Students' Perceptions of Teacher Biases: Experimental Economics in Schools*

Amine Ouazad† Lionel Page‡

October 2011

Abstract

We put forward a new experimental economics design with monetary incentives to estimate students' perceptions of grading discrimination. We use this design in a large field experiment which involved 1,200 British students in grade 8 classrooms across 29 schools. In this design, students are given an endowment they can invest on a task where payoff depends on performance. The task is a written verbal test which is graded non anonymously by their teacher, in a random half of the classrooms, and graded anonymously by an external examiner in the other random half of the classrooms. We find significant evidence that students' choices reflect perceptions of biases in teachers' grading practices. Our results show systematic gender interaction effects: male students invest less with female teachers than with male teachers while female students invest more with male teachers than with female teachers. Interestingly, female students' perceptions are not in line with actual discrimination: Teachers tend to give better grades to students of their own gender. Results do not suggest that ethnicity and socioeconomic status play a role.

*We would like to thank Sandra Black, Brian Jacob, Stephen Machin, Sandra McNally, Guy Michaels, Ana Maria Santacreu, Felix Weinhardt, and especially Robert Slonim for comments on preliminary versions of this paper. We would also like to thank the audience of the Education Group at the London School of Economics, the University of Texas in Austin, the Copenhagen Business School, the 2011 Econometrics Society Australasian Meeting, the 2011 Australia New Zealand Workshop on Experimental Economics (ANZWEE). We thank Andrew Mellon and Oliver Clifton-Moore, from the U.K. Department for Education who provided financial and administrative support. We thank Jim Riches for his excellent coordination work, the Data Dissemination Unit of the Department for Education, the University of Westminster, and Amal Alia's excellent research assistance. We thank INSEAD for financial, administrative, media and computing support. The usual disclaimers apply. This paper does not represent the views of the Department for Education.

†INSEAD, London School of Economics, and CREST-INSEE.

‡Queensland Institute of Technology and University of Cambridge.

1 Introduction

There is an extensive literature studying the determinants of educational achievement. There is, in particular, an interest in the factors which foster racial, ethnic, or gender gaps in education. Most studies focus on the effectiveness of educational inputs such as teacher quality (Rockoff 2004, Hanushek & Rivkin 2006), peer effects (Epple & Romano 2010, Black, Devereux & Salvanes 2009), or parental characteristics (Black et al. 2009). A number of these inputs have a significant impact on student achievement. Yet, student effort also impacts achievement, and effort may respond strategically to educational inputs. For instance, Fryer & Torelli (2010) and Akerlof & Kranton (2000) suggest that students' behavior responds to a change in peer group characteristics[1] in a way that impacts educational achievement. And students' behavior may also respond to teacher characteristics. In psychology, the stereotype threat literature (Steele 1997, Steele & Aronson 1995) argues that female and minority students' fear that teachers' judgments will confirm racial or gender stereotypes may lead to lower performance.

Unfortunately, little economics research exists to document students' perceptions of teachers and the effect of these perceptions on effort and achievement. Recent literature on teacher's grading practices has found consistent, even if sometimes small, biases along the lines of gender, race, and ethnicity. Lavy (2008) finds that in Israel, male students are systematically given lower grades in all fields when graded non anonymously at the high-school matriculation exam and finds that these results are sensitive to the gender of the teacher. Dee (2007) also found that teachers give better grades to students of their own gender. In England, Gibbons & Chevalier (2007), using administrative data that includes a broad range of student characteristics but not teacher characteristics, found teacher biases depending on race and gender. In India, using an experimental design which randomly assigns exam contents to student characteristics, and where success at the exam is tied to financial rewards, Hanna & Linden (2009) finds that lower caste students get lower grades and thus lower rewards. In Sweden, Hinnerich, Hoglin & Johanneson (2011) also estimated teacher biases in grading using an experimental design and found significant teacher biases by student ethnicity but not by student gender.

Students also express a belief in teacher biases in subjective survey data. In the United States,

[1] Fryer & Torelli (2010) suggests that, for minority students only, higher grades have a causal negative impact on the number of friendships. Akerlof & Kranton (2002) shows that group identity affects student effort.

the educational literature has looked at students' subjective questionnaire answers on teacher biases (for instance, Wayman (2002)). There is, however, widespread skepticism in economics as to what subjective survey data identifies (Bertrand & Mullainathan 2001), skepticism which is somehow stronger when looking at subjective survey data on perceptions of racial or gender biases (Antecol & Kuhn 2000, Antecol & Cobb-Clark 2008).

This paper designs and implements a large scale experiment in the classroom, with monetary incentives, across 29 English schools with 1,200 grade 8 students, to estimate how students perceive their teacher's grading practices. In the experiment we gave students a substantial monetary endowment, and asked how much of their endowment they would like to devote to a written test. The money that is not devoted to the test is kept by the student. The money devoted to the test can double if all test answers are right, but the payoff is lower than the initial endowment if more than half of the answers are wrong. Interestingly, the teacher's grading practice can be partly discretionary, as these exam questions do not have a formally right or wrong answer. We then compare the amount of the endowment devoted to the test when students know that they will be graded non anonymously by their teacher, and when students know that they will be graded anonymously by an external examiner. Students and teachers are fully aware of the structure of the experiment, i.e. there is no deception involved (Davis & Holt 1993). The experiment was carried out in controlled conditions – no interactions between students, large classroom, scripted experimental instructions – in the classroom with students and their usual teacher, close to the definition of an artefactual experiment (Levitt & List 2009). Importantly, the set of students taking part in the experiment reflects the overall composition of the student population in England.

Our results suggest a strong gender effect: male students tend to invest less when graded by a female teacher than the anonymous examiner, and female students tend to invest more when graded by a male teacher than when graded by anonymous external examiners. This suggests that students believe that there are teacher biases in grading practices. Male students anticipate tougher grading from a female teacher, and reduce their investment when graded by a female teacher. Conversely, female students seem to anticipate more lenient grading when the teacher is male and increase their investment accordingly. Also, teachers gave better grades to students of their own gender. Hence, male students' choices are consistent with female teachers' grading practice, but female students' choices are not consistent with male teachers' grading practices. Interestingly, there is no significant

effect of ethnicity and/or of socioeconomic status on students' perceptions of teachers' grading.

This experiment adds to the large number of studies using the methodology of laboratory experiment in the field (Harrison & List 2004) . The number of field experiments is expanding particularly fast in the economics of education (Bettinger & Slonim 2006, Bettinger & Slonim 2007, Hoff & Pandey 2006, Fryer 2010) as classrooms provide a convenient setting where conditions can be controlled while preserving external validity.

Our experimental design takes the form of a variant of the trust game (Berg, Dickhaut & McCabe 1995) in the classroom. In the trust game, a truster can decide to send money to a trustee. The amount sent is multiplied by some factor, and the trustee can chose to send back everything, nothing or any amount in between back to the truster. The trust game has been used to measure trust and perceptions of trustworthiness in different social contexts (Bohnet, Greig, Herrmann & Zeckhauser 2008, Bohnet & Zeckhauser 2004). Experiments using trust games have also been specifically designed to estimate individuals' perceptions of discrimination and discriminatory behavior, for instance in Israel (Fershtman & Gneezy 2001). A key difference between this paper's experiment and a trust game is that, in line with usual grading situations, we removed any teachers' monetary incentives to diminish the rewards of the students.

The paper is structured as follows. Section 2 presents the experimental design, the additional administrative data, the internal validity of the experiment, and presents descriptive statistics on students' choices and payoffs in the experiment. Section 3 estimates the effect of the non anonymous condition on student choices, by teacher and student gender, and by socioeconomic status & ethnicity. We estimate students' subjective probability of success at the test by estimating a structural expected utility model on our experimental data. The section also describes teachers' actual grading practices and comparesthese with students' perceptions of teachers' grading practices. Section 7 discusses the internal and external validity of the experiment, the importance of non monetary incentives. Section 8 concludes by discussing the policy implications of our results.

2 Experimental Design

We design a 90-minute experiment that comprises of two sessions and a questionnaire: A first session, where students know that they will be graded anonymously by the external examiner. A second

session, where a random half of the students know that they will be graded non anonymously by their teacher and another random half of the students know that they will be graded anonymously by the external examiner. After these two sessions, students fill a survey questionnaire.

2.1 Background Information

Around 1,200 grade 8 students across 29 schools in London, Manchester and Liverpool took part in the experiment. Students and schools came from all parts of the ability distribution. Participating schools had a wide variety of achievement levels and a wide variety of social backgrounds. In England a common measure of achievement in secondary education is the number of five or more GCSEs (General Certificate of Secondary Education) with grades from A to C, called 'good' GCSEs. The highest performing school was an all-girls Church of England school which had 75% of students with five or more GCSEs grade C or above. The median school was a mixed community school, with 54% of students having five or more good GCSEs. Finally, the lowest performing school was a mixed community school, which had 38% of students with five or more good GCSEs.

