Journal of Public Economics 105 (2013) 116–130

Contents lists available at SciVerse ScienceDirect

Journal of Public Economics

journal homepage: www.elsevier.com/locate/jpube

Students' perceptions of teacher biases: Experimental economics in schools[☆]

Amine Ouazad [a], Lionel Page [b,*]

[a] INSEAD, France
[b] Queensland University of Technology, Australia

ARTICLE INFO

Article history:
Received 11 November 2011
Received in revised form 7 May 2013
Accepted 12 May 2013
Available online 28 June 2013

Keywords:
Field experiment
Education
Discrimination
Teacher biases

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 1200 British students in grade 8 classrooms across 29 schools. In this design, students are given an endowment that they can invest on a task where payoff depends on performance. The task is a written verbal test which is graded nonanonymously 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 suggest systematic gender effects: students invest more with male teachers. Moreover, if we use the choices made with an external examiner as a benchmark, this result seems to come from two effects which complement each other: when comparing students' choices with their teacher to those made with an external examiner, we find that male students invest less with female teachers while female students invest more with male teachers.

1. Introduction

There is an extensive literature studying the determinants of educational achievement. There is, in particular, an interest in the factors that foster racial, ethnic, or gender gaps in education. Most studies focus on the effectiveness of educational inputs such as teacher quality (Rockoff, 2004; Hanushek and Rivkin, 2006), peer effects (Epple and Romano, 2010; Black et al., 2009), or school quality (Card and Krueger, 1992; Betts, 1995). Recently, considerable interest in the role of student effort in the classroom drives research at the frontier of the economics of education (Fryer, 2010; de Fraja et al., 2010; Bettinger, 2012).[1]

Interestingly, literature in psychology has emphasized that student effort responds to teachers' expectations and behavior. Research in psychology of education emphasizes that students' perceptions of their teachers' expectations plays a key role in their motivation (Maehr and Midgley, 1991; Meece et al., 2006). In particular, several studies have also found that a perceived good relationship with their teacher is correlated with a motivation to engage in school activities (Urdan and Schoenfelder, 2006). Students who perceive their teacher as supportive are more likely to invest higher level of effort (Wentzel, 1997) while the students who perceive teachers as harsh and cold have lower academic achievements (Wentzel et al., 2010).[2] Differential treatment by teachers can lead to perceptions of such differential treatment which then influences students' self-expectancies, and these self-expectancies cause future achievement (Jussim et al., 2009). In particular, such expectations can generate gender, racial, and ethnic inequalities when students perceive differential treatments along these dimensions. The study of Worrall et al. (1988) argues that (i) students perceived a differential treatment in favor of girls but not in favor of high achievers and (ii) teachers have better perceptions of girls and high-achievers. The study of McKown and Weinstein (2002) found that teachers are more likely to underestimate the ability of African American students and that those students tend to perceive this bias and, as a consequence, invest less effort at school.

[☆] We would like to thank two anonymous referees, the audience of the Education Group at the London School of Economics, the University of Texas in Austin, the Copenhagen Business School, the 2011 meeting of the Asia-Pacific Economic Science Association, the 2011 Econometrics Society Australasian Meeting, the 2011 Australia New Zealand Workshop on Experimental Economics (ANZWEE), the Labor and Public Policy Seminar series at the Paris School of Economics. 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.

* Corresponding author.
E-mail address: lionel.page@qut.edu.au (L. Page).
URL: http://staff.qut.edu.au/staff/page4/ (L. Page).

[1] Bettinger (2012) and Fryer (2010) have estimated whether financial incentives can boost student effort and achievement, particularly for low-achieving students.

[2] These psychological studies argue that the causality go from students' beliefs about their teacher to their motivation and thereby to their achievement.

0047-2727/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jpubeco.2013.05.002

Economics has a growing literature which, over the recent years, has found consistent teacher biases along the lines of gender, race, and ethnicity. On the gender side, accumulated evidence suggests that teacher assessments tend to favor girls. Lavy (2008) finds that in Israel, male students are systematically given lower grades in all fields when graded nonanonymously at the high-school matriculation exam and finds that these results are sensitive to the gender of the teacher. Dee (2007) found that teachers have better perceptions of students of their own gender, which, given the large share of female teachers in the classroom tends to disfavor boys.[3] On the racial and ethnic side, the literature has found that teachers rate minority students worse than other students. In Sweden, Hinnerich et al. (2011) estimated that ethnic minority students get lower grades in a field experiment.

However, while psychologists have hypothesized an effect of students' perceptions of teachers' differential treatment on effort, little experimental research exists on this question. In economics, the first major hurdle is to measure students' beliefs: questionnaires may ask about students' perceptions (in the United States, Wayman (2002)), but economists are typically skeptical about questionnaire answers (Bertrand and Mullainathan, 2001), skepticism which is stronger when looking at survey data on perceptions of racial or gender biases (Antecol and Kuhn, 2000; Antecol and Cobb-Clark, 2008). The second major hurdle is to estimate the causal impact of student beliefs on effort. Feldman and Theiss (1982) estimate the impact of students' perceptions of teachers by providing fictitious information about teachers.[4] However, the norm in economics is to avoid deceiving participants (Davis and Holt, 1993).[5] To our knowledge, there is little research that credibly identifies the causal impact of beliefs on student effort using an experimental design.

