155)
TQS = 3 × Non-Blind × Repetition				0.192	0.157
				(0.158)	(0.175)
TQS = 4 × Non-Blind × Repetition				0.276**	0.131
				(0.139)	(0.152)
TQS = 5 × Non-Blind × Repetition				0.485**	0.284
				(0.197)	(0.227)
Constant	0.686***	0.515***	0.737***	0.721***	0.663***
	(0.128)	(0.127)	(0.125)	(0.116)	(0.116)
Observations	7,566	7,566	6,219	7,566	6,219

Notes: ***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

Interaction variables between TQS and Repetition (TQS × Repetition) were not included in the table as they all yielded mostly statistically insignificant results, and those between TQS and Non-Blind (TQS × Non-Blind), despite some yielding statistically significant results, as they weren't directly relevant to the research question.

Table 3: Estimation of the Repetition Bias in Maths Scores

	Third-term marks in Maths				
	(1)	(2)	(3)	(4)	(5)
Repetition	0.035	0.061	−0.008	0.152	0.158
	(0.130)	(0.133)	(0.130)	(0.122)	(0.134)
Non-Blind	0.001	−0.140***	0.093***	0.060	0.012
	(0.024)	(0.034)	(0.033)	(0.044)	(0.063)
Repetition × Non-Blind	0.008	0.063	0.113	−0.044	0.100
	(0.049)	(0.066)	(0.074)	(0.075)	(0.119)
Heterogeneous effects for gender:					
Girl		−0.074**			−0.251***
		(0.035)			(0.039)
Girl × Non-Blind		0.274***			0.186***
		(0.047)			(0.054)
Girl × Repetition		0.017			0.048
		(0.073)			(0.080)
Girl × Non-Blind × Repetition		−0.075			−0.069
		(0.097)			(0.108)
Heterogeneous effects for punishments:					
Punishment			−0.514***		−0.272***
			(0.049)		(0.049)
Punishment × Non-Blind			−0.323***		−0.233***
			(0.064)		(0.065)
Punishment × Repetition			0.292***		0.225***
			(0.086)		(0.086)
Punishment × Non-Blind × Repetition			−0.008		−0.006
			(0.116)		(0.117)
Heterogeneous effects for teacher questionnaire scores:					
TQS = 1				0.097	0.017
				(0.062)	(0.075)
TQS = 2				0.312***	0.199***
				(0.067)	(0.082)
TQS = 3				0.632***	0.532***
				(0.060)	(0.074)
TQS = 4				0.822***	0.680***
				(0.053)	(0.066)
TQS = 5				1.226***	1.061***
				(0.054)	(0.060)
TQS = 1 × Non-Blind × Repetition				0.124	0.026
				(0.129)	(0.153)
TQS = 2 × Non-Blind × Repetition				0.085	0.064
				(0.143)	(0.169)
TQS = 3 × Non-Blind × Repetition				0.111	0.133
				(0.158)	(0.192)
TQS = 4 × Non-Blind × Repetition				0.234*	0.170
				(0.139)	(0.167)
TQS = 5 × Non-Blind × Repetition				0.265	0.259
				(0.200)	(0.267)
Constant	0.767***	0.806***	0.778***	0.699***	0.869***
	(0.128)	(0.128)	(0.123)	(0.116)	(0.119)
Observations	7,579	7,597	4,970	7,597	4,970

Notes: ***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

Interaction variables between TQS and Repetition (TQS × Repetition) were not included in the table as they all yielded mostly statistically insignificant results, and those between TQS and Non-Blind (TQS × Non-Blind), despite some yielding statistically significant results, as they weren't directly relevant to the research question.

Table 4: Quantile regression for French scores

	Dependent variable:					
	French Scores					
	(1)	(2)	(3)	(4)	(5)	DID
	5th perc.	25th perc.	50th perc.	75th perc.	95th perc.	
Repetition × Non-Blind	−0.075	−0.177**	−0.104	−0.056	0.116	−0.118**
	(0.109)	(0.069)	(0.083)	(0.067)	(0.123)	(0.049)
Repetition	−0.520***	−0.634***	−0.707***	−0.837***	−0.814***	−0.010
	(0.069)	(0.039)	(0.071)	(0.052)	(0.104)	(0.131)
Non-Blind	−0.014	0.113***	0.082*	0.110***	−0.006	0.036
	(0.059)	(0.037)	(0.046)	(0.038)	(0.073)	(0.024)
Constant	−1.495***	−0.540***	0.167***	0.815***	1.629***	−0.686***
	(0.032)	(0.021)	(0.038)	(0.031)	(0.071)	(0.128)
Observations	7,566	7,566	7,566	7,566	7,566	7,566

Note: *p<0.1; **p<0.05; ***p<0.01

Table 5: Quantile regression for Maths scores

	Dependent variable:					
	Maths Scores					
	(1)	(2)	(3)	(4)	(5)	DID
	5th perc.	25th perc.	50th perc.	75th perc.	95th perc.	
Repetition × Non-Blind	0.106	−0.046	−0.040	−0.014	0.236**	0.008
	(0.101)	(0.074)	(0.059)	(0.069)	(0.103)	(0.049)
Repetition	−0.553***	−0.685***	−0.832***	−0.810***	−0.872***	0.035
	(0.081)	(0.052)	(0.043)	(0.046)	(0.082)	(0.130)
Non-Blind	−0.110*	0.030	0.087***	0.075**	−0.090**	0.001
	(0.064)	(0.035)	(0.029)	(0.033)	(0.041)	(0.024)
Constant	−1.438***	−0.517***	0.203***	0.827***	1.700***	−0.767***
	(0.056)	(0.014)	(0.020)	(0.025)	(0.037)	(0.128)
Observations	7,597	7,597	7,597	7,597	7,597	7,597