Table 1 shows that the demographic composition of our schools does not strongly differ from the characteristics of the English student population. Our schools have more ethnic diversity than the average English secondary school, and have slightly lower achievement. This is due to the number of schools in the London area. There are about 194 grade 8 students on average in our schools, which is a slightly lower number of grade 8 students than in the overall population. We have 13% of free meal students in our experiment, compared to 17% of free meal students in the population of English students. We have fewer White students in our sample than in the population of English students (64% versus 84%), and slightly more male students in the sample than in the grade 8 population (54% versus 51%). Overall achievement scores at grade 6 national examinations (also known as Key Stage 2 in England) are slightly lower than the national average.

2.2 The First Session

Prior to the experiment, parents sign a parental agreement[2] that clearly spells out the conditions of the experiment, including the use of monetary incentives. Head teachers and teachers agree with the format of the experiment.

[2]Only one out of 1,200 students' family refused to sign the agreement.

We go to each school with four experts in education. Two experts are presenters, and two experts are anonymous external graders. The presenters are recruited from a larger set of former principals, inspectors, or teachers and are specifically trained to present the experiment to students in the same way in each classroom. We flip one coin to randomize the allocation of external examiners to classrooms and one coin to randomize the allocation of presenters to classrooms. Presenters do not grade and graders do not present.

The experiment proceeds as follows. In each school, we work with two classes of approximately 20 students. The experiment starts and ends at the same time in both classrooms. The experiment takes place in large classrooms. The teacher of the classroom is present from the beginning of each experiment, but keeps silent. The teacher is either the main teacher of the grade or the English teacher. Before entering the classroom, students are handed a table number. They then enter the classroom in silence and sit at the table corresponding to their number. Students are only identified by their number and never by their name – thus the experimental procedure is anonymous. Numbers are assigned randomly so that students are not able to choose where they want to sit. This limits the potential for cheating and peer effects. Sealed envelopes containing the questions and the answer sheets are on each table.

A presenter, in each classroom, reads the experimental instructions aloud. The timeline presented in the appendix (page 38) is strictly followed. The experiment is about defining words presented in a paragraph that contains the word. An example question, "archaeologist", is then read aloud by the presenter. A few students provide potential answers, and the presenter does not say which answer is better than the others. Each question is a word definition, as in the previous example.

We purposedly chose a task, defining words, where there is no formal right or wrong answer. This potentially gives teachers the possibility of adopting different grading practices with different students. Choosing a task where grading practices depend on the teacher is critical for the study of students' behaviour whenpotentially facing a teacher bias. In practice, we observe that word definitions are graded differently by different graders. Indeed, a grader can, for instance, choose to give the point to students who give the definition that is consistent with the context only. For instance, "demonstration" has two different meanings, depending on the context. The word "demonstration" is presented in a paragraph where it means "a public meeting or a march protesting

against something." Graders decide in each case whether the acceptable answer should be consistent with the context. We do not provide guidelines. Graders can require definitions that are full sentences, graders can also sanction definitions based on examples, such as examples of "species" rather than a definition of "species."

The presenter then tells students that he will give them £2. Students are able to keep this endowment or students can choose to buy questions at a cost of 20p each.

A right answer leads to a gain of 40p, whereas a wrong answer leads to no money. There are 10 potential questions, so that a student can get up to £4. Students do not know the questions ex-ante, and cannot choose which questions they want to answer. The presenter describes a couple of scenarios, e.g. the student chooses to buy 4 questions, gets 3 questions right. The presenter asks students to calculate how much they would get. The payoff is $2 - 4 \times 0.20 + 3 \times 0.40 = 2.40$ pounds. Thus the presenter makes sure that students understand the game. The payoff of a student who buys n questions and gets $k \leq n$ answers right is:

$$c(n, k) = 2 - 0.20 \cdot n + 0.40 \cdot k$$

Finally, students then choose the number of bought questions by circling a number between 0 and 10 at the bottom of the envelope. Students are informed that this choice cannot be changed later on.

How many questions do you want to buy?										
0	1	2	3	4	5	6	7	8	9	10

Once the number of questions to buy has been circled, students open the envelope containing the answer sheet. They have 20 minutes to write down in silence their definitions. Students answer questions 1 to 4 if they chose 4 questions. They cannot choose the specific questions to answer.

We chose a reasonably long duration of 20 min to ensure that students do not need to consider a time constraint when making their choices.

The words are taken from all subjects, from science, geography, history, and English.[3]Also, the

[3]When we carried out the experiment, the words were species, monologue, ridge, gravity, paranoia, eroded, unemployment, recycling, demonstrations, tax. These words come the last ten years of English national examinations (Key Stage 3).

design is such that both difficult and easy questions are present.[4] In some cases of students with special educational needs, an adult reads the text – but not any answer – quietly to the student.

Envelopes are then collected and given to the anonymous external marker. This completes the first round.

It is important to stress that no feedback is given at the end of the first round. Feedback on outcomes is only provided at the end of the second round, once students have left the classroom. Payoffs are handed at the end of the experiment for all students, regardless of their choices, to avoid differences in choices due to impatience (Bettinger & Slonim 2007).

2.3 The Second Session

Students are then told that there will be a second round, with the same guidelines, and a different set of questions. Each student gets a new envelope and a new endowment. In one randomly selected classroom, the "treatment" classroom, students are told that answers will be corrected by their teacher. In that classroom, students write their name and their teacher's name at the top of envelope. The rest of the session then proceeds as before: students choose a number of questions from 0 to 10, and then have 20 minutes to fill in the answer sheet. Words of the second round are different than words of the first round.[5]

Envelopes are collected. Students leave the classroom and keep the paper bearing their table number. Envelopes are given either to the teacher or the external examiner, who corrects them. The presenter calculates the payoffs, fills envelopes with the monetary payoffs. Envelopes bear the student's number. Envelopes are handed to the student. This completes the second round.

To summarize, students of the treatment classroom are in the nonanonymous condition in the second round, and students of the control classroom are in the anonymous condition in the second round.

Anonymous Condition *Grading is performed by an external examiner who does not see the student or his/her name. Students are identified only by their table number.*

[4]For instance, monologue was an especially difficult word (with a low success rate), gravity was a particularly easy one (with a very high success rate), paranoia was difficult, unemployment and recycling were easy, demonstrations was difficult (in the context of the excerpt), and tax was found to be moderately difficult.

[5]The words were customary, stone's throw, wrestling, earthquake, single, charisma, fictional character, legacy, rhyme, curfew.

Non anonymous Condition *Students write their name and their teacher's name on the envelope and answers are corrected by the teacher.*

And in the first round, students are always in the anonymous condition.

	Control Classroom	Treatment Classroom
Round 1	Anonymous Condition	Anonymous Condition
Round 2	Anonymous Condition	Non anonymous Condition

Both classrooms start and end the experiment at the same time, which prevents contamination of the control by the treatment.

We observe each student's choice and outcome twice. In the treatment classroom, we observe students' choice and outcome once in the anonymous setting, and once in the non anonymous setting. In the control classroom, we observe students' choice and outcome twice in the anonymous setting.

2.4 Complementary Data: Survey Questionnaire, Administrative Data and Teachers' and External Examiners' Grading

At the end of the second round, students fill a survey questionnaire after the second round, and before envelopes are handed, hence payoffs do not affect answers to the questionnaire.[6] Questions of the survey questionnaire assess students' stated perceptions of the role of hard work, luck, their perceptions of the teacher's fairness, whether different ethnicities have equal opportunities, and whether they feel that their effort at school is not rewarded. We also ask students how they perceive their own ability, and how much weekly pocket money they get. The average weekly pocket money we estimated using our data was close to the average amount from a survey by Halifax Bank[7]. Only a small number of students reported weekly pocket money conditional on good behavior or conditional on participation in the duties of the house – cleaning their room, washing the dishes, etc. [8]

[6]Because of experimental constraints, half of the students filled the survey questionnaire.

[7]Halifax Pocket Money Survey 2008, available at http://www.lloydsbankinggroup.com/media/pdfs/halifax/2008/August/25_08_08_Halifax_pocket_money_survey_2008.pdf .

[8]Presenters also lead a discussion about students' feelings about the experiment; whether they enjoyed it, what they felt the purpose of the experiment was. students said they enjoyed the game, the presence of monetary rewards;

We merge the experimental results with administrative data on students, from the English National Pupil Database.[9] As a requirement to participate to the study, every school gave an agreement to provide the name and the national unique pupil number of the students participating to the experiment. In practice, 85% of schools provided us with a complete list of the names and numbers of the students. We are able to match those students to their test score on national examinations in 2009, just one year before the experiment, and to get their ethnicity, gender, free meal status. When the data is not available, we code ethnicity and gender through classroom observation and names. For ethnicity, we break down the sample into white students and nonwhite students; and also into narrower categories: White, Asian, Black, Mixed, or Other. The free meal status is given to students whose parents or carers are on income-based job seeker's allowance, income support, and other welfare benefits. It is a proxy for economic deprivation which comprises about 17% of the student population.

Finally, students' answers and teachers' grades were coded question by question, for each round and in each condition, so that the final file includes the whole sequence of right and wrong answers.