We design an experiment to estimate whether (i) students believe in teacher biases and whether (ii) these beliefs impact their effort level. 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 verbal test. A student who chooses to devote more money to the test has a longer test and hence gets more feedback from the teacher. 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. Grading practices are discretionary, as these exam questions do not have a well defined right or wrong answer. Classrooms are randomly assigned a *treatment* condition, where students know that they will be graded nonanonymously by the teacher, or a *control* condition, where students know that they will be graded anonymously by an external examiner. The external examiner is never in contact with the student, never appears nor are his gender or race mentioned. We then compare the amount of the endowment devoted to the test in the treatment and in the control classroom. Students and teachers are fully aware of the structure of the experiment, i.e. there is no deception involved (Davis and Holt, 1993). Standard economic theory[6] provides predictions regarding the optimal amount of the endowment devoted to the test. Differences in students' choices across the control and the treatment groups suggest differences in students' perceptions of teachers' grading practices. We implement this experiment with monetary incentives, across 29 English schools with 1200 grade 8 students. The experiment was carried out in controlled conditions – no interactions between students, large classroom, scripted experimental instructions – with students' usual teacher, close to the definition of an artefactual experiment (Levitt and List, 2009). Importantly, the set of students taking part in the experiment reflects the overall composition of the student population in England.

Results first suggest that money devoted to the test is not significantly different in the treatment and in the control conditions. This indicates that students do not have, on average, overly optimistic or overly pessimistic perceptions of the external grader compared to their teacher. Because of this result, differences in treatment effects by gender and ethnicity will be particularly informative.

The average effect indeed masks a strong gender effect: students tend to invest more with male teachers than with female teachers. On average, because students invest more with a male teacher they choose a test that is 9% longer with a male teacher. When breaking down the sample by student gender, and using the external examiner as a benchmark, male students invest less when graded by a female teacher — they choose a test that is 10% shorter, and female students invest more when graded by a male teacher — they choose a test that is 14% longer when assessed by a male teacher. In contrast, we found little impact of nonanonymous grading on nonwhite students' choices. Nonwhite students did not devote more or less money to the test when assessed nonanonymously by the teacher than when assessed anonymously by the teacher. There was no evidence of teachers grading differently nonwhite students than other students.[7]

This paper contributes to three separate literatures. First, the paper contributes to the literature on gender dynamics in the classroom. Standard economic theory applied to our results explains male students' behavior by their expectations that female teachers will have tougher grading practices. Indeed, in the experiment, female teachers gave worse grades to males. Hence, male students' choices are consistent with female teachers' grading practice. Therefore results are consistent with a mechanism in which differential treatment of male students leads them to exert lower effort, and thus students get less feedback from the teacher. These results could help explain male students' underperformance in English by their expectations of differential treatment. To our knowledge, this is a new result in the economics of education literature.[8] For female students, standard economic theory would explain their behavior by their expectations that male teachers give them higher grades. However, while on this point the evidence is only suggestive, we find that, if anything, male teachers give female students lower grades, in line with the literature that argues that teachers have better perceptions of same-gender students' performance (Dee, 2007). Female students' behavior could be explained either by a misperception of male teachers' grading practices toward them, or, alternatively they could react to a perceived discrimination by investing more.[9]

Second, the paper contributes to the literature on racial dynamics in the classroom. We found that nonwhite students did not have significantly better or worse perceptions of teachers' grading practices. In contrast, literature in education and psychology suggests that African American students may respond negatively to teachers' expectations (Ronald, 1998). This explanation stems in great part

[3] In England, Gibbons and Chevalier (2007) 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,

[4] In contrast, other studies in psychology – using observational data and multilevel regressions (McKown and Weinstein, 2002) – do not deceive students, but then the concern is that estimates are not causal.

[5] This is both for ethical reasons and to preserve the trust of participants in the experimental designs.

[6] The paper presents predictions and estimation using the subjective expected utility framework (Savage, 1954). Prospect theory (Kahneman and Tversky, 1979) would also lead to similar stylized facts, namely that the share of the endowment devoted to the test is increasing with the student's subjective probability of getting a test answer right.

[7] There is no significant variation in perceptions or grading for specific nonwhite ethnicities.

[8] Such a mechanism was hypothesized in psychology by Jussim et al. (2009).

[9] Female students could set an achievement goal which requires more effort when the teacher discriminates. Goal setting is an important behavioral mechanism in the classroom (Meece et al., 2006). In the behavioral economics literature, Camerer et al. (1997) shows that cab drivers set a goal for the day, and tend to work until that goal is reached.

from a reading of the stereotype threat literature (Steele and Aronson, 1995; Steele, 1997) whereby African American students reduce their effort because of their fear of confirming racial stereotypes. But experimental evidence is needed in the literature to establish whether in such a case ethnic minority students respond to teachers' differential treatment. We find no significant evidence of student perception of ethnic bias in grading, nor of actual teacher ethnic/racial bias in grading.

Finally, this experiment adds to the number of very recent economics of education studies using the methodology of laboratory experiments in the field (Harrison and List, 2004; Bettinger and Slonim, 2006; Bettinger and Slonim, 2007; Hoff and Pandey, 2006; Fryer, 2010). The number of field experiments is expanding particularly fast in the economics of education as classrooms provide a convenient setting where conditions can be controlled while preserving external validity. Recent work by Levitt et al. (2012) also expresses a willingness to more tightly integrate the economics of education with the economic theory of risk and uncertainty.

The paper is structured as follows. Section 2 presents the theoretical framework that allows for the elicitation of students' perceptions. Section 3 presents the experimental design and descriptive statistics on choices and payoffs. Section 4 estimates the effect of the nonanonymous condition on student choices, by teacher and student gender, and by ethnicity. We estimate students' subjective probability of success at the test using a structural expected utility model. The section also describes teachers' actual grading practices. Section 5 discusses the internal and external validity of the experiment, the role of the teacher's subject, and the importance of non-monetary incentives. Section 6 concludes.

2. Theoretical framework

2.1. Lottery choice and subjective perceptions of grading practices

We consider a student who is endowed with a sum of money S. The student chooses to purchase n questions between zero and N. Each question costs c. The student makes his choice and answers the questions. A grader marks the answers. Each question yields a payoff ω if the grader marks the answer as correct, and 0 if the grader marks the answer as incorrect. Let $k \in \{0,1,2, \ldots,n\}$ be the number of correct answers. The monetary payoff is payoff (k,n).

$$\text{payoff}(k,n) = (S - n \times c) + k \times \omega.$$

The uncertainty is on k, while S, n, c, ω are known. No question ($n = 0$) implies no variance in the outcome payoff$(0,n) = S$, and more questions implies a larger maximum payoff $S - n \times c + n \times \omega$ and a lower minimum payoff $S - n \times c$.