Note: *p<0.1; **p<0.05; ***p<0.01

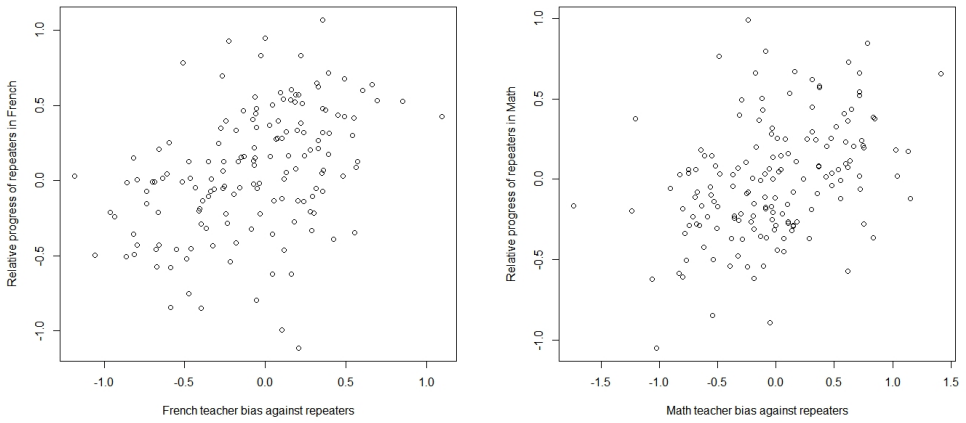
4.2 Effect of bias on progress and spillovers

We estimate equation 6 for French in the first column of table 6. The dependent variable is repeaters' relative progress between the beginning and the end of grade 6. The variable of interest is the bias of the grade 6 teacher, which is measured at the end of the year. All regressions control for the achievement gap between repeaters and non-repeaters measured at the beginning of grade 6. Since both repeaters' bias and relative progress are generated and not observed variables, we computed bootstrap estimates of the coefficients' standard errors. In particular, we employed a one-step bootstrapping method, drawing a random sample of classes with replacement 1000 times.

Teachers' biases have a high and significant effect on repeaters' relative progress: being assigned to a French teacher who is one standard deviation more biased against repeaters would decrease repeaters' relative progress by 0.306 SD, with the initial achievement gap between repeaters and not being held equal.

We estimate the same equation for maths in the third column of table 6. The effect is similar and even larger for maths. If we take two classes with the same initial achievement gap between repeaters and not, having a math teacher who is 1 SD more biased against repeaters is associated with a decrease of 0.696 SD in repeaters' relative progress in Math. We provide a graphical intuition for this result by plotting the teacher's bias against the relative progress of repeaters in their class (Figure 3).

Figure 3: Correlation between teachers' biases against repeaters and repeaters' relative progress over grade 6 in Math and French.



(a) Correlation for French

(b) Correlation for Math

In terms of internal validity, one may wonder how is it possible to have a non-significant average bias and still find a significant effect of teachers's bias on students' progress. However, we can see the great deal of variation in bias against repeaters (see Figure 3): despite the null average across all sample, the variation across classes in teachers' biased assessments might affect repeaters relative progress in these classes.

In terms of external validity, we base ourselves mainly on the study by Bénabou et al. 2004, who estimated the gap between *Zones d'education prioritaire* (ZEPs) and other schools once the priority education policy is implemented. They show that the introduction of ZEPs has no significant effect on the success of pupils. Therefore, our results on the effect of bias on student progress can be considered to be not or only marginally influenced by the fact that students are in ZEPs.

In column (2) in table 6, we run a DDD regression adding the spillovers effect of having a biased teachers in maths on the progress in French. Conversely, in column (4), we see the spillover effects of a biased French teacher on the progress in Maths. Spillovers may happen as teachers talk informally, and in each term they discuss together the assessment of the student during class council meetings. We find that having a biased French teacher has a significant effect on the progress in Math (column 4): the two biases sum up and the progress of the repeater is hindered when both happen at the same time. However, having a biased math teacher has only a mild and non- significant effect on the progress in French.

The inclusion of the bias of the teacher of the other subject does not change the effect of the gender bias on progress in one subject. This confirms that the gender bias of literacy and math teachers are independent.

Table 6: Effect of teachers' bias toward repeaters on their progress and spillover effects of bias.

	Dependent variable:			
	Relative progress in French		Relative progress in Maths	
	(1)	(2)	(3)	(4)
Repeaters bias in Maths		−0.086	0.696***	0.691***
		(0.060)	(0.031)	(0.030)
		[0.0627]	[0.0392]	[0.0382]
Repeaters bias in French	0.306***	0.313***		0.108***
	(0.060)	(0.060)		(0.029)
	[0.0538]	[0.0552]		[0.0286]
Achievement gap in French	−0.031	−0.025		
	(0.031)	(0.031)		
	[0.0348]	[0.0343]		
Achievement gap in Maths			−0.028**	−0.024*
			(0.014)	(0.014)
			[0.0122]	[0.0114]
Constant	0.015	0.021	−0.026**	−0.023*
	(0.027)	(0.027)	(0.013)	(0.013)
	[0.0287]	[0.0288]	[0.0117]	[0.0115]
Observations	178	178	178	178
R^2	0.153	0.162	0.744	0.762

Notes: The observation unit is a class. Each variable is averaged at a class level: the achievement gap is the difference between repeaters and not-repeaters in blind scores taken at the beginning of grade 6. Relative progress is the difference among the progress of repeaters and not repeaters, progress being defined as the difference between the blind scores at the end and at the beginning of grade 6. Repeaters bias is the difference between the non-blind and the blind score for repeaters versus non repeaters. *p<0.1; **p<0.05; ***p<0.01

4.3 Interpretation of the bias

Blind scores are designed and evaluated by the French Ministry of Education. Teachers may decide to use different formats for the non-blind score evaluations, even if the school curriculum and the competencies they aim to test are the same. For example, some teachers may rely more on homework or oral presentations, others on written tests that are closer to the blind scores. This would not be problematic if all teachers gave the same quantity of homework, as this would simply be constant across classes. Repeaters and non-repeaters' diligence for homework is likely to differ, since for example repeaters tend to come from more disadvantaged backgrounds (for example, they might not have a proper space to do the homework). The estimated teacher bias captures also these differences in teachers' evaluation methods.