2.5 Descriptive statistics

Students' choices are summarized in Table 3. Over the two rounds, students choose an average of 6.3 questions, with a standard deviation of 3.2.Students with higher prior grade 6 test scores bought more questions. A 1 standard deviation increase in prior standardized score increases the number of questions bought by 0.5 in the first round and by 0.8 in the second round. The correlation between prior score and the number of questions is significantly stronger in the second round.

On average, students had 3.57 good answers, representing a success rate of 54%. Thus the questions are neither too easy, nor too hard. Students get in the envelope an average of £4.33. They earn a bit more than if they had not bought any question. Thus if students were risk neutral and success rate did not depend on the number of questions chosen, students should actually have bought 10 questions in each session Table 3 shows the distribution of payoffs in the first and the second round.

our most significant finding is that the presence of monetary rewards made most students interested in understanding and defining words, including students who would not otherwise be easily motivated. Students declared that defining words was neither too easy nor too hard.

[9]This database is central in most papers estimating school quality in England, see for instance Machin & McNally (2005) and Kramarz, Machin & Ouazad (2010).

3 Results

3.1 Students' Perception of Teachers' Grading Practices: Identification Strategy

To identify the effect of the non anonymous condition on students' perceptions of teachers' grading practices, we study how their choice of investment (number of questions bought) vary between the anonymous and the non anonymous condition (defined in section 2.3 on page 8). Students are graded in the non anonymous condition in the second round of the treatment classroom only, hence the 2-round design of our experiment allows us to get a within-student estimate of the treatment effect. The effect is estimated by comparing the change in the number of questions chosen in the first and the second round in the treatment and in the control group. A within-student estimate leads to more precise estimates than an estimate relying on one round of observation.

This amounts to estimating the following regression:[10]

$$\begin{aligned} \text{Questions}_{i,t} &= \textit{constant} + \alpha \cdot \text{Round } 2_{i,t} + \gamma \cdot \text{Treatment}_{i,t} \\ &\quad + \delta \cdot \text{Round } 2 \times \text{Treatment}_{i,t} \\ &\quad + u_i + \varepsilon_{i,t} \end{aligned} \tag{1}$$

$\text{Questions}_{i,t}$ is the number of questions bought by student i in round $t = 1, 2$. The coefficient of interest is δ, the effect of the non anonymous condition on the number of questions bought. Because the treatment is randomly assigned (see section 4.1 on page 18), we can model u_i as a random effect. The random effects estimator is a more efficient estimator than the fixed effects estimator. Estimation with fixed effects confirms that the results are robust to the use of fixed effects. α controls for the difference in average behavior between the second and the first round, a difference which is partly due to students experiencing the first round and learning about the task. Interestingly, α also controls for learning when students of different characteristics learn differently.[11]

[10] Results based on a Poisson count data model right-censored at 10 with student fixed effects yield very similar results.

[11] To see that, assume that $\alpha = a + a_i$, where $E(a_i) = 0$, and a is a constant. a_i is the student-specific learning control, with $E(a_i) = 0$. Then the residual is $\eta_{i,t} = \varepsilon_{i,t} + a_i \text{Round } 2_{i,t}$. Algebra shows that the randomization of the treatment ensures that $\text{Treatment}_{i,t}$ and $\text{Round } 2 \times \text{Treatment}_{i,t}$ are independent of $\eta_{i,t}$. Hence the treatment effect δ is consistently estimated by OLS.

The average effect is not significantly different from zero (0.035 with a standard error of 0.15), which suggests that students do not make wild assumptions on the behavior of the external grader compared to their teacher. More interestingly, the non significant average effect masks considerable variation in the way students of different characteristics have responded to the non anonymous condition. We estimate equation 1 on different subsamples defined by the teacher's and student's gender, the student's ethnicity and free meal status.

3.2 Perceptions by Student Gender

Student gender has been shown to be one of the key variable affecting the grading practices of teachers in the previous literature. Our results indicate that students do form beliefs over teachers' leniency/toughness in grading, beliefs which differ according to their own gender and the gender of their teacher.

Effects by teacher and student gender are presented in Table 4. Each cell is a separate regression according to baseline specification 1.[12] When graded by a male teacher, female students tended to buy 0.843 more question when graded by the teacher than when graded by the external examiner. The treatment effect is statistically significant at 5%. When graded by a female teacher, male students tended to buy 0.601 less question than when graded by the external examiner. Overall, since the number of female teachers was higher than the number of male teachers, students graded by a male teacher bought significantly more questions in the non anonymous condition than in the anonymous condition (+0.576).

3.3 Perceptions by Parental Income and by Student Ethnicity

A key question is whether students from different ethnic and social backgrounds perceive teacher biases against their group. This question is particularly relevant for ethnic minorities and students from low social backgrounds. A negative perception of their teachers could cause a lower investment in the education process and deepen inequalities in educational achievement.

To test for an effect of students' socio-economic background on students' perceptions, we use students' free school meal eligibility. Free meal eligibility is based on parental income & recipiency

[12] This allows the coefficient α, measuring students' 'learning' in-between the two rounds, to differ across genders. A single regression where the Round 2 $\times$ Treatment$_{i,t}$ variable is interacted with students' and teachers' gender has also been carried out, yielding very similar results.

of welfare benefits and represents about 17% of the student population. It is therefore a good proxy for poverty and deprivation. The bottom part of Table 7 estimates result for free meal and non free meal students. As for table 4, each cell is the effect of the non anonymous condition for a separate regression. Results suggest that there is no effect of poverty status on the number of questions bought.

Table 7 also displays the same analysis for white and for nonwhite students. As mentioned in the introduction, the stereotype threat literature (Steele & Aronson 1995, Steele 1997) finds that African American students' fear of confirming racial stereotypes of underachievement may negatively affect their achievement. Another psychology literature suggests that even arbitrary group affiliation may affect the way people treat others (Tajfel 1982). We find no such effect of ethnicity on students' choices. There is no effect regardless of whether we consider the whole nonwhite category or whether we consider a breakdown of nonwhite students by racial subgroup.[13] These results are significant as they suggest that students from all different ethnic background believe that they have equal chances in the educational system in England. This is confirmed in the answers from the survey questionnaire. When answering the question "Do you think that pupils with the same ability but different ethnicities are equally likely to succeed at school", students from ethnic minorities overwhelmingly answered positively.

3.4 Estimating Students' Subjective Probabilities of Success

Previous analyses found an effect of the non anonymous condition on the number of questions bought. The difference-in-differences estimation allows us to estimate an average effect in terms of number of questions. Whilst the magnitude of this effect is informative, it is unclear how this difference translates in terms of beliefs. Clues as to what beliefs translate into students' investment choices can be found by considering that choosing a number of questions to buy between 0 and 10 is making a trade-off between risk and return.

We estimate a structural model of choice where students choose the number of question which maximizes their utility. Doing so we are able to convert the treatment effects of Table 4 into differences in subjective probability of success with their respective teachers.

[13] Indian, Pakistani, Black, and Black Caribbean students have very different achievement levels in England. We find no effect when considering these subgroups.

We formalize the student's choice in an expected utility framework where students choose a trade-off between the risk and the return of buying more questions. We assume a random utility model where the utility of choosing n questions is:

$$U_n = E\left[u\left(c(n,k)\right)\right] + \varepsilon_n \tag{2}$$

where $k \leq n$ is the number of right answers, $c(n,k) = 2 - 0.20 \cdot n + 0.40 \cdot k$ is the payoff when n questions are bought and k answers are right, u is the Von-Neumann Morgenstern utility function defined on the payoff, and ε_n a random factor. Assuming that students form a subjective probability $\hat{p}$ of getting a right answer on any question, the subjective probability of getting k answers right when buying n questions is $P(k|n) = \left(\begin{array}{c} n \\ k \end{array} \right) \hat{p}^k (1-\hat{p})^{n-k}$. As mentioned in section 2.2, our observation suggests that students did not need more than 20 minutes to fill the answer sheet.

The probability $P(n;\hat{p};r)$ of choosing n questions depends on his subjective probability of a right answer $\hat{p}$ and his relative risk-aversion r. $E(n) = \sum_{n=0}^{10} P(n;\hat{p};r) \cdot n$ is the average number of questions bought for students who believe that the subjective probability of a right answer is $\hat{p}$ and r is relative risk aversion. The average number of questions bought increases when the subjective probability $0 \leq \hat{p} \leq 1$ of a right answer increases, and the number of questions bought decreases when risk aversion r increases.

The subjective probability $\hat{p}$ of a right answer depends on whether the observation belongs to the treatment or control classroom, whether the observation is in the second round, and whether the observation is for treatment classroom in the second round. That gives a specification for $\hat{p}$ which is similar to the baseline specification of equation 1. There is a different $\hat{p}$ for each round and for the control and treatment classrooms.

$$\begin{aligned} \hat{p}_{i,t} &= a + b \cdot \text{Round 2} + c \cdot \text{Treatment}_{i,t} \\ &\quad + d \cdot \text{Round 2} \times \text{Treatment}_{i,t} \end{aligned} \tag{3}$$

To make things amenable to estimation, we assume that the utility function exhibits constant relative risk aversion (CRRA), so that $u(c) = \frac{c^{1-r}}{1-r}$, and r is relative risk-aversion. We estimate the parameters $\hat{p}, r$ by maximum likelihood, assuming that ε_n is i.i.d. extreme value distributed as in

Andersen, Fountain, Harrison & Rutström (2010) . Fechner errors or normally distributed errors can also be used, without significant changes in the point estimates presented below.