Hence, the student's optimal choice of n is a trade-off between risk and return. A number of standard economic frameworks can represent this choice, with similar implications. Here, for simplicity, we assume that the student maximizes the expected utility of the payoffs.[10] The expected utility of choosing n questions is noted $U(n)$.

$U(n)$ is the expectation of the utility of the payoffs $u(\text{payoff }(k,n))$ minus the costs $\delta(n)$ – cognitive and psychological costs – of choosing n questions.

$$U(n) = \sum_{k=0}^{n} p(k,n) \cdot u(\text{payoff}(k,n)) - \delta(n). \tag{1}$$

The student forms a subjective probability $p(k,n)$ of getting k right answers out of n.[11] His choice n maximizes his utility:

$$n^* = \text{argmax}_n U(n).$$

The subjective probability $p(k,n)$ of k right answers out of n ultimately depends on the student's subjective probability π of getting an answer right to any question out of the n questions chosen.

$$p(k,n) = \binom{k}{n} \pi^k (1-\pi)^{n-k}$$

π depends on student characteristics (e.g. confidence), grading conditions (anonymous or nonanonymous grading, the grader's characteristics such as gender, ethnicity, and age).

The number of questions chosen, n^* is a weakly increasing function of the subjective probability π. Indeed, as the subjective probability π increases, the probability of having a large number of answers right increases, and the probability of having a small number of answers right decreases. Formally, as π increases, the subjective probability of k answers right out of n, $p(k,n)$, increases for $k \geq n\pi$ and decreases for $k \leq n\pi$. An increase in π puts greater weight on the higher payoffs $u(\text{payoff}(k,n))$ for $k \geq n$ and smaller weight on the smaller payoffs $u(\text{payoff}(k,n))$ for $k \leq n$. Hence the optimal number of questions n^* weakly increases as π increases.

Overall, a student chooses a larger number of questions n^* than another student (i) because of a higher π, (ii) because of lower risk-aversion or lower non-monetary cost $\delta(\cdot)$. In turn, the subjective probability π depends on student self confidence and grading conditions.

2.2. Identifying differences in subjective perceptions using a randomized design

We identify the causal impact of grading conditions on choices n^* by randomly assigning students to one of two grading conditions.

Anonymous condition Grading is performed by an external examiner who does not see the student nor his/her name. The student does not get information on the grader's characteristics.

Nonanonymous condition Grading is performed nonanonymously by the teacher.

We write the characteristics of the student $X_{student}$ and the grading condition $Y_{condition}$, and the number of questions chosen $n^* = n^*(X_{student}, Y_{condition})$. The randomization of the assignment of students to each condition means that the distribution of student characteristics (which affect confidence, risk aversion, and non-monetary costs) are identical across the two conditions. The effect of the grading condition is simply estimated as the difference in the average number of questions chosen across the two conditions.

$$\begin{aligned}\text{Treatment} = {} & E(n^*(X_{student}, Y_{condition})|Y_{condition} = \{\text{Non anonymous condition}\})) \\ & -E(n^*(X_{student}, Y_{condition})|Y_{condition} = \{\text{Anonymous condition}\}).\end{aligned}$$

Because the number of questions chosen n^* is a weakly increasing function of π, a positive treatment effect implies a positive impact of nonanonymous grading by the teacher on the subjective probability π.

[10] Expected utility (von Neumann and Morgenstern, 1944), subjective expected utility theory (Savage, 1954), and prospect theory (Kahneman and Tversky, 1979) all yield similar predictions as the ones outlined in this paper. The crucial property of our experiment is that the number of questions chosen is increasing with the student's probability of getting a right answer.

[11] Specifically, the Subjective Expected Utility Theorem (Savage, 1954) states that the student's choice can be represented as an expected utility if the preference relation satisfies the continuity and extended independence axioms. Then, subjective probabilities $p(k,n)$ exist and $U(n)$ is written as $\sum_{k=0}^{n} p(k,n) \cdot u(r(k,n))$. Eq. (1) adds the psychological costs $\delta(n)$.

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

	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

In the anonymous condition, the student does not get information on the grader's characteristics. We assume therefore here that grader characteristics are independent of student characteristics and grading conditions, i.e. that $E(n^*(X_{student},Y_{condition})|Y_{condition} = \{\text{Anonymous grading,Grader Char.}\})$ is equal to $E(n^*(X_{student},Y_{condition})\ Y_{condition} = \{\text{Anonymous grading}\})$. The empirical validity of such an assumption is discussed in Section 5.1.[12]

Using the same dataset, the treatment effect Treatment (Male) of male students is estimated by taking the difference in the average number of questions chosen by male students in the nonanonymous condition and by male students in the anonymous condition. Because the assignment of students to grading conditions is random, the assignment of *male* students to grading conditions is also random. A positive treatment effect Treatment (Male) > 0 is indicative of male students' perceptions of more favorable grading practices by their teacher in the nonanonymous condition.

Finally, the treatment effect can be estimated when conditioning on a specific teacher characteristic. Conditional on a male teacher, the random assignment of students to conditions makes the distributions of student characteristics asymptotically identical across conditions. Hence the treatment effect Treatment (Male teacher) > 0 is indicative of a perception of more favorable grading practices by male teachers.

2.3. Interpretation of the experiment

The experimental design of this paper shares a number of features with students' behavior in the classroom outside of the experimental framework. de Fraja et al. (2010) and Bishop (2006) describe student behavior at school as determined by a trade-off between the return and the cost of effort. Literature has pointed out that the returns of students' effort are uncertain and risky,[13] and that these returns depend on teacher behavior in the classroom.[14] Jensen (2010) shows how perceived returns to education affect schooling decisions, and that these perceptions may be inaccurate.