Moreover, we are not able to disentangle the effect of teacher bias in giving grades from the teacher's biased behaviour in class. These two effects happen at the same time during grade 6, thus the progress we measure may be affected by both. It would be interesting to measure the progress of students using the difference between the blind score given at the end of grade 6 and the blind score given at the end of grade 9, as C. Terrier 2014 does. The latter is the "brévet", the final assessment of the knowledge acquired at the end of middle school. However, this blind score is inherently different from the ones in grade 6: it is high-stake for students, as their grades allow them to benefit from a merit scholarship for the rest of their schooling. Thus the difference between grade 6 blind scores and "brévet" does not only capture the progress of the student, but also the different stakes. Repeaters tend to come from disadvantaged socio-economical background, thus the incentive of the scholarship may be more important than for students coming from less disadvantaged backgrounds.

Another possible threat is that blind and non-blind scores are not measured exactly in the same period, as the blind test is taken at the very end of the school year, while the teacher's grades are given between April and June and averaged for the final grade sheet. From the descriptive statistics, it is apparent that repeaters tend to repeat 6 grade with a statistically significantly lower probability than non-repeaters. It might happen that teachers do not want to make a repeater repeat again, so they decide to be more lenient in grading repeaters in the last term. For this reason, we verify that our DiD specification yields consistent results when estimated with blind and non-blind scores of the 1st term.

5 Robustness checks

5.1 Standardization of the scores

Following Breda and Ly 2015, we standardize the scores, initially ranging between 0 and 20, to $N(0,1)$. We transform these in order to keep the ordinal information but getting rid of possible difference in grading the blind and non-blind scores (for example, different ranges of grades in blind and non-blind scores, but similar ordering). Differently from the ENS admission test studied by Breda and Ly 2015, however, 6th grade children are not explicitly ranked, but some of them may simply try to score higher than a threshold, i.e. for example the threshold for a passing grade. Since scores are standardized, the expected value is $E(S) = 0$. Thus, defining p as the share of repeaters and applying the law of iterated expectations,

$$\begin{aligned} &E(S) = pE(S|repeaters) + (1-p)E(S|non\text{-}repeaters) = 0 \\ &\iff E(S|non\text{-}repeaters) = \frac{E(S) - pE(S|repeaters)}{1-p} = -\frac{pE(S|repeaters)}{1-p} \end{aligned}$$

This binds the coefficients of our regression: for example, if we only run the regression $S = \delta_0 + \delta_1 R + u$, δ_1 is simply:

$$\begin{aligned} \delta_1 &= E(S|repeaters) - E(S|non\text{-}repeaters) = E(S|repeaters) + \frac{pE(S|repeaters)}{1-p} \\ &= \frac{(1-p)E(S|repeaters) + pE(S|repeaters)}{1-p} = \frac{1}{1-p} \end{aligned}$$

However, in our main regression $S = \beta_0 + \beta_1 NB + \beta_2 R + \beta_3 NB \times R + u$, however (column 1 of Table 2 and 3), our coefficient of interest β_3 is:

$$\begin{aligned} \beta_3 =& E(S|R=1, NB=1) - \beta_0 - \beta_1 - \beta_2 \\ =& E(S|R=1, NB=1) - E(S|R=0, NB=1) - [E(S|R=1, NB=0) - E(S|R=0, NB=0)] \end{aligned}$$

The expected value $E(S|R=0) = \frac{pE(S|R=1)}{1-p}$, but if we add another condition (for example, $E(S|R=0, NB=1)$, or also $E(S|R=0, girl=1)$ when we add controls), the expected value can be higher or lower than $\frac{pE(S|R=1)}{1-p}$.

5.2 1st Term Replication

To better interpret the bias, it is pivotal to understand whether it is mainly driven by statistical discrimination or it is taste-based. In the first case, statistical discrimination happens when teachers have imperfect information, thus use their prior ideas to get an idea of the pupil. For example, if teachers think that repeaters are lazy students, when they meet a repeater they will first think of her as lazy, but during the year they may get to know her better and update their priors. Taste-based discrimination, instead, is not driven by imperfect information: teachers remain biased even after some time spent with the student. The difference is not crystal clear, since teachers themselves probably do not know if one or the other mechanism prevails in their case. However, if the bias was driven by statistical discrimination, we would expect the bias at the end of the year to be lower and with a lower variance than the one at the beginning of the year.

Re-running the equation 1 using data from the first term, the bias at the beginning of the year is not statistically significant neither for French nor for maths (See table 5). This points to the direction that bias against repeaters in French is mainly taste based.