Standard errors are clustered by classroom. The coefficient of interest here is d, the effect of the treatment on the subjective probability of a right answer. We also parameterize risk aversion by gender, to control for potential differences in risk attitudes by gender.

$$r_i \quad = \quad constant + g \cdot \text{Male}_i$$

where g measures the difference in risk aversion between male and female students. Our assumption that risk aversion is stable between the two rounds and across treatment and control is supported by the data: A regression for a different level of risk aversion for each round gives point estimates that are not statistically different.

Results are presented in Table 8. Risk aversion estimates suggest that students are risk loving, i.e. they have negative risk aversion. Such a result is not uncommon in situations where participants are given an endowment to play with. This is due to the so called house money effect (Thaler & Johnson 1990), the fact to play with an amount of money recently received. In our experiment, students are not playing with their own money but rather with an endowment of £2 in each round.

The subjective probability of a right answer is estimated to be 62% (column 1) over the whole sample. This is above the estimated success rate of 52 and 57% in the first and second round respectively, indicating some degree of overconfidence

Results also show that students have a significantly higher subjective p when graded by a male teacher. According to our results, students believe that a question graded by a male teacher is 6 percentage points more likely to be deemed right. Students also believe that a question graded by a female teacher is 3.5 percentage points less likely to be deemed right. This is consistent with the non-structural estimates of Table 4.

Our results indicate that the gender effects observed in the difference in differences model can be linked with very substantial differences in subjective beliefs. In the non anonymous condition, female students behave as if they had an increase of 10 percentage point in their subjective probability of success when the teacher is a male. Conversely, male students behave as if they had a 16.5 percentage point decrease in their subjective probability of success. These results confirm the significant effect

of the non anonymous treatment on the students subjective beliefs in their chance of success. Female students behavior suggests that they believe that their chance of success is significantly higher with a male teacher. Conversely, male students seem to believe that they are significantly less likely to succeed if the teacher is a female.

3.5 Grading Practices

We chose not to perform double grading of answer sheets in order to preserve teachers' anonymity and thus avoid teachers' strategic response to double grading. However, comparing the number of right answers across the anonymous and non anonymous condition is not appropriate if one wants to compare grading practices across external examiners and teachers. Indeed, both grading practices and students' choices vary across the two conditions.

To solve this issue, we compare grades given in the two conditions, question by question, which substantially alleviates the previous issue. The control and the treatment groups are randomly allocated, hence comparing grading question by question across the two conditions is likely to give us a good estimate of the teacher's grading practice vis a vis the external examiner.

Table 9 shows p_{teacher}, the fraction of right answers when corrected by the teacher and $p_{\text{external examiner}}$, the fraction of right answers when corrected by the external examiner. For the first question, the teacher graded the answer right in 48% of cases, and the external examiner graded the answer right in only 39% of cases. The difference is 8 percentage points and strongly significant.

Overall, for all questions, the teacher marked the answer right with a 6 percentage point higher probability than the external examiner. The difference is significant at 5% for several questions, but is only significant at 10% overall.

Previous literature on teacher biases has found a tendency for teachers to advantage female students (Lavy 2008). To assess whether teachers' grading practices differ over different subset of students, we regressed the probability of a right answer on student gender, a non anonymous condition dummy, the prior grade 6 score, and interactions between the non anonymous condition and the prior score, and between the non anonymous condition and the teacher's gender.

$$\begin{aligned}\text{Question } k \text{ Right}_{i,round\,2} &= \text{constant} + a \cdot \text{Male}_i + b \cdot \text{Non Anonymous Condition}_i \\ &\quad + c \cdot \text{Grade 6 Score}_i\end{aligned}$$

$$+d \cdot \text{Non Anonymous Condition}_i \times \text{Grade 6 Score}_i$$
$$+f \cdot \text{Non Anonymous Condition} \times \text{Male}_i$$
$$+g \cdot \text{Non Anonymous Condition} \times \text{Male}_i \times \text{Female Teacher}_i$$
$$+g \cdot \text{Non Anonymous Condition} \times \text{Female}_i \times \text{Male Teacher}_i + \varepsilon_i \quad (4)$$

where, as before, i indexes students, and ε_i is the residual. Prior grade 6 score is broken down into quartiles, so that Grade 6 Score$_i$ is a set of dummies for the second, third, and fourth quartile of prior achievement.

Table 10 presents the results for three words. Results for other words are available from the authors and do not significantly differ. Again, students are more likely to get the answer right when corrected in the non anonymous condition: Teachers' likelihood of giving the point is 7 to 22 percentage points higher. And male teachers were even more lenient for words 'customary' and 'single', increasing this likelihood by another 8 to 16 percentage points. Male students are less likely to get the answer right in the non anonymous condition on some questions, a finding consistent with Lavy (2008), who finds that male students tend to get lower grades when graded non anonymously.

More able students are more likely to get the answer right, a student in the top quartile of the grade 6 scores is from 21 to 24 percentage points more likely to get the answer right. This is the same effect in the anonymous and the non anonymous condition, revealing that teachers grade students of different ability levels the same way as the external examiner; the difference between the external examiner and different types of teachers is that teachers give higher grades on average.

All in all, results on teachers' grading practices by gender partially match our main results, presented in section 3. Male students' choices are consistent with the perception of an actual bias. In classrooms where their teacher was female, male students invested less when they knew that the teacher would grade their paper knowing their name (the *non anonymous condition*). Female students' choices, on the other hand, are inconsistent with teachers' actual bias. Our results suggest that female students' choices would be consistent with male teachers giving them higher grades, while they actually receive higher grades from female teachers.

Overall our results confirm previous studies showing that teachers grading practices vary over different subset of students. Our experimental design allows us to test whether in return students

form beliefs about the existence of such differentials in grading practices.

4 Discussion

4.1 Internal Validity

A possible concern about our results is whether our randomization process was successful. In spite of our random allocation of the treatment and presenters by coin toss, one could wonder whether we have successfully eliminated systematic differences in students and presenters characteristics between the treatment group and the control group. To test for this, we first compare the characteristics of students between the treatment and the control group, including their gender, ethnicity, and prior grade 6 score. The results, displayed in Table 2 indicate that there is indeed no significant differences between the characteristics of the students in the treatment group and the students in the control group.

As a second check of the internal validity of the experiment, we perform a placebo test by noticing that there should be no treatment effect in the first round. There would be an effect if presenters or classroom effects rather than teachers are driving the treatment effects. The sixth row of Table 2 shows that the number of questions chosen in the first round is not significantly different between the control and the treatment classroom. Also the last two rows show that there is no treatment effect in the first round in schools which, in the second round, have a male teacher in the non anonymous condition. This indicates that the different effects observed accross teachers from different genders does not come from systematic differences in the characteristics of their students.[14]

4.2 External validity

When discussing our results, it is important to consider in what extent the effect found in this experiment can be generalised. A key question is whether the teachers participating in our experiment have specific characteristics which would make them very unrepresentative of the the overall population of male and female teachers in England. In field experiments it is typically difficult to estimate representative parameters on a non randomly selected subpopulation.

[14] Also, the average difference between the treatment and the control group is the same in the first and in the second round.

Several elements concur to suggest that there is no reason to doubt of the external validity of our results. First, we have selected schools from different regional areas with very different ethnic and socio-economic compositions. The selection process specifically aimed to form a sample for which the experimental results could be used to inform the policy makers in England. The section 2.1 describes how the selected schools have a large range of characteristics and how as a consequence the population of students participating to the experiment does not markedly differ from the population of English pupils. Second, the effect is present for every male teacher of the sample, and the treatment effect is large – above one additional question – for two teachers of the sample. Third, male teachers are observed in very different schools across the sample: in community schools, voluntary aided schools, grammar schools, and specialist schools, in London, Manchester and Liverpool.[15] Fourth, teacher gender is not correlated with students' prior achievement. The p-value of the t-test of the equality of prior scores for students graded by a male teacher versus students graded by a female teacher is 0.1546; the absence of such correlation is important since we find treatment effects that depend on ability – so that the male teacher effect that we find is not due to some correlation between prior achievement and the teacher's gender.

Overall there is no indication that the students and teachers participating in the experiment have characteristics unrepresentative of the population of students and teachers in England .