Measuring the impact of student beliefs in a differential treatment on effort is however difficult with observational data because differences in effort – for instance measured as the number of hours of homework – are also indicative of differences in confidence, risk aversion, and cost of effort. Using a clearly defined set of monetary incentives with randomized conditions in the field (an "artefactual experiment" (List, 2006)) has the advantage of yielding potentially internally valid results in a relevant context. In particular, the random assignment of the two grading conditions provides an identification strategy that controls for the above-mentioned confounding factors. However the external validity of results obtained with such monetary incentives can be discussed (Davis and Holt, 1993), specifically the comparison between student effort and investment in questions (cf. Section 5).

3. Experimental design

3.1. Selection of schools

Around 1200 grade 8 students across 29 schools in London, Manchester and Liverpool took part in the experiment during the 2009–2010 academic year. 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.

[12] In particular students may have a prior probability that the external grader is male, female, white, or nonwhite; such prior probability may vary across schools or classrooms. We discuss whether such effects play a role empirically in our experimental results and suggest that they are unlikely to be significantly driving our results.

[13] For instance, students have imperfect knowledge of the returns to a college degree (Manski, 1993; Arcidiacono et al., 2010).

[14] There is an extensive literature on the impact of teacher quality on students' outcome (Hanushek and Rivkin, 2006).

The experiment takes the form of a 90-minute experiment that comprises two sessions and a questionnaire: The first session, where students know that they will be graded anonymously by the external examiner; the second session, where a random half of the students know that they will be graded nonanonymously 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.

3.2. Two sessions

The experiment is performed in two sessions. In the first session, students in both the control and the treatment classroom are in the anonymous condition. In the second session, the control classroom is in the anonymous condition, while the treatment classroom is in the nonanonymous condition.

	Control classroom	Treatment classroom
Session 1	Anonymous condition	Anonymous condition
Session 2	Anonymous condition	Nonanonymous condition

The 2-session design allows us to perform two placebo tests in the first session. First, if student assignment to the conditions is truly random, there should be no treatment effect in the first session. Second, if student assignment to the conditions is random and students' choices are not affected by the framing of the experiment, the treatment effect in the first session should not depend on the characteristics of the teacher in the second session in the nonanonymous condition.

3.3. The first session

Prior to the experiment, parents sign a parental agreement[15] 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. No deception is used in the experiment in regard to the teachers and students involvement. Given the sensitive nature of the object of study, i.e. students' perception of teachers' grading practices, only limited information is given on exact purpose of the experiment, in order to avoid *Hawthorne effects* (Mayo, 1949): where participants change their behavior when in an experimental setting, for instance as a response to experimenters' potential expectations.[16] The experiment is presented as a study of students' decision making processes.

Each school is visited by a team of 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 ("form teacher" in Britain) of the grade or the English teacher. This was checked before starting the experiment. 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 56) 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 graders 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' behavior when potentially 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 20 pounds each.

A right answer leads to a gain of 40 pounds, 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. 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:

$$r(n,k) = 2{-}0.20{\cdot}n + 0.40{\cdot}k.$$

Finally, students 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 min 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.

[15] Only one out of 1200 students' family refused to sign the agreement.

[16] We made it clear that this study was not about estimating particular teachers' biases. Indeed, when we later analyze teacher biases, the biases of any particular teacher are not estimated. Evidence on Hawthorne effects is largely debated, for instance in Levitt and List (2011)there. Finally, for Hawthorne effects to be at play here, students would have to expect that teachers' behavior would change due to the Hawthorne effect.

The words are taken from all subjects, from science, geography, history, and English.[17] Also, the design is such that both difficult and easy questions are present.[18] 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 session.

It is important to stress that no feedback is given at the end of the first session. Feedback on outcomes is only provided at the end of the second session, once students have left the classroom. Payoffs are handed at the end of the experiment for all students, regardless of their choices; there is indeed evidence that students exhibit high discount rates (Gruber, 2000), and handing out pay-offs at the end of the experiment avoids differences in choices due to impatience (Bettinger and Slonim, 2007).

3.4. The second session

Students are then told that there will be a second session, 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 min to fill in the answer sheet. Words of the second session are different than words of the first session.[19]

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 session.

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 nonanonymous setting. In the control classroom, we observe students' choice and outcome twice in the anonymous setting.

3.5. Complementary data: Survey questionnaire, administrative data and Teachers' and external examiners' grading

At the end of the second session, students fill a survey questionnaire after the second session, and before envelopes are handed, hence payoffs do not affect answers to the questionnaire.[20] 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.[21] 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.[22]

We merge the experimental results with administrative data on students, from the English National Pupil Database.[23] 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 session and in each condition, so that the final file includes the whole sequence of right and wrong answers.

3.6. Descriptive statistics

Students' choices are summarized in Table 3. Over the two sessions, 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 session and by 0.8 in the second session. The correlation between prior score and the number of questions is significantly stronger in the second session.

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. Table 3 shows the distribution of payoffs in the first and the second session.

4. Results

4.1. Estimation of students' perception of teachers' grading practices

We identify the effect of the nonanonymous condition on students' perceptions of teachers' grading practices by estimating the impact of the nonanonymous condition on students' investment choice (number of questions bought). The novelty here compared to Section 2.2 is that we use the 2 sessions of our experiment 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 session in the treatment and in the control group. A within-student estimate leads to more precise estimates than an estimate relying on one session of observation.

[17] 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).

[18] 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.

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

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

[21] 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.

[22] 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; 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.

[23] This database is central in most papers estimating school quality in England, see for instance Machin and McNally (2005a) and Kramarz et al. (2010).