Table 7: Estimation of Repetition Bias in Maths and French Scores in 1st term

	Dependent variable:	
	Scores in French	Scores in Math
Repetition	0.235^{***}	-0.118^{***}
	(0.038)	(0.046)
Non-Blind	-0.003	0.005
	(0.045)	(0.049)
Repetition × Non-Blind	0.018	0.027
	(0.051)	(0.044)
Constant	1.024^{***}	0.641^{***}
	(0.027)	(0.029)
Observations	8,061	8,095

Notes: Fixed effects for class included. Errors clustered at school level. *p<0.1; **p<0.05; ***p<0.01

5.3 Number of observations

The high standard errors in our main regression leads us to say that the bias is not statistically significant in Maths and mildly significant in French. Ideally, if we could design an experiment, we would like to allocate half of the population to the treatment and half to the control group, in order to to minimize the standard error of the coefficient. In our case, 28% of our sample are repeaters, 72% are not. Especially when we investigate heterogeneous effects, the number of repeaters in each category (for example, repeaters *and* female) are even lower. Even if we always have number of observations compatible with using inferential statistical tools (always above 150 observations in each subgroup), this clearly drives up the standard errors of our coefficients. Clearly, we are in the context of a natural experiment, so we are not able to choose how to split between treated and control groups.

5.4 Quasi-random assignment of students

The quasi-random assignment of students to more or less biased teachers is a key assumption for the causal interpretation of the impact of bias on relative progress of repeaters. As mentioned in the identification section 2.3, the institutional framework of the French school system makes this assumption credible.

However, following Terrier 2020, we exploit the framework by Pei, Pischke, and Schwandt 2019 to argue in favour of the assumption. Even if it is not possible to directly test on the randomness of the assignment, we can do a left- and right-hand side balancing test.

The right-hand-side test consist in considering each class as one observation and regressing the bias against repeaters on the class-level difference between repeaters and non-repeaters characteristics.

If the coefficient is significant, it means that repeaters and non-repeaters with a given characteristic (for example, high social background) are not equally likely to be assigned a biased teacher, i.e. there is a violation of the identifying assumption.

This is not the case for the assignment of Maths teacher. The gaps between repeaters and non repeaters in performances, social background, biparental families do not have an impact on the probability of being assigned to a more biased teacher. The need-based scholarship is given to the poorest children, but the gap between repeaters and non repeaters in the scholarship recipience does not have an effect on the repeaters bias, as the same goes for the gap between the number of first-born children among repeaters and non-repeaters. The only statistical significant estimate is the one for the gap in employment of parents of repeaters and non-repeaters.

In French, instead, the achievement gap at the beginning of the year is estimated to have a significant effect on the repeaters bias, and so does the different social background of repeaters and non-repeaters. This means that the repeaters bias is incorporating information on students' initial performance. It is likely that the initial performance is correlated with the student's progress (for example, kids with initial low achievement rates have more room for improvement). Thus, there is an endogeneity issue, mitigated by the fact that low-achieving students are quasi-randomly assigned by the school headmaster to more or less biased teachers.

The estimated coefficient of the social background is significant too. As we saw in section 4.1.3, repeaters who also come from poor families tend to face bias, thus these two kinds of bias are not easy to disentangle. However, this may be mitigated by the fact that blind test focus more on comprehension, while grades given by French teachers may take into account the children's vocabulary, which is correlated with the social background of the parents.

The left-hand side balancing test consists of placing the variable on the left-hand side (LHS) of the regression instead of the outcome variable. In our case, this would mean regressing the repeaters bias and the initial achievement gap on the dummies for socio-economic status, biparental family, scholarship recipient, first child and having at least one parent employed. A zero coefficient on the causal variable of interest then confirms the identifying assumption: moreover, this test is more powerful than the righ-hand side one (Pei, Pischke, and Schwandt 2019). This difference in power is greater when the available variable is a noisy measure of the true underlying confounder. This is our case: for example, we only have a binary variable (High SES) instead of a continuous one to control for a multidimensional issue like social status and poverty.

In our case, no regression has statistically significant estimates for the impact of the repeaters bias and the achievement gap, except only a statistically significant at 10% level estimate of the correlation between repeaters bias and the dummy for being a first child for French. The other exception is the correlation of repeaters bias and achievement gap in Maths with the employment of the main parent. We get a statistically significant result for the right-hand side.

5.5 Balanced checks of attrition

Two types of attrition exist in this research: attrition at the class level, classes cease to exist after the first term, and attrition at the individual level, i.e. pupils who drop out after either the first or second term. There is no attrition at the class level in our data. Attrition at the individual level exists. Table 12 in the appendix tests if the percentage of repeaters or non-repeaters missing in a class is correlated with the degree of bias of their teacher's grading by regressing the percentages on the grade bias. For each of repeaters and non-repeaters, we ran four regressions where the percentage of missing variables act as dependent variables. The percentage of missing non-repeaters yielded statistically significant estimates of coefficients at the 5% level for Term 2 Maths grade bias and only 10% level for Term 3 French grade bias. These coefficients imply a potential attrition bias in our research, but since we do not used data from the second Term, we should only worry about the Term 3 attrition in French.

6 External validity

Our dataset is collected in the context of a relatively deprived educational district: almost two-thirds of the schools that entered the study were located in a "priority education" zone ("Zone d'Education Prioritaire", ZEP) — a label that distinguishes historically deprived areas. The policy of priority education zones was launched in 1981 by the Ministry of National Education for two main reasons: the rise in unemployment and social segregation is leading some areas to experience a concentration of social and economic difficulties and this environment had a negative impact on students' schooling, regardless of other individual and socio-economic factors. The policy of ZEP was supposed to result in more resources being distributed to those particularly disadvantaged areas. The policy was supposed to last only 4 years with a clear objective: to raise the academics level of pupils in the ZEPs. After successive reforms in 1990 and 2006, the number of ZEP has gradually increased from 350 to 1189. Now, 1 in 5 pupils belong to a ZEP school. The geographical area where our data was collected is therefore not insignificant. Precisely, our experiment is geographically based in the district of Créteil, which includes all suburbs located to the east of Paris: 10% of the 352 State-run middle schools from the Créteil district are included in our database.