4.3 Gender Results by Teachers' Subject

A possible confounding factor for our result on the effect of the teachers' gender could be that male and female teachers in our sample tend to teach different disciplines. Female teachers are for instance more likely to be English teachers than male teachers. Statistically, female students outperform male students in all disciplines, but at grade 9 (GCSE) exams, the gender gap is larger in English, and the Humanities. As a consequence, a male student could form lower expectation about his chances of success with an English teacher because English teachers are 'tougher' graders for him than teachers of other subjects, e.g. mathematics. As a consequence, we could observe a negative effect of female teachers on boys' investment. In this case, the subject area of the teacher would be driving the results rather than the gender of the teacher.

The experimental data includes the subject taught by the teacher. The different subjects are:

[15] We preserve the anonymity of schools in the paper.

English, mathematics, history, humanities, business studies, information and communications technology (ICT). We estimate Table 4 on two subsets. The first subset is made of students who are in schools where the teacher of the non anonymous condition is an English or a Humanities teacher. These are two subjects where the gender gap at grade 9 is higher than the gender gap in the other subjects. The second subset is made of students who are in schools where the teacher of the treatment classroom is a mathematics, business studies, or ICT teacher.

Results for the subset of English teachers and humanities teachers are shown in Table 5. Results are not significantly different from the results for the overall sample reported in Table 4, and are, if anything, stronger. The effect of the nonanonymous condition for male students graded by a female teacher is -1.202 questions. The effect of the nonanonymous condition for female students graded by a male teacher is 1.469 questions. Interestingly, in Table 6, where the effect is estimated on mathematics, business studies, or ICT teachers, gender interaction effects are not significant.

4.4 Monetary versus Non monetary Incentives

A key assumption to link our results to students' differences in perceptions of their teachers grading practices is that the monetary payoffs of the game was the main motivation of students' choices. However, one could wonder whether non monetary incentives could play a key role in students' choices. A student may want to please or impress the presenter (Levitt & List 2007), please the teacher relatively more than the presenter, signal his/her ability (Feltovitch & Harbaugh 2002), signal hard work or conform to group norms when graded by the teacher (Austen-Smith & Fryer 2005). Several elements indicates that such non monetary incentives are unlikely to be driving the results.

First, the monetary incentives given in the experiment are substantial for 13 year old students. Students can earn up to £8, which represents 1.25 times students' average weekly pocket money (around £6), and represents around a third of the weekly disability living allowance in the U.K.[16]. In 2003, 2.5 million individuals in the U.K. were receiving the disability living allowance, which is partly a substitute for unemployment benefits (Benitez-Silva, Disney & Jimenez-Martin 2010). From our personal experience and the feedback we received from students, the prospect to win real

[16]Source: UK government's digital service for people in England and Wales accessible at http://www.direct.gov.uk/en/disabledpeople/financialsupport/dg_10011925.

money was a key motivator for students and it prompted them to think carefully about the best option to maximize their payoffs.

Second, it must be stressed that non monetary incentives in themselves would not necessarily biasing the results. If some students have a desire to please the presenter similar to the one to please the teacher, the randomness of the assignment to the treatment and control, and the within-student design control for this. Non monetary incentives could naturally be stronger in the second round when students are marked by their teachers, but even in this situations, these non monetary incentives bias our results only if they are systematically different over different subgroup of students. If they tend to be the same for all students, they should be averaged out in the difference in differences estimation due to the random allocation of students.

Third, to check for the possibility that different subset of students may have different level of non monetary incentives, we use the answers from the post experiment survey. The survey includes a question about the desire of the student to value the relationship with the teacher independently of the monetary incentives of the experiment : "A good relationship with the teacher matters (Strongly Disagree... to Strongly Agree)." To see whether the answer to this question is correlated with the effect of the non anonymous condition for female students when graded by male teachers, we proceed in the following way. We focused on the sample of female students in schools where the teacher was male, and, for each question, we split the sample into two parts. Students whose answer is below the median answer (they disagree more than the median student), and students whose answer is above the median answer (they agree more than the median student). We then estimated the treatment effect for those two subgroups, question by question. The results are presented in Table 11. Remarkably, the treatment does not differ by the answer to the survey questionnaire, and treatment effects are still significant and positive for each subgroup. This indicates that the expressed differences in the belief that a good relationship with the teacher matters did not drive the observed differences in number of question chosen by female students.

5 Conclusion

Using a deception-free incentive-compatible experimental design in 29 English schools with 1,200 students, we estimated the effect of students' perceptions of teacher biases on student investment

in the classroom. Our results suggest that students from low-income families and minority ethnic backgrounds do not believe in systematic teacher biases. This result is significant given that in some countries, including the United States, studies have found that minority students state beliefs in detrimental teacher biases (Wayman 2002). Our result may either indicate that such biases do not exist to the same extent as in England, or that our experiment gives us a better indication of students' underlying beliefs than traditional survey questionnaires. Unlike surveys, our design provides students with monetary incentives to reveal their beliefs.

Previous economics of education literature on teacher biases suggests that in some contexts teachers give better grades to students of their own gender (Dee 2007). We find that students' perceptions strongly depend on their gender and their teacher's gender. Male students invest less when graded by a female teacher, and female students invest more when graded by a male teacher. These results imply that male students have lower expectations about their chances of success when graded by a female teacher while female students have higher expectations about their chances of success when graded by a male teacher. Interestingly, an analysis of teachers' grading practices shows that these belief only partially match teachers' actual behavior. Indeed, teachers are more lenient with students of their own gender. Male students' choices are in line with the fact that male teachers give them lower grades, but female students' choices are not consistent with male teachers' grading practice.

A breakdown by teachers' subjects reveals that gender interaction effects are driven by the subset of English and Humanities teachers, and that there is little effect for other subjects, i.e. mathematics, business studies, and ICT. Interestingly, gender gaps in achievement at school are much stronger in English and the Humanities. Indeed, in virtually all fields, including mathematics and science, girls outperform boys at GCSE examinations,[17] i.e. a higher fraction of girls achieve 5 A-C GCSEs, so-called 'good GCSEs'. At grade 2 and grade 6 national examinations (Key Stages 1 and 2), girls significantly outperform boys in English, but boys are only slightly ahead of girls in mathematics (Machin & McNally 2005). All in all, our results are consistent with the possibility that gender interactions play a stronger role in English and the Humanities classes, and shape educational outcomes more strongly.

Overall, results shed new light on the nature of gender interactions in the classroom. Students'

[17] General Certificate of Secondary Education (GCSE) are taken in grade 9.

responses to teachers' characteristics are an important determinant of their effort, all the more that students' actions need not be consistent with teachers' actions and perceptions. Importantly, the two effects we find go in the same direction: they both increase the gender gap in student investment; Indeed, with a male teacher, the gap between boys' and girls' effort increases because girls invest more; with a female teacher, the gap increases because boys invest less.

The growing gender gap in education has become a concern for policy makers (Weaver-Hightower 2003). Further research may help explain what shapes students' perceptions, whether and how misperceptions can be corrected, and how much these perceptions affect student effort and investment in other contexts.

References

Akerlof, G. A. & Kranton, R. E. (2000), 'Economics and identity', *Quarterly Journal of Economics* pp. 1–39.

Akerlof, G. A. & Kranton, R. E. (2002), 'Identity and schooling: Some lessons for the economics of education', *Journal of Economic Literature* **40**(4), 1167–1201.

Andersen, S., Fountain, J., Harrison, G. W. & Rutström, E. E. (2010), 'Estimating subjective probabilities', pp. 1–59.

Antecol, H. & Cobb-Clark, D. A. (2008), 'Identity and racial harassment', *Journal of Economic Behavior & Organization* **66**(3-4), 529–557.

Antecol, H. & Kuhn, P. (2000), 'Gender as an impediment to labor market success: Why do young women report greater harm?', *Journal of Labor Economics* **18**(4), 702–728.

Austen-Smith, D. & Fryer, R. G. (2005), 'An economic analysis of "acting white"', *Quarterly Journal of Economics* pp. 551–583.

Benitez-Silva, H., Disney, R. & Jimenez-Martin, S. (2010), 'Disability, capacity for work and the business cycle: an international perspective', *Economic Policy* **July**.

Berg, J., Dickhaut, J. & McCabe, K. (1995), 'Trust, reciprocity, and social history', *Games and Economic Behavior* **10**, 122–142.

Bertrand, M. & Mullainathan, S. (2001), 'Do people mean what they say?', *American Economic Review* **91**(2), 67–72.

Bettinger, E. & Slonim, R. (2006), 'Using experimental economics to measure the effects of a natural educational experiment on altruism', *Journal of Public Economics* **90**(8-9), 1625–1648.

Bettinger, E. & Slonim, R. (2007), 'Patience among children', *Journal of Public Economics* **91**, 343–363.

Black, S. E., Devereux, P. J. & Salvanes, K. G. (2009), 'Like father, like son? a note on the intergenerational transmission of iq scores', *Economics Letters* **105**(1), 138–140.

Bohnet, I., Greig, F., Herrmann, B. & Zeckhauser, R. (2008), 'Betrayal aversion: Evidence from brazil, china, oman, switzerland, turkey, and the united states', *American Economic Review* **98**(1), 294–310.

Bohnet, I. & Zeckhauser, R. (2004), 'Trust, risk and betrayal', *Journal of Economic Behavior & Organization* **55**(4), 467–484.