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 in 1st session	6.46	6.33	0.453
	(0.12)	(0.12)	
	[597]	[557]	
Questions bought in 1st Session, School with Male Teacher in 2nd Session	6.37 (0.21) [225]	6.21 (0.20) [204]	0.564
Questions bought in 1st Session, School with Female Teacher in 2nd Session	6.30 (0.157) [225]	6.60 (0.146) [204]	0.160

Confidence intervals in parenthesis. Number of observations in brackets.

This amounts to estimating the following regression, where δ is the coefficient of interest:[24]

$$\begin{aligned}\text{Questions}_{i,t} = &\ constant + \alpha\cdot\text{Session } 2_{i,t} + \gamma\cdot\text{Treatment}_{i,t}\\ &+\delta\cdot\text{Session } 2\times\text{Treatment}_{i,t}\\ &+u_i+\varepsilon_{i,t}\end{aligned} \qquad (2)$$

Questions$_{i,t}$ is the number of questions bought by student i in session $t = 1, 2$. The coefficient of interest is δ, the effect of the nonanonymous condition on the number of questions bought. Because the treatment is randomly assigned (see Section 5.1 on page 24) 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 session, a difference which is partly due to students experiencing the first session and learning about the task. Interestingly, α also controls for learning when students of different characteristics learn differently.[25]

The average effect represents the difference between students' perception of their teachers and their perception of the anonymous graders. We used presenters with teaching experience, with a variety of ages. There is no a priori reason to think that students' perceptions of anonymous graders are neutral. While we did not give any information on the anonymous graders, students could have different beliefs about anonymous graders. Such beliefs could stem from the interpretation of the presenter's attitude or words. In particular, students could use presenters' characteristics to infer the likely characteristics of the graders.[26] In addition the anonymous grader only participates in a one shot interaction which could also affect students' perception of anonymous graders' behavior if students perceive that their teachers also use the grades to motivate them in repeated interactions.[27]

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

The average effect – i.e. the average difference in the number of questions purchased in the treatment and control classrooms – is a benchmark of students' reaction to the anonymous grader relative to their teachers. A small average effect indicates that students have reasonable perceptions of the anonymous grader. Hence, the average effect is not the primary effect of interest in this study. Of much more significant interest to us is the study of the existence of difference across subgroups. Under the assumption that different groups of students do not have systematically different beliefs of the external examiner's grading practices in the anonymous condition, different effects between groups of students reveal differences in their beliefs of their teacher's grading practices (please see Section 5.1 for further discussion).[28]

When estimating specification 2, the average effect of the nonanonymous condition is not significantly different from zero (0.035 with a standard error of 0.15), which suggests that students did not make wild assumptions on the behavior of the external grader compared to their teacher. More interestingly, the nonsignificant average effect masks considerable variation in the way students of different characteristics have responded to the nonanonymous condition. We estimate specification 2 on different subsamples defined by the teacher's and student's gender, the student's ethnicity and free meal status.[29]

4.2. Perceptions by student gender

Student gender has been shown to be one of the key variables affecting teachers' grading practices in previous literature (Dee, 2005, 2007). Our results, which focus first on students, 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 presents the coefficient δ of interest from a separate regression as in specification 2.[30] When graded by a male teacher, female students tended to buy 0.843 more question when graded by the

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

[25] 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 $\eta_{i,t} = \varepsilon_{i,t} + a_i$ Round $2_{i,t}$. Algebra shows that the randomization of the treatment ensures that Treatment$_{i,t}$ and Round 2 × Treatment$_{i,t}$ are independent of $\eta_{i,t}$. Hence the treatment effect δ is consistently estimated by OLS.

[26] For this reason we independently randomized both the allocation of the presenters and the graders.

[27] For instance, students may expect that the teacher uses prior classroom behavior in addition to the actual answers when grading.

[28] Appendix B presents empirical evidence that students' perceptions of the anonymous grader are not significantly correlated with the gender of the student or of the anonymous grader.

[29] Estimation on subsamples and estimation on the entire sample with dummies for the subgroups does not yield significantly different estimates. Appendix C presents similar estimates when pooling observations and interacting the treatment dummy with students' and teachers' characteristics.

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

Table 4
Main result — Effect of nonanonymous grading by the teacher by teacher and student gender.

	Teachers			
Students	All	Male	Female	Δ = Male − Female
All	0.036 (0.150)	0.576 (0.233)**	−0.318 (0.197)	0.894 (0.297)**
Observations	2292	856	1396	2292
Male	−0.086 (0.232)	0.487 (0.312)	−0.601 (0.268)**	1.088 (0.446)**
Observations	1031	486	801	1031
Female	0.359 (0.230)	0.843 (0.371)**	0.110 (0.268)	0.733 (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 questions. Being graded by a male teacher increases the number of questions bought by 0.576 questions.
**: Significant at 5%. *: Significant at 10%.
This table reports the effect of the nonanonymous condition for each group of students and each group of teachers. Coefficients of the first five rows are the coefficients of separate regressions $questions_{i,t} = \alpha \text{Session } 2_{i,t} + \delta \text{Session } 2 \times \text{Treatment}_{i,t} + u_i + \varepsilon_{i,t}$.

Table 5
Effect by ethnicity and by free meal eligibility.

Students	Treatment effect
White	0.097 (0.178)
Observations	1614
Nonwhite	−0.100 (0.284)
Observations	678
Eligible for free meals	0.238 (0.390)
Observations	290
Noneligible for free meals	0.007 (0.163)
Observations	2002

Each coefficient comes from a separate regression for the treatment effect on each subsample.
No significant coefficient in the table.
This table reports the effect of the nonanonymous condition for each group of students and each group of teachers. Coefficients are the coefficient δ of regression $questions_{i,t} = \alpha \text{Session } 2_{i,t} + \delta \text{Session } 2 \times \text{Treatment}_{i,t} + u_i + \varepsilon_{i,t}$.

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 students was slightly higher than the number of male students, the average student graded by a male teacher bought significantly more questions in the nonanonymous condition than in the anonymous condition (+0.576).