The ZEPs have given rise to a large body of social science literature that allows us to develop a more accurate picture of the context from which our data is drawn. Among the information that can inform the external validity of our results, we know that classes are smaller with about two fewer students per class in 1999 with 17% more teachers for primary schools and 9% for secondary schools (Jeljoul, Lopes, and Degabriel 2001).

Thus, in terms of external validity, we should note that our results do not apply to the whole population of pupils in France, but to those in "priority education" middle schools. In terms of interpretation, for example, teachers assigned to deprived areas are on average younger than teachers in more advantaged schools (Prost 2012). This is not a minor issue. Teachers' beliefs have a large impact on their behaviour and their unconscious system of judging students (Rosenthal and Jacobson 1968). These beliefs, even unconscious ones, can be related to the age of the teachers (Boraita 2015). In "Priority Education Zones", the average age of teachers is well below the national average (Prost 2012). It is reasonable to assume that younger teachers, who have just completed their studies, are more likely to be aware of the latest scientific studies on the effectiveness of repetition. In fact, some researchers have shown that their knowledge of the state of the art on the issue of repetition has a direct impact on teachers' beliefs (Boraita 2015; Marcoux and Crahay 2008). Thus, we might be estimating a lower bound of the actual bias against repeaters.

There is an additional level of selection in our sample: the dataset was collected in the context of an experimentation for a program of parent–school meetings (Avvisati et al. 2014). Schools' headmasters needed to volunteer for having the program in their schools. We can for example imagine that only more active headmasters participated in the experiment. In this case, the decision to participate may be related to the attitude towards biased teachers. For example, more active heads of schools may both punish more biased teachers and decide to take up the program offered by Avvisati et al. 2014. This would mean that there are other schools in which teachers are less punished for being biased, thus are allowed to be more biased, thus we are measuring a lower bound of the bias against repeaters.

In terms of how these results can be generalized, the results at the national exam taken at the end of middle school in our sample are actually very similar to those obtained on average by the schools classified as ZEPs both in the district and outside of it in 2008–2009. This may mean that at least from an academic point of view our sample is representative of the population of deprived schools.

7 Conclusion

This paper continues the work previously done on studying teachers' biases in grading by focusing on bias against students that have repeated a year of their education using the same methods and data as Terrier 2020. While the results didn't prove that such a bias always exists, at least not in all subjects, i.e. Maths, it showed that class repetition can bias teachers' grading when combined with other shortcomings from the pupil, such as disciplinary sanctions. This was achieved via a difference-in-difference-in-difference (DDD) approach and a series of regressions controlling for multiple covariates including gender, disciplinary sanctions, and teacher questionnaire scores. Moreover, we found that the effect of teachers' bias against repeaters on repeaters' relative progress is relevant, while spillovers effects of the bias from one subject's teacher to another are negligible.

References

Alexander, K, D Entwisle, and N Kabbani. 2001. "The dropout process in life course perspective : early risk factors at home and school." *Teacher College Record,* no. 103, 760–822.

Avvisati, Francesco, Marc Gurgand, Nina Guyon, and Eric Maurin. 2014. "Getting parents involved: A field experiment in deprived schools." *Review of Economic Studies* 81 (1): 57–83.

Bénabou, R., M. Gurgand, F. Kramarz, and C. Prost. 2004. "Zones d'éducation prioritaire : quels moyens pour quels résultats ? Suivi d'un commentaire de Marc Gurgand." *Economie et Statistique* 380 (1): 3–34. https://doi.org/10.3406/estat.2004.7676.

Blank, R. M. 1991. "The effects of double-blind versus single-blind reviewing: Experimental evidence from the American economic review." *The American Economic Review* 81 (5): 1041–1067.

Boraita, F. 2015. "Les croyances de futurs enseignants sur le redoublement au regard de leurs connaissances sur ses effets et de leurs conceptions psychopédagogiques." *Revue des sciences de l'éducation* 41 (3): 483–508.

Braxmeyer, Nicole, Jean-Claude Guillaume, and Jean-Francois Levy. 2004. "Les pratiques d'évaluation des enseignants au collège." *Note évaluation,* no. 13, 1–5.

Breda, T., and S. T. Ly. 2015. "Professors in core science fields are not always biased against women: Evidence from France." *American Economic Journal: Applied Economics* 7 (4): 53–75.

Breda, Thomas, and Mélina Hillion. 2016. "Teaching accreditation exams reveal grading biases favor women in male-dominated disciplines in France." *Science* 353 (6298): 474–478.

Burkam, David T, Laura LoGerfo, Doug Ready, and Valerie E Lee. 2007. "The differential effects of repeating kindergarten." *Journal of Education for Students Placed at Risk* 12 (2): 103–136.

CNESCO. 2014. "Lutter contre les difficultés scolaires : le redoublement et ses alternatives ?" *Conférence de consensus.*

Cooley Fruedwirth, J., S. Navarro, and Y. Takahashi. 2011. "How the timing of grade retention affects outomes : identification and estimation of time-varying treatment effects." *Human Capital and Economic Opportunity Working Group.*

Cornwell, Christopher, David B Mustard, and Jessica Van Parys. 2013. "Noncognitive skills and the gender disparities in test scores and teacher assessments: Evidence from primary school." *Journal of Human resources* 48 (1): 236–264.

Crahay, M. 2007. "Peut-on lutter contre l'échec scolaire ?"

Davaillon, Alice, and Emmanuelle Nauze-Fichet. 2004. "Les trajectoires scolaires des enfants« pauvres»." *Education et formations* 70:41.