Davis, D. D. & Holt, C. A. (1993), 'Experimental economics'.

Dee, T. (2007), 'Teachers and the gender gaps in student achievement', *Journal of Human Resources* .

Epple, D. & Romano, R. E. (2010), 'Peer effects in education: A survey of the theory and evidence', *Handbook of Social Economics* pp. 1–186.

Feltovitch, N. & Harbaugh, R. (2002), 'Too cool for school? signalling and countersignalling', *RAND Journal of Economics* **33**(4), 630–649.

Fershtman, C. & Gneezy, U. (2001), 'Discrimination in a segmented society: An experimental approach', *The Quarterly Journal of Economics* .

Fryer, R. (2010), 'Financial incentives and student achievement: Evidence from randomized trials', *NBER Working Paper Series* .

Fryer, R. G. & Torelli, P. (2010), 'An empirical analysis of 'acting white'', *Journal of Public Economics* **94**(5-6), 380–396.

Gibbons, S. & Chevalier, A. (2007), 'Teacher assessments and pupil outcomes', *Centre for the Economics of Education Working Paper* **December**.

Hanna, R. & Linden, L. (2009), 'Measuring discrimination in education', *NBER Working Paper Series* .

Hanushek, E. A. & Rivkin, S. G. (2006), 'Teacher quality', *Handbook of the Economics of Education, Volume 2* **2**.

Harrison, G. W. & List, J. A. (2004), 'Field experiments', *Journal of Economic Literature* **XLII**, 1009–1055.

Hinnerich, B. T., Hoglin, E. & Johanneson, M. (2011), 'Ethnic discrimination in high school grading: Evidence from a field experiment', pp. 1–36.

Hoff, K. & Pandey, P. (2006), 'Discrimination, social identity, and durable inequalities', *American Economic Review* **96**(2), 206–211.

Kramarz, F., Machin, S. & Ouazad, A. (2010), 'Using compulsory mobility to identify the relative contribution of pupils and schools to test scores', pp. 1–58.

Lavy, V. (2008), 'Do gender stereotypes reduce girls' or boys' human capital outcomes? evidence from a natural experiment', *Journal of Public Economics* pp. 1–23.

Levitt, S. D. & List, J. A. (2007), 'What do laboratory experiments measuring social preferences reveal about the real world?', *Journal of Economic Perspectives* **21**(2), 153–174.

Levitt, S. & List, J. (2009), 'Field experiments in economics: The past, the present, and the future', *European Economic Review* **53**(1), 1–18.

Machin, S. & McNally, S. (2005), 'Gender and student achievement in english schools', *Oxford Review of Economic Policy* **21**(3), 357–372.

Rockoff, J. (2004), 'The impact of individual teachers on student achievement: Evidence from panel data', *The American Economic Review* **94**(2), 247–252.

Steele, C. M. (1997), 'A threat in the air. how stereotypes shape intellectual identity and performance', *American Psychologist* **52**(6), 613–629.

Steele, C. M. & Aronson, J. (1995), 'Stereotype threat and the intellectual test performance of african americans', *Journal of Personality and Social Psychology* pp. 1–15.

Tajfel, H. (1982), 'Social psychology of intergroup relations', pp. 1–41.

Thaler, R. H. & Johnson, E. J. (1990), 'Gambling with the house money and trying to break even: The effects of prior outcomes on risky choice', *Management Science* **36**(6), 643–660.

Wayman, J. C. (2002), 'Student perceptions of teacher ethnic bias: A comparison of mexican american and non-latino white dropouts and students', *The High School Journal* **85**(3).

Weaver-Hightower, M. (2003), 'The "boy turn" in research on gender and education', *Review of Educational Research* **73**(4), 471.

Table 1: Student Characteristics

	Sample				School		Year 8 Population	
	Mean	S.D.	Min	Max	Mean	S.D.	Mean	S.D.
School Demographics								
Students per school	44.06	9.86	20.00	60.00	194.50	52.46	202.04	66.59
Students per classroom	22.34	4.99	10.00	30.00	-	-	-	-
Student Demographics								
Free meal	0.13	0.34	0.00	1.00	0.27	0.44	0.17	0.37
White	0.64	0.48	0.00	1.00	0.69	0.30	0.84	0.37
Nonwhite	0.31	0.46	0.00	1.00	0.31	0.46	0.16	0.37
- Black	0.07	0.25	0.00	1.00	0.04	0.20	0.02	0.13
- Asian	0.10	0.30	0.00	1.00	0.08	0.27	0.06	0.23
- Mixed	0.04	0.20	0.00	1.00	0.05	0.21	0.02	0.15
Male	0.54	0.50	0.00	1.00	0.50	0.50	0.51	0.50
Prior Achievement (Grade 6)								
Test Score	54.16	43.11	0.00	99.00	-	-	59.55	17.13

Source: Experimental data for columns 1 to 4, and student Level Annual School Census (PLASC), Department for Education for columns 5 to 8.

Table 2: Randomization of the Treatment

	Treatment group	Control group	p-value of the difference
Randomization			
Free school meal	0.512	0.547	0.618
	(0.02)	(0.02)	
	[597]	[557]	
Key Stage 2 score	87.27	86.46	0.361
	(0.63)	(0.63)	
	[597]	[557]	
White	0.682	0.659	0.524
	(0.02)	(0.03)	
	[597]	[557]	
Male	0.513	0.547	0.352
	(0.02)	(0.02)	
	[597]	[557]	
Classroom Size	38.4	37.9	0.909
	(2.50)	(2.73)	
	[597]	[557]	
Placebo Tests			
Questions Bought	6.46	6.33	0.453
in 1st Round	(0.12)	(0.12)	
	[597]	[557]	
Questions Bought in 1st Round,	6.37	6.21	0.564
School with Male Teacher in 2nd Round	(0.21)	(0.20)	
	[225]	[204]	
Questions Bought in 1st Round,	6.30	6.60	0.160
School with Female Teacher in 2nd Round	(0.157)	(0.146)	
	[225]	[204]	

Confidence intervals in parenthesis. Number of observations in brackets.

Table 3: Choices and Outcomes

	Mean	S.D.	Min	Max
First Round				
Questions Purchased	6.39	2.93	0.00	10.00
Good answers	3.43	2.32	0.00	10.00
Fraction Right	0.52	0.23	0.00	1.00
Payoff (£)	2.09	0.59	0.00	4.00
Second Round				
Questions Purchased	6.25	3.45	0.00	10.00
Good answers	3.73	2.70	0.00	10.00
Fraction Right	0.57	0.26	0.00	1.00
Payoff (£)	2.24	0.66	0.00	4.00

Table 4: Main Result – Effect of Non anonymous Grading by the Teacher by Teacher and Student Gender

	Teachers			
Students	All	Male	Female	Δ = Male − Female
All	0.036	0.576	-0.318	0.894
	(0.150)	(0.233)**	(0.197)	(0.297)**
Observations	2,292	856	1,396	2,292
Male	-0.086	0.487	-0.601	1.088
	(0.232)	(0.312)	(0.268)**	(0.446)**
Observations	1,031	486	801	1,031
Female	0.359	0.843	0.110	0.733
	(0.230)	(0.371)**	(0.268)	(0.413)*
Observations	873	278	595	873

Each coefficient comes from a separate regression for the treatment effect on each subsample.
Reading: Being graded by the teacher increases the number of questions bought by 0.036 question. Being graded by a male teacher increases the number of questions bought by 0.576 question.
***: Significant at 1%. **: Significant at 5%. *: Significant at 10%.

This table reports the effect of the non anonymous condition for each group of students and each group of teachers. Coefficients of the first five rows are the coefficients of separate regressions $questions_{i,t} = \alpha \text{Round } 2_{i,t} + \delta \text{Round } 2 \times \text{Treatment}_{i,t} + u_i + \varepsilon_{i,t}$.

Table 5: Main Result – Effect of Non anonymous Grading by the Teacher by Teacher and Student Gender – For English & Humanities Teachers

	Teachers			
Students	All	Male	Female	Δ = Male − Female
All	0.041	0.844	-0.403	1.247
	(0.231)	(0.287)***	(0.318)	(0.462)**
Observations	811	285	526	811
Male	-0.163	0.473	-1.202	1.675
	(0.345)	(0.341)	(0.553)**	(0.682)**
Observations	402	169	233	402
Female	0.686	1.469	0.379	1.090
	(0.282)**	(0.499)***	(0.553)	(0.606)*
Observations	405	112	293	405

Each coefficient comes from a separate regression for the treatment effect on each subsample.
Reading: Being graded by the teacher increases the number of questions bought by 0.036 question. Being graded by a male teacher increases the number of questions bought by 0.576 question.
***: Significant at 1%. **: Significant at 5%. *: Significant at 10%.

This table reports the effect of the non anonymous condition for each group of students and each group of teachers. Coefficients of the first five rows are the coefficients of separate regressions $questions_{i,t} = \alpha \text{Round } 2_{i,t} + \delta \text{Round } 2 \times \text{Treatment}_{i,t} + u_i + \varepsilon_{i,t}$.