4.3. 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 education 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 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 5 estimates result for free meal and nonfree meal students. As for Table 4, each cell is the effect of the nonanonymous condition for a separate regression. Results suggest that there is no effect of poverty status on the number of questions bought.

Table 5 also displays the same analysis for White and for non-White students. As mentioned in the introduction, the stereotype threat literature (Steele and 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). Our results do not suggest such effects of ethnicity on students' choices. There is no effect regardless of whether we consider the whole non-White category or whether we consider a breakdown of non-White students by racial subgroup.[31] 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.

4.4. Estimating students' subjective probabilities of success

Previous analyses found an effect of the nonanonymous condition on the number of questions bought using a regression that made no particular assumption on the particular utility function that drives student choices. The theoretical framework for the experiment (Section 2) does not rely specifically on an expected utility framework for instance. But choosing a particular utility function – and estimating it – allows us to translate a difference in the number of questions chosen into differences in subjective probabilities of getting an answer right. This matters because if students react strongly (large change in the number of questions chosen) to a small change in subjective beliefs, even small differences in the perceptions of male and female teachers could trigger large changes in student behavior.

Hence, we estimate a structural model of choice where students choose the number of question which maximizes their utility, in an expected utility framework. Doing so we are able to convert the treatment effects of Table 4 into differences in subjective probability of success with their respective teachers. We assume a random utility model where the utility of choosing n questions is:

$$U_n = E[u(c(n,k))] + \varepsilon_n \tag{3}$$

where $k \le n$ 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 $P(k|n) = \binom{n}{k} \hat{p}^k (1-\hat{p})^{n-k}$.[32]

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 $0 \le \hat{p} \le 1$ of a right

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

[32] This model does not include the 20 min duration of the session as a time constraint in the students' decisions. This modeling choice is supported by our observations (see Section 3.3) which suggests that students did not need more than 20 min to fill the answer sheet.

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 session, and whether the observation is for treatment classroom in the second session. That gives a specification for $\hat{p}$ which is similar to the baseline specification of Eq. (2). There is a different $\hat{p}$ for each session and for the control and treatment classrooms.

$$\hat{p}_{i,t} = a + b \cdot \text{Session 2} + c \cdot \text{Treatment}_{i,t} + d \cdot \text{Session 2} \times \text{Treatment}_{i,t}. \tag{4}$$

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 et al. (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.

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

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

Results are presented in Table 6. 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 and Johnson, 1990), the fact of playing 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 session.

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 session 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 nonanonymous 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 nonanonymous treatment on students' subjective beliefs in their chances 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.

4.5. 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 nonanonymous condition

Table 6
Estimation of the expected utility model.

	Whole sample			Male teacher	Male teacher	Female teacher	Female teacher	White	Nonwhite	Free meal	Nonfree meal
Dependent variable: p											
Constant	0.626 (0.006) ***	0.629 (0.007) ***	0.627 (0.011) ***	0.638 (0.018) ***	0.637 (0.018) ***	0.616 (0.015) ***	0.606 (0.014) ***	0.650 (0.016) ***	0.641 (0.019) ***	0.595 (0.040) ***	0.656 (0.014) ***
Treatment			0.014 (0.019)	−0.020 (0.031)	−0.019 (0.031)	0.036 (0.023)	0.033 (0.022)	0.008 (0.026)	−0.021 (0.037)	0.093 (0.049) *	−0.020 (0.023)
Session 2			−0.012 (0.010)	−0.040 (0.020) **	−0.062 (0.040)	0.006 (0.012)	−0.026 (0.015)	−0.013 (0.016)	−0.006 (0.019)	−0.029 (0.024)	−0.003 (0.014)
Treatment × Session 2			0.001 (0.016)	0.061 (0.027) **	−0.074 (0.076)	−0.035 (0.021) *	0.017 (0.025)	0.036 (0.022)	0.008 (0.032)	0.020 (0.042)	0.024 (0.020)
Treatment × Session 2 × Female					0.100 (0.049) **						
Treatment × Session 2 × Male							−0.165 (0.053) ***				
Dependent variable: r											
Constant	−0.574 (0.015) ***	−0.520 (0.056) ***	−0.565 (0.019) ***	−0.544 (0.034) ***	−0.546 (0.034) ***	−0.583 (0.033) ***	−0.640 (0.032) ***	−0.639 (0.024) ***	−0.611 (0.055) ***	−0.604 (0.054) ***	−0.653 (0.026) ***
Male		−0.094 (0.088)									
Number of observations	2292	2292	2292	856	856	1396	1396	946	469	290	1083

Reading: Students graded nonanonymously 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 nonanonymously 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 nonanonymous 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%.

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, starting with the first question; which substantially alleviates the previous issue of the endogenous selection of questions. 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 7 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 nonanonymous condition dummy, the prior grade 6 score, and interactions between the nonanonymous condition and the prior score, and between the nonanonymous condition and the teacher's gender.

$$\begin{aligned}\text{Question } k \text{ Right}_{i,round\ 2} = {} & \text{constant} + a\cdot\text{Male}_i + b\cdot\text{Non Anonymous Condition}_i \\ & +c\cdot\text{Grade 6 Score}_i \\ & +d\cdot\text{Non Anonymous Condition}_i \times \text{Grade 6 Score}_i \\ & +f\cdot\text{Non Anonymous Condition} \times \text{Male}_i \\ & +g\cdot\text{Non Anonymous Condition} \times \text{Male}_i \times \text{Female Teacher}_i \\ & +g\cdot\text{Non Anonymous Condition} \times \text{Female}_i \\ & \times\text{Male Teacher}_i + \varepsilon_i \end{aligned} \quad (5)$$

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 8 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 nonanonymous 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 nonanonymous condition on some questions, a finding consistent with Lavy (2008), who finds that male students tend to get lower grades when graded nonanonymously.