Dong, Y. 2010. "Kept back to get ahead ? Kindergarten retention and academic performance." *European Economic Review* 54:219–236.

Falch, T., and L. R. Naper. 2013. "Educational evaluation schemes and gender gaps in student achievement." *Economics of Education Review* 36:12–25.

Gary-Bobo, R., and J. Robin. 2012. "6. Le redoublement est-il inefficace et nuisible : Débats et difficultés d'analyse." *Regards croisés sur l'économie* 12:98–113.

Givord, Pauline. 2020. "How a student's month of birth is linked to performance at school," no. 221, https://doi.org/https://doi.org/https://doi.org/10.1787/822ea6ce-en. https://www.oecd-ilibrary.org/content/paper/822ea6ce-en.

Goldin, C., and C. Rouse. 2000. "Orchestrating impartiality: The impact of blind auditions on female musicians." *American Economic Review* 90 (4): 715–741.

Goos, M. 2013. "Grade retention. The role of the national educational policy and the effects on students' academic achievement, psychosocial functioning, and school career."

Jeljoul, M., A. Lopes, and R. Degabriel. 2001. "Quelle priorité dans l'attribution des moyens à l'éducation prioritaire ?: L'éducation prioritaire." *Education et formations,* no. 61, 83–94.

Koenker, Roger. 2005. *Quantile Regression.* Econometric Society Monographs. Cambridge University Press.

Lavy, Victor. 2008. "Do gender stereotypes reduce girls' or boys' human capital outcomes? Evidence from a natural experiment." *Journal of public Economics* 92 (10-11): 2083–2105.

Lavy, Victor, and Edith Sand. 2018. "On the origins of gender gaps in human capital: Short-and long-term consequences of teachers' biases." *Journal of Public Economics* 167:263–279.

Lorcerie, Françoise. 1998. "Sur la scolarisation des enfants d'immigrés en France." *Insaniyat* 38 (19).

Marcoux, G., and M. Crahay. 2008. "Mais pourquoi continuent-ils à faire redoubler ? Essai de compréhension du jugement des enseignants." *Revue des Sciences de l'Education* 30 (3): 501–518.

Payet, Jean-Paul, Henriot-Van Zanten, et al. 1996. "Ecole et immigration." *Revue française de pédagogie* 117 (1): 5–6.

Pei, Zhuan, Jörn-Steffen Pischke, and Hannes Schwandt. 2019. "Poorly measured confounders are more useful on the left than on the right." *Journal of Business & Economic Statistics* 37 (2): 205–216.

Prost, C. 2012. "La politique d'éducation prioritaire : quel bilan ?" *Regards croisés sur l'économie* 12:114–126.

Reynolds, Arthur J, Katherine A Magnuson, and Suh-Ruu Ou. 2010. "Preschool-to-third grade programs and practices: A review of research." *Children and Youth Services Review* 32 (8): 1121–1131.

Robinson, Joseph Paul, and Sarah Theule Lubienski. 2011. "The development of gender achievement gaps in mathematics and reading during elementary and middle school: Examining direct cognitive assessments and teacher ratings." *American Educational Research Journal* 48 (2): 268–302.

Rosenthal, R., and L. Jacobson. 1968. "Pygmalion in the Classroom." *Urban Review* 3:16–20.

Seibel, C. 1984. "Genèses et conséquences de l'échec scolaire." *Revue Française de Pédagogie* 67.

Terrier. 2020. "Boys lag behind: How teachers' gender biases affect student achievement." *Economics of Education Review* 77:101981.

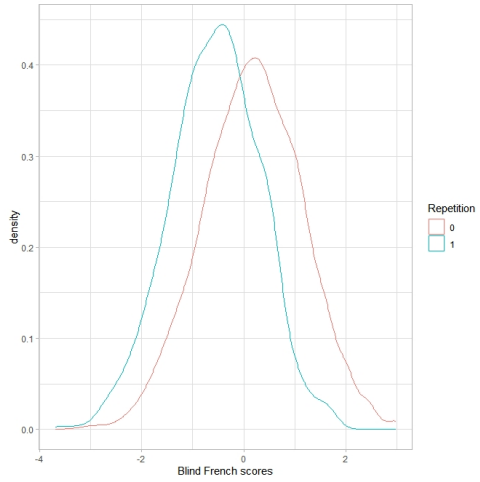
Terrier, C. 2014. "Giving a little help to girls? Evidence on grade discrimination and its effect on students' achievement." *PSE Working Papers* 36.

Wing, Coady, Kosali Simon, and Ricardo A Bello-Gomez. 2018. "Designing difference in difference studies: best practices for public health policy research." *Annual review of public health* 39.

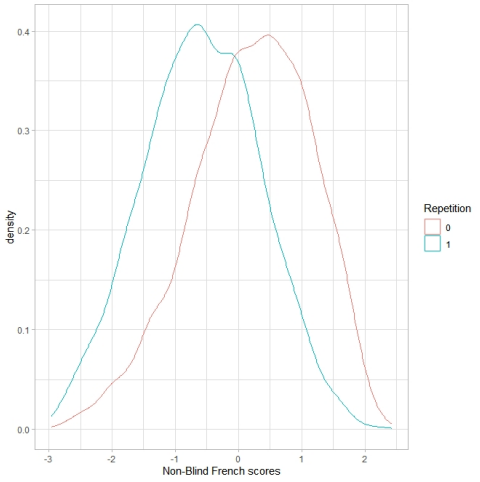
Wooldridge, Jeffrey M. 2010. *Econometric analysis of cross section and panel data.* MIT press.