Table 6: Main Result – Effect of Non anonymous Grading by the Teacher by Teacher and Student Gender – Teachers of Mathematics, Business Studies, Information and Computer Technology

	Teachers			
Students	All	Male	Female	Δ = Male − Female
All	0.048	0.437	-0.251	0.688
	(0.195)	(0.319)	(0.251)	(0.460)
Observations	1,481	571	870	1,481
Male	-0.038	0.488	-0.319	0.807
	(0.306)	(0.442)	(0.299)	(0.462)
Observations	671	317	568	671
Female	0.110	0.446	-0.124	0.570
	(0.352)	(0.518)	(0.299)	(0.698)
Observations	468	166	302	468

Each coefficient comes from a separate regression for the treatment effect on each subsample.
Reading: Being graded by the teacher increases the number of questions bought by 0.036 question. Being graded by a male teacher increases the number of questions bought by 0.576 question.
***: Significant at 1%. **: Significant at 5%. *: Significant at 10%.

This table reports the effect of the non anonymous condition for each group of students and each group of teachers. Coefficients of the first five rows are the coefficients of separate regressions $questions_{i,t} = \alpha \text{Round } 2_{i,t} + \delta \text{Round } 2 \times \text{Treatment}_{i,t} + u_i + \varepsilon_{i,t}$.

Table 7: Effect by Ethnicity and by Free Meal Eligibility

Students	Treatment Effect
White	0.097
	(0.178)
Observations	1,614
Nonwhite	-0.100
	(0.284)
Observations	678
Eligible for Free Meals	0.238
	(0.390)
Observations	290
Non Eligible for Free Meals	0.007
	(0.163)
Observations	2,002

Each coefficient comes from a separate regression for the treatment effect on each subsample.
***: Significant at 1%. **: Significant at 5%. *: Significant at 10%.
This table reports the effect of the non anonymous condition for each group of students and each group of teachers. Coefficients are the coefficient δ of regression $questions_{i,t} = \alpha \text{Round } 2_{i,t} + \delta \text{Round } 2 \times \text{Treatment}_{i,t} + u_i + \varepsilon_{i,t}$.

Table 8: Estimation of the Expected Utility Model

	– Whole Sample –			Male Teacher	Male Teacher	Female Teacher	Female Teacher	White	Nonwhite	Free Meal	Non Free Meal
Dependent variable : p											
Constant	0.626	0.629	0.627	0.638	0.637	0.616	0.606	0.650	0.641	0.595	0.656
	(0.006) ***	(0.007) ***	(0.011) ***	(0.018) ***	(0.018) ***	(0.015) ***	(0.014) ***	(0.016) ***	(0.019) ***	(0.040) ***	(0.014) ***
Treatment			0.014	-0.020	-0.019	0.036	0.033	0.008	-0.021	0.093	-0.020
			(0.019)	(0.031)	(0.031)	(0.023)	(0.022)	(0.026)	(0.037)	(0.049) *	(0.023)
Round 2			-0.012	-0.040	-0.062	0.006	-0.026	-0.013	-0.006	-0.029	-0.003
			(0.010)	(0.020) **	(0.040)	(0.012)	(0.015)	(0.016)	(0.019)	(0.024)	(0.014)
Treatment × Round 2			0.001	0.061	-0.074	-0.035	0.017	0.036	0.008	0.020	0.024
			(0.016)	(0.027) **	(0.076)	(0.021) *	(0.025)	(0.022)	(0.032)	(0.042)	(0.020)
Treatment × Round 2 × Female					0.100						
					(0.049) **						
Treatment × Round 2 × Male							-0.165				
							(0.053) ***				
Dependent variable : r											
Constant	-0.574	-0.520	-0.565	-0.544	-0.546	-0.583	-0.640	-0.639	-0.611	-0.604	-0.653
	(0.015) ***	(0.056) ***	(0.019) ***	(0.034) ***	(0.034) ***	(0.033) ***	(0.032) ***	(0.024) ***	(0.055) ***	(0.054) ***	(0.026) ***
Male		-0.094									
		(0.088)									
Number of Observations	2,292	2,292	2,292	856	856	1,396	1,396	946	469	290	1,083

Reading: Students graded non anonymously by a male teacher believe the probability of a right answer is 6.1 percentage points higher than when they are graded anonymously by an external examiner. Students graded non anonymously by a female teacher believe the probability of a right answer is 6.1 percentage points higher than when they are graded anonymously by an external examiner.
Notations: p is the subjective probability of a right answer, r is risk aversion. Both are parameterized so that we estimate the effect of the non anonymous condition on the probability of a right answer, keeping risk aversion constant. Maximum likelihood standard errors clustered by student. ***: significant at 1%, **: significant 5%, *: significant at 10%.

Table 9: Comparing Grading Practices - The Teacher vs External Markers

Question	Word	p_{Teacher}	$p_{\text{External Examiner}}$	Difference	p-value
1	customary	0.48	0.39	0.08	0.01
2	stone's throw	0.36	0.33	0.03	0.30
3	wrestling	0.75	0.76	-0.01	0.71
4	earthquake	0.84	0.77	0.07	0.01
5	single	0.64	0.47	0.17	0.00
6	charisma	0.34	0.23	0.11	0.00
7	fictional character	0.74	0.76	-0.02	0.63
8	legacy	0.43	0.47	-0.04	0.41
9	rhyme	0.63	0.52	0.11	0.02
10	curfew	0.52	0.45	0.07	0.17
Overall		0.57	0.52	0.06	0.06

$p_{Teacher}$ is the fraction of answers deemed right by the teacher. $p_{External\,Examiner}$ is the fraction of answers deemed right by the external examiner. The p-value is the p-value of the t-test of the significance of the difference of the fractions in the non anonymous groups and in the anonymous groups.

Table 10: Comparing Grading Practices - The Teacher vs External Markers

VARIABLES	(1) customary	(2) customary	(3) customary	(4) single	(5) single	(6) single	(7) rhyme	(8) rhyme	(9) rhyme
Male	0.108**	0.108**	0.108**	0.060	0.060	0.060	0.098	0.098	0.098
	(0.047)	(0.047)	(0.047)	(0.053)	(0.053)	(0.053)	(0.073)	(0.074)	(0.074)
Nonanonymous condition	0.072	0.053	0.095	0.175**	0.140*	0.171**	0.202*	0.223**	0.250**
	(0.063)	(0.063)	(0.067)	(0.072)	(0.073)	(0.077)	(0.104)	(0.103)	(0.106)
Nonanonymous × Male Student	-0.122*	-0.133*	-0.036	-0.097	-0.120	0.046	-0.111	-0.095	-0.189
	(0.068)	(0.068)	(0.080)	(0.074)	(0.075)	(0.085)	(0.104)	(0.106)	(0.118)
Nonanonymous × Male Teacher		0.086*			0.158***			-0.109	
		(0.050)			(0.052)			(0.079)	
Nonanonymous × Male Student × Female Teacher			-0.177***			-0.225***			0.022
			(0.067)			(0.072)			(0.108)
Nonanonymous × Female Student × Male Teacher			-0.018			0.083			-0.204*
			(0.073)			(0.072)			(0.107)
2nd Quartile of Prior Score	-0.060	-0.060	-0.060	0.039	0.039	0.039	0.220**	0.220**	0.220**
	(0.062)	(0.062)	(0.062)	(0.073)	(0.073)	(0.073)	(0.101)	(0.101)	(0.102)
3rd Quartile of Prior Score	0.119*	0.119*	0.119*	0.194**	0.194**	0.194**	0.194*	0.194*	0.194*
	(0.068)	(0.068)	(0.068)	(0.077)	(0.077)	(0.077)	(0.102)	(0.102)	(0.102)
4th Quartile of Prior Score	0.220***	0.220***	0.220***	0.198***	0.198***	0.198***	0.217**	0.217**	0.217**
	(0.065)	(0.065)	(0.065)	(0.070)	(0.070)	(0.070)	(0.095)	(0.095)	(0.095)
Nonanonymous × 2nd Quartile of Prior Score	0.153*	0.138	0.127	0.088	0.060	0.053	-0.225	-0.204	-0.201
	(0.090)	(0.090)	(0.090)	(0.101)	(0.101)	(0.101)	(0.147)	(0.147)	(0.147)
Nonanonymous × 3rd Quartile of Prior Score	-0.024	-0.046	-0.049	0.019	-0.024	-0.025	-0.067	-0.029	-0.011
	(0.096)	(0.097)	(0.097)	(0.104)	(0.105)	(0.105)	(0.146)	(0.151)	(0.149)
Nonanonymous × 4th Quartile of Prior Score	0.108	0.091	0.081	-0.012	-0.042	-0.047	0.009	0.027	0.032
	(0.092)	(0.092)	(0.092)	(0.098)	(0.098)	(0.098)	(0.132)	(0.133)	(0.132)
Observations	846	846	846	702	702	702	353	353	353
R-squared	0.054	0.057	0.062	0.058	0.069	0.072	0.056	0.062	0.065

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

The dependent variable is 1 if the answer was deemed right by the grader (either the teacher in the non anonymous condition, or the external grader in the anonymous condition). Observations come from the second round of the experiment , where students are randomly assigned to the anonymous or the non anonymous condition.

Table 11: Treatment Effect for Female Students graded by a Male Teacher
– by Answer to the survey questionnaire

	Treatment Effect of Male Teachers for Female Students (Additional Number of Questions)
Good relationship with the teacher matters	
More than the median student	0.801
	(0.442)*
Less than the median student	0.892
	(0.467)*
The advice and help of my teacher have played an important role in my progress	
More than the median student	0.849
	(0.432)**
Less than the median student	0.834
	(0.482)*
Number of Observations	278

Appendix A: Survey Questionnaire

Comparing Survey Questionnaire Answers and Students' Choices

Survey questionnaire may substitute for the experiment if survey questionnaire answers are sufficiently predictive of the number of questions purchased.

Table 12 shows the results of the regression of the number of questions bought on the answers to the survey questionnaire. Each questionnaire answer except pocket money is coded from -1 (Strongly disagree) to +1 (Strongly agree). The answers predict 6% of the variance of the number of questions bought. A perception that hard work determines success is positively correlated with a larger number of questions bought. If the answer goes from 0 (Neither agree nor disagree) to 1 (Strongly agree), the number of questions bought increases by 0.768. A perception that luck determines success is negatively correlated with the number of questions bought. This is likely explained by the fact that a stated stronger role of luck increases the subjective variance of the payoff. If the answer goes from 0 (Neither agree nor disagree) to 1 (Strongly agree), the number of questions bought goes down by 0.619 question. Similarly, a perception that ethnicities have equal opportunities, or that a good relationship with the teacher matters is positively correlated with the number of questions bought.

Pocket money is not correlated with the number of questions bought. Indeed, the effect of pocket money can theoretically increase or lower the marginal utility of money for the student: higher amounts of pocket money may be correlated with higher income, or, on the contrary, may be correlated with less parental investment in the child's education.

Table 12: Survey Questionnaire

	Mean	S.D.	Min	Max	N
According to you, how important are the following in helping you to do well at school?					
Luck	0.05	0.65	-1.00	1.00	633
Hard Work	0.82	0.31	-0.67	1.00	635
Good relationship with the teacher	0.33	0.52	-1.00	1.00	636
Do you think that pupils with the same ability but different ethnicities are equally likely to succeed at school?					
Strongly disagree (-1) to Strongly agree (1)	0.68	0.50	-1.00	1.00	631
Do you think that, as a student, you are					
Very weak (-1) to Good (1)	0.32	0.24	-0.67	1.00	631
Do you think your teachers expect you to do well at school?					
Not at all (-1) to Yes very much (1)	0.68	0.40	-1.00	1.00	629
Sometimes my effort at school is not given a proper reward					
Strongly disagree (-1) to Strongly agree (1)	0.28	0.48	-1.00	1.00	633
The advice and help of my teacher have played an important role in my progress					
Not at all (-1) to Yes very much (1)	0.50	0.45	-1.00	1.00	630
Pocket money per week	5.89	7.87	0.00	60.00	679

Each answer to the survey questionnaire except pocket money was coded from -1 (Strongly Disagree) to +1 (Strongly Agree).

Table 13: Correlations of Students' Choice with the Answers to the Survey Questionnaire

	Dependent variable: Questions
Hard work determines success	0.753
	(0.325)**
Luck determines success	-0.579
	(0.144)***
Ethnicities have equal opportunities	0.567
	(0.190)***
Good relationship with the teacher matters	0.379
	(0.193)**
Advice of the teacher helped	0.285
	(0.229)
Thinks teacher has high expectations	-0.437
	(0.255)*
Sometimes my effort at school is not given a proper reward	-0.445
	(0.191)**
Self-perception of Ability	1.285
	(0.416)***
Pocket Money	-0.003
	(0.011)
F-Statistic	8.579
R-squared	0.068
N	1,073

Each answer to the survey questionnaire except pocket money was coded from -1 (Strongly Disagree) to +1 (Strongly Agree). The grade 6 score is standardized to a mean of zero and a standard deviation of 1. The dataset contains two choices per student and one survey questionnaire answer per student. Thus, standard errors are clustered by student.

Appendix B: Experimental Procedure

Detailed Timeline

1. We determine randomly which experimenters are assigned to which classroom.
2. Students are assigned a random number.
3. Students sit at the table corresponding to the number.
4. The presenter introduces the experiment to students.
5. Students answer the example definition.
6. The presenter gives one among many possible answers for the example definition.

First round

7. Students choose how many questions they would like to buy, from no question to 10 questions.
8. Once this choice is made, Students can open the envelope and have 20 minutes to provide answers.
9. Envelopes are collected.

Second round

10. Students get a second envelope.
11. In the non anonymous group, the presenter states that questions will be corrected by their teacher. Students confirm that they know the teacher and are asked to confirm the subject that he/she teaches.
12. In the non anonymous group, students write down their name and their teacher's name on the envelope.
13. Students choose how many questions they would like to buy, from no question to 10 questions.

14. Once this choice is made, students can open the envelope and have 20 minutes to write down their answers.

15. Envelopes are collected.

16. Students can leave the classroom**.**

17. The teacher/the external marker grades the papers.

18. Payoffs are calculated and distributed in envelopes bearing the student's number.

Instructions

Today we would like to conduct an experiment with you, where you have the opportunity to win some real money. Over the next 90 minutes, we will ask you to complete two quizzes which are based on the definitions of words. The more questions you answer correctly, the more money you can win. I'd like you to relax, have fun and enjoy this experiment. I'm now going to explain what you have to do. If you have any questions, please ask me at the end.

In front of you will see an instruction sheet for the game. We will give you £2 to start with which you will use to take part in the quiz – you won't see the actual money until the end of the 90 minute round. Inside the envelope there are 10 questions where you will be asked to define different words. If you want to answer a question – you will have to pay 20p to play. If you get it right, you win 40p – so you double your money. If you get the question wrong, you don't win anything, but you forfeit the 20p you used to play. Remember – You're only playing with the money we give you.

Let's have a practice. On the instruction sheet, you'll see the practice question: 'Many people from pirates to archaeologists – have devoted their lives to a quest. What is an archaeologist?' Now I'll give you 2 minutes to have a go at this question – you must remain silent during these 2 minutes. Remember that this is just a practice – no money will be awarded.

2 minutes

A definition could be 'An archaeologist is somebody who studies past human societies such as the things they built and the environment they lived in. An archaeologist may excavate sites and recover evidence of past societies'.

Now the real experiment will involve 10 separate questions like this and you will have 20 minutes to answer them. At the end of the 20 minutes, the quiz will be marked anonymously – that means that your name will not appear on the quiz sheet, but you can be identified by the number on your desk. Please keep hold of this desk number as it identifies you so you collect your winnings.

There is no clear-cut definition of the word, so a range of answers could be accepted.

To recap, in order to play, you will have to pay 20p per question. If you decide to answer all 10 questions and you answer them all correctly, you could win £4 – doubling your money. If you get any questions wrong, you will forfeit the 20p you paid to answer it. You don't have to answer a question if you don't want to. You will receive your winnings at the end of the experiment.

Now I'd like you to turn over the envelope and fill in your desk number on the sheet and the name of your teacher. You'll see another example on the sheet about wrestling. At the bottom of the sheet, you must choose how many questions you would like to answer – you can choose from 0 to 10. Please choose now.

In a moment, I'll ask you to remove the quiz paper from the envelope and begin. You will have 20 minutes to answer the questions – and you must remain silent during these 20 minutes. I'll give you a warning when there are 5 minutes left.

You may now take out your quiz paper and begin. You have 20 minutes. Good luck.

20 minutes

Ask the young people to put the quiz back in the envelope and collect them all. Tell the young people that the quizzes will now be marked and the results will be given at the end when they will receive any winnings.

Hand out the envelopes again.

OK, now we are going to ask you to try a second quiz paper. Like last time, you will be given 10 questions on the definition of words and you will have 20 minutes to answer them.

I'm going to give you another £2 to play with – remember it costs 20p to answer each question and you can win 40p for every correct answer. Please turn over the envelope and fill in your desk number and your teacher's name. At the bottom of the sheet, you must choose how many questions you would like to answer – you can choose from 0 to 10. Please choose now.

- **Anonymous condition** Again your paper will be marked anonymously – You must remain

silent again from now on.

- **Non Anonymous condition** This time, your questions will be marked by your teacher. For this reason I will ask you to fill in your name instead of the desk number. You should also give the name of your teacher. Your teacher will be able to assess your answer using his/her knowledge of your vocabulary.

Please remove the quiz from inside the envelope. You have 20 minutes, I will warn you when there are 5 minutes remaining - you may begin. Good luck.

20 minutes

Ask the young people to put the quiz back in the envelope and collect them all. Tell the young people that the quizzes will now be marked and the results will be given at the end after the experiment.