Table 7
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. $p_{\text{External Examiner}}$ 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 nonanonymous groups and in the anonymous groups.

Results suggest that ethnicity did not play a significant role in teachers' grading. Column (4) regresses the probability of marking the answer as correct on gender, ethnicity dummies, as well as on a non-anonymous condition × non-White dummy. The effect of the latter is nonsignificant. Results of Columns 9 and 15 suggest no significant impact for the other questions either. Hence neither non-White students' perceptions of grading practices nor the actual grading practices are significantly different for nonwhite students.

The results of the impact of ability on grading suggest that, while teachers graded higher ability students more favorably, more able students were not more likely to get a positive outcome in the nonanonymous condition. 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 nonanonymous condition, revealing that teachers grade students of different ability levels the same way as the external examiner. This suggests that the teacher is not using his knowledge of the student's prior achievement in the classroom to grade the answers.

If one interprets the results presented in Section 4 as reflecting students' perceptions of teachers' grading, male students' choices are consistent with female teachers' grading practices. 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 *nonanonymous condition*). Female students' choices, on the other hand, are hard to rationalize with teachers' actual grading practices. Our results suggest that female students' choices would be consistent with male teachers giving them higher grades.

5. Discussion

5.1. Internal validity

A possible concern for our results is whether randomization was successful. In spite of our random allocation of the treatment and presenters by coin tosses, one could wonder whether we have successfully eliminated systematic differences in students and presenters characteristics between treatment and 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 are 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 session, when all students are in the nonanonymous condition. 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 session* 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 nonanonymous condition. This indicates that the different effects observed across teachers from different genders may not come from systematic differences in the characteristics of their students.[33]

More fundamentally, our interpretation of the results by teacher gender relies on the assumption that students have neutral beliefs about the characteristics of the external examiner characteristics (in

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

Table 8
Comparing grading practices — The teacher vs external markers.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
Variables	Customary	Customary	Customary	Customary	Customary	Single	Single	Single	Single	Single	Rhyme	Rhyme	Rhyme	Rhyme	Rhyme
Male	0.108**	0.108**	0.108**	0.022	0.048	0.060	0.060	0.060	−0.013	0.011	0.098	0.098	0.098	0.031	0.048
	(0.047)	(0.047)	(0.047)	(0.034)	(0.034)	(0.053)	(0.053)	(0.053)	(0.037)	(0.037)	(0.073)	(0.074)	(0.074)	(0.052)	(0.053)
Nonwhite				0.086*	0.086*				0.019	0.009				−0.007	−0.000
				(0.052)	(0.051)				(0.058)	(0.059)				(0.079)	(0.078)
Nonanonymous condition	0.072	0.053	0.095	0.071*	0.008	0.175**	0.140*	0.171**	0.165***	0.131**	0.202*	0.223**	0.250**	0.072	0.126
	(0.063)	(0.063)	(0.067)	(0.041)	(0.054)	(0.072)	(0.073)	(0.077)	(0.044)	(0.063)	(0.104)	(0.103)	(0.106)	(0.062)	(0.093)
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)		
Nonanonymous × Nonwhite				−0.048	−0.034				−0.058	−0.044				0.090	0.078
				(0.075)	(0.074)				(0.082)	(0.081)				(0.115)	(0.117)
2nd quartile of prior score	−0.060	−0.060	−0.060		−0.064	0.039	0.039	0.039		0.044	0.220**	0.220**	0.220**		0.225**
	(0.062)	(0.062)	(0.062)		(0.062)	(0.073)	(0.073)	(0.073)		(0.073)	(0.101)	(0.101)	(0.102)		(0.102)
3rd quartile of prior score	0.119*	0.119*	0.119*		0.100	0.194**	0.194**	0.194**		0.192**	0.194*	0.194*	0.194*		0.193*
	(0.068)	(0.068)	(0.068)		(0.068)	(0.077)	(0.077)	(0.077)		(0.077)	(0.102)	(0.102)	(0.102)		(0.102)
4th quartile of prior score	0.220***	0.220***	0.220***		0.207***	0.198***	0.198***	0.198***		0.191***	0.217**	0.217**	0.217**		0.212**
	(0.065)	(0.065)	(0.065)		(0.066)	(0.070)	(0.070)	(0.070)		(0.070)	(0.095)	(0.095)	(0.095)		(0.094)
Nonanonymous × 2nd Quartile of prior score	0.153*	0.138	0.127		0.154*	0.088	0.060	0.053		0.089	−0.225	−0.204	−0.201		−0.240
	(0.090)	(0.090)	(0.090)		(0.090)	(0.101)	(0.101)	(0.101)		(0.101)	(0.147)	(0.147)	(0.147)		(0.148)
Nonanonymous × 3rd Quartile of prior score	−0.024	−0.046	−0.049		0.005	0.019	−0.024	−0.025		0.030	−0.067	−0.029	−0.011		−0.057
	(0.096)	(0.097)	(0.097)		(0.096)	(0.104)	(0.105)	(0.105)		(0.103)	(0.146)	(0.151)	(0.149)		(0.147)
Nonanonymous × 4th Quartile of prior score	0.108	0.091	0.081		0.133	−0.012	−0.042	−0.047		0.001	0.009	0.027	0.032		0.017
	(0.092)	(0.092)	(0.092)		(0.092)	(0.098)	(0.098)	(0.098)		(0.098)	(0.132)	(0.133)	(0.132)		(0.131)
Observations	846	846	846	847	847	702	702	702	702	702	353	353	353	353	353
R-squared	0.054	0.057	0.062	0.007	0.054	0.058	0.069	0.072	0.023	0.056	0.056	0.062	0.065	0.012	0.055

Robust standard errors in parentheses.

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

*** $p < 0.01$.

** $p < 0.05$.

* $p < 0.1$.

particular gender), or at least that students do not have systematically different beliefs about the external examiner.[34] A limitation of our experiment is that we do not observe the students' perceptions of (or priors on) the external examiner. We therefore do not know how students' beliefs vary across students and across schools.[35] First, one possibility would be for instance that students are more likely to believe that the external examiner is a female if their teacher is female and conversely that the examiner is a male if their teacher is male. To test for such a possibility, we checked whether students' choices differ systematically in the first round (when they are marked anonymously by the external grader) between schools with male teachers and schools with female teachers.[36] We do not find any significant difference between these two situations ($p = 0.61$). Second, following on the same principle, we estimated the main regression on the whole sample, with an interaction dummy between the teacher gender and the treatment effect to check whether our effects are only driven by differences in students' perception of the external grader across classrooms with male teachers and classrooms with female teachers.[37] Doing so does not affect the results about the interaction between teachers and students gender. The corresponding coefficients keep the same magnitude and are both significant at 5%. Third, perceptions of the external grader may differ across schools, thus leading to a heterogeneous effect of male teachers across schools.[38] To check for this possible heterogeneity, we estimated a separate treatment effect of male teachers for each school by interacting the treatment effect with each school effect, as described in the appendix. Although there is some degree of treatment heterogeneity, the positive effect of male teachers is present for all schools of the sample.

5.2. External validity

Our experimental design is a response to the quandary that experimenters face between external and internal validity. A lab experiment has the advantage of providing clear monetary incentives that model students' typical trade-off between return and cost of effort.[39] But our results will shed light on actual students' perceptions and the impact of such perceptions on their effort and achievement at school if the results can be generalized to other settings than this particular task.

We tried to mitigate such external validity concerns as much as possible by (i) selecting participants from the specific population of interest: secondary students in England, (ii) conducting the experiment in a usual and relevant setting for them and for our study, their school; finally (iii) by facing participants – the students – with a task based on a type of vocabulary test whose words are taken from their curriculum exams, as words are taken from the previous ten years of the exam at the end of secondary school.[40]

Although our experiment is an artefactual experiment (List, 2006) – a lab experiment taking place in the field – the latter two characteristics (ii) & (iii) make the experiment closer to field experiments than to pure lab experiments. However, as an artefactual experiment, the design differs from real world situations in critical ways.

First, we are not observing actual student effort,[41] such as homework, or paying attention during classes. Our experiment sets up a situation where investment in a task is costly while the return to this investment is risky. In such a situation students face a trade-off between higher returns and the higher associated risk. Our motivation to opt for such a setting is that it reflects the underlying trade-off faced by students when they have to choose to expend effort now for a future and uncertain reward in the form of exam marks, college entry and labor market prospects. However it is not necessarily the case that students would react exactly the same way in a real effort task as in a task with monetary costs and rewards. Further research looking at effort tasks could therefore provide additional insights about how students are likely to react to the perceptions of lower chances of success due to teacher biases.

Second, we are not able to reproduce the time frame of the trade-off between costly investments and future rewards faced by students.[42] In the lab or in artefactual experiments – as in this experiment – rewards tend to be given soon after investment choices. In particular in our experiment, rewards are given at the end of the experiment for all students. Thus, although our results do not depend on students' impatience, students are not experiencing a long time period between investment and reward. On the contrary, students have to invest their time and effort now for rewards obtained days or weeks later such as exam outcomes, or even years later when considering labor market opportunities. This time dimension clearly adds a layer of complexity to students' choices that is not present in our experiment. Outside the experiment, students could dismiss present teachers biases if they believe that later evaluation (at the university or on the labor market) will be free from such biases.[43]

In addition to these issues, the Appendix addresses additional concerns by providing additional empirical results. First, one may wonder whether our sample of teachers was representative and therefore whether our results are a fair representation of what we could have found in other English schools. Second, one may wonder whether our teacher gender results may not be due to the larger representation of English and humanities teachers among female teachers. Third, one could wonder whether students guessed the anonymous marker's gender — even if our instructions and our experimenters never revealed such characteristic. On each of these concerns, the Appendix provides evidence that these are unlikely to have affected our results.

5.3. Monetary versus non-monetary incentives

Relating the experimental results to students' perceptions of teachers' grading practices requires us to believe that monetary payoffs are credible determinants of students' choices. One could wonder whether non-monetary incentives play a role in students' choices. A student may want to please or impress the presenter (Levitt and List, 2007), please the teacher relatively more than the presenter, signal his/her ability (Feltovitch and Harbaugh, 2002), signal hard work or conform to group norms when graded by the teacher (Austen-Smith

[34] The mechanism can be integrated in the formal model (Section 2.2) by noting the student's subjective probability that the grader is male γ_{male}. When in the anonymous condition, $0 \leq \gamma_{male} \leq 1$. The difference between the anonymous and the nonanonymous conditions is twofold: (i) in the nonanonymous condition, the student is certain of the grader's gender, either $\gamma_{male} = 1$ when the teacher is male, or $\gamma_{male} = 0$ when the teacher is female, and, importantly (ii) the student knows that his name (& thus gender) is revealed to the grader.

[35] Although asking students about their perceptions of the external grader would raise a similar set of issues as for questions related to their perceptions of the teacher; such issues are common to survey questionnaires, e.g. the social desirability bias (Bertrand and Mullainathan, 2001).

[36] In the notations of Footnote 34, the prior γ_{male} could be different in schools where the teacher was male, and in schools where the teacher was female.

[37] A regression on the whole sample constrains the prior γ_{male} to be equal across schools, while regressions on subsamples allows the prior γ_{male} to vary across subsamples.

[38] In the formalization of Footnote 1, γ_{male} would differ across schools.

[39] This is not unlike Voors et al. (2011), which carries out public good experiments in 35 villages in Sierra Leone.

[40] Words are taken from the previous ten years of the Key Stage 3 reading booklets, as described in Section 3.1.

[41] On issues surrounding the measurement of student effort, see for instance de Fraja et