Appendix

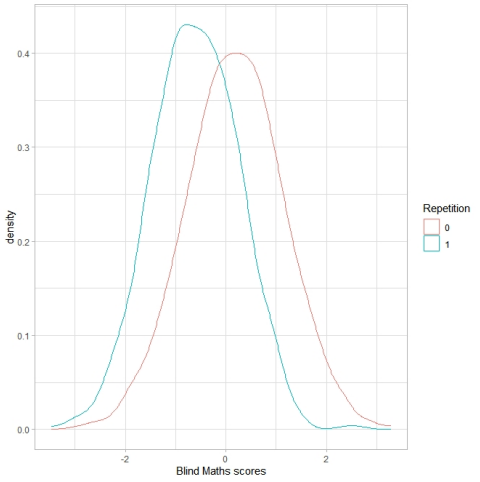
Figure 4: Distribution of 3rd term blind and non-blind scores of repeaters and non-repeaters.



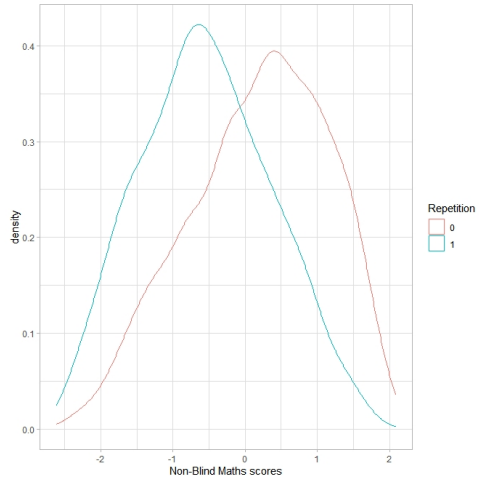
(a) Distribution of blind French scores



(b) Distribution of non-blind French scores



(c) Distribution of blind Maths scores



(d) Distribution of non-blind Maths scores

Notes: The distribution of densities of scores shows a clear shift between repeaters (in blue) and not repeaters (in red). The fact that the magnitude of the shift seems similar for blind and not blind scores confirm our main result that there is not a significant bias in teachers' evaluation.

Table 8: Heterogeneous effect of bias against repeaters along the gender axis for French in 3rd term.

	Dependent variable
	Scores
Repetition	0.206^{***}
	(0.059)
Non-Blind	0.076
	(0.060)
Non-Blind × Repetition	-0.161^{**}
	(0.071)
Boy	-0.344^{***}
	(0.030)
Boy × Non-Blind	-0.082^{**}
	(0.039)
Boy × Repetition	−0.034
	(0.081)
Boy × Non-Blind × Repetition	0.099
	(0.085)
Constant	0.859^{***}
	(0.035)
Observations	7,566

Notes: Fixed effects for class included. Errors clustered at school level. *p<0.1; **p<0.05; ***p<0.01

Table 9: Heterogeneous effect of bias against repeaters for pupils with good conduct grades and honors for French in 3rd term.

	Dependent variable:	
	Scores	
	(1)	(2)
Repetition × Non-Blind	−0.077	−0.073
	(0.075)	(0.048)
Repetition	−0.429***	−0.086**
	(0.086)	(0.042)
Non-Blind	−0.112	−0.124*
	(0.108)	(0.070)
Good Conduct	0.638***	
	(0.063)	
Non-Blind × Good Conduct	0.250**	
	(0.097)	
Repetition × Good Conduct	−0.295***	
	(0.097)	
Repetition × Non-Blind × Good Conduct	0.120	
	(0.083)	
Honors		0.885***
		(0.042)
Non-Blind × Honors		0.328***
		(0.048)
Repetition × Honors		−0.300***
		(0.086)
Repetition × Non-Blind × Honors		0.134
		(0.095)
Constant	0.047	0.183***
	(0.057)	(0.037)

Notes: Fixed effects for class included. Errors clustered at school level. *p<0.1; **p<0.05; ***p<0.01

Table 10: Heterogeneous effect for monoparental and biparental children for French

	Dependent variable:	
	French scores	
	(1)	(2)
Repetition × Non-Blind	−0.075	−0.136**
	(0.092)	(0.058)
Repetition	−0.615***	−0.714***
	(0.070)	(0.043)
Non-Blind	−0.010	0.050*
	(0.053)	(0.027)
Biparental	0.699***	
	(0.128)	
Biparental × Non-Blind	0.060	
	(0.059)	
Repetition × Biparental	−0.099	
	(0.082)	
Repetition × Biparental × Non-Blind	−0.060	
	(0.109)	
Monoparental		−0.132***
		(0.044)
Non-Blind × Monoparental		−0.060
		(0.059)
Repetition × Monoparental		0.099
		(0.082)
Repetition × Monoparental × Non-Blind		0.060
		(0.109)
Observations	7,566	7,566
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 11: Heterogeneous effects along the socio-economic axis for French.

	Dependent variable:	
	French scores	
	(1)	(2)
Repetition × Non-Blind	-0.122^{**}	-0.142^{***}
	(0.050)	(0.052)
Repetition	-0.643^{***}	-0.002
	(0.036)	(0.039)
Non-Blind	0.059	0.039
	(0.058)	(0.058)
White-collar	0.941^{***}	
	(0.056)	
White-collar × Non-Blind	-0.096^{*}	
	(0.056)	
White-collar × Repetition	-0.135	
	(0.128)	
White-collar × Repetition × Non-Blind	-0.109	
	(0.138)	
Unemployed		-0.165^{**}
		(0.065)
Non-Blind × Unemployed		-0.027
		(0.066)
Repetition × Unemployed		0.075
		(0.098)
Repetition × Non-Blind × Unemployed		0.111
		(0.084)
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 12: Balance check of attrition at the class level.

	Term 2		Term 3	
	Maths	French	Maths	French
	(1)	(2)	(3)	(4)
Dep var. % repeaters missing	−0.020	−0.008	0.015	0.008
	(0.031)	(0.031)	(0.012)	(0.006)
Dep var. % non-repeaters missing	0.167**	0.043	0.005	−0.052*
	(0.083)	(0.037)	(0.018)	(0.029)
Number of observations	1121	1140	1121	1140

Notes: This table reports estimates from a class-level regression of the percentage of repeaters (respectively non-repeaters) with a missing score on the repeater bias. This is done for both repeaters and non-repeaters. In columns 1 and 2, the dependent variable is the percentage of repeaters (respectively non-repeaters) for whom the non-blind score is missing at the end of term 2 (Math in column 1 and French in column 2). Columns 3 and 4 are the same for term 3. Stars correspond to the following p-values: *p<0.1; **p<0.05; ***p<0.01

Table 13: Right hand side balancing test for Maths.

	Dependent variable:					
	Repeaters bias in Maths					
	(1)	(2)	(3)	(4)	(5)	(6)
Achievement gap in Maths	−0.043	−0.046	−0.048	−0.045	−0.044	−0.049
	(0.034)	(0.035)	(0.036)	(0.036)	(0.036)	(0.036)
High SES gap		0.048	0.047	0.062	0.058	0.065
		(0.106)	(0.107)	(0.108)	(0.109)	(0.108)
2 parents in household gap			0.033	0.041	0.037	−0.009
			(0.078)	(0.078)	(0.079)	(0.081)
Need-based scholarship gap				0.065	0.064	0.079
				(0.073)	(0.074)	(0.073)
First child gap					−0.027	−0.023
					(0.072)	(0.072)
At least 1 parent employed gap						0.202**
						(0.098)
Constant	−0.006	−0.003	−0.002	−0.003	−0.004	0.009
	(0.031)	(0.032)	(0.032)	(0.032)	(0.032)	(0.033)
Observations	178	178	178	178	178	178
R^2	0.009	0.010	0.011	0.015	0.016	0.040
F Statistic	1.549	0.871	0.638	0.676	0.566	1.192

Notes: Fixed effects for schools included in all regressions. Errors clustered at school level. The degrees of freedom of the F-Statistics are respectively: df = 1; 176 for the 1st regression, df = 2; 175 for the second one, df = 3; 174 for the third one, df = 4; 173 for the fourth one, df = 5; 172 for the fifth one, df = 6; 171 for the sixth one. *p<0.1; **p<0.05; ***p<0.01

Table 14: Right hand side balancing test for French.

	Dependent variable:					
	Repeaters bias in French					
	(1)	(2)	(3)	(4)	(5)	(6)
Achievement gap in French	-0.098^{**}	-0.097^{**}	-0.098^{**}	-0.098^{**}	-0.102^{**}	-0.112^{***}
	(0.042)	(0.041)	(0.041)	(0.041)	(0.041)	(0.042)
High SES gap		-0.018^{**}	-0.022^{**}	-0.022^{**}	-0.023^{**}	-0.025^{***}
		(0.009)	(0.009)	(0.009)	(0.009)	(0.009)
2 parents in household gap			0.008	0.008	0.002	-0.004
			(0.006)	(0.006)	(0.007)	(0.009)
Need-based scholarship gap				-0.0002	-0.003	-0.004
				(0.008)	(0.008)	(0.008)
First child gap					0.012^{*}	0.009
					(0.006)	(0.007)
At least 1 parent employed gap						0.010
						(0.009)
Observations	178	178	178	178	178	178
R^2	0.306	0.325	0.336	0.336	0.374	0.359
F Statistic	1.864^{***}	1.970^{***}	1.993^{***}	1.926^{***}	2.198^{***}	1.993^{***}

Notes: Fixed effects for schools included in all regressions. Errors clustered at school level. The degrees of freedom of the F-Statistics are respectively: df = 1; 176 for the 1st regression, df = 2; 175 for the second one, df = 3; 174 for the third one, df = 4; 173 for the fourth one, df = 5; 172 for the fifth one, df = 6; 171 for the sixth one. *p<0.1; **p<0.05; ***p<0.01

Table 15: Left hand side balancing test for Maths.

	Dependent variable:				
	High SES	2 parents	Scholarship	First Child	Employed
	(1)	(2)	(3)	(4)	(5)
Repeaters bias in Maths	0.228	2.063	0.934	0.979	3.656^{**}
	(1.237)	(1.849)	(1.140)	(1.674)	(1.781)
Achievement gap in Maths	0.578	1.605^{***}	0.268	1.134	2.079^{***}
	(0.359)	(0.595)	(0.362)	(0.812)	(0.750)
Constant	-3.054^{***}	-11.359^{***}	-2.309^{***}	-8.966^{***}	-14.393^{***}
	(0.276)	(0.431)	(0.207)	(0.560)	(0.467)

Note: *p<0.1; **p<0.05; ***p<0.01

Table 16: Left hand side balancing test for French.

	Dependent variable:				
	High SES	2 parents	Scholarship	First Child	Employed
	(1)	(2)	(3)	(4)	(5)
Repeaters bias in French	−1.583	1.064	0.350	2.580*	1.913
	(1.057)	(1.590)	(1.300)	(1.457)	(1.685)
Achievement gap in French	−0.149	0.176	0.355	0.696	1.406
	(0.401)	(0.917)	(0.623)	(0.883)	(1.102)
Constant	−3.714***	−11.929***	−2.078***	−8.732***	−14.058***
	(0.384)	(0.814)	(0.549)	(0.710)	(0.969)
Note:					*p<0.1; **p<0.05; ***p<0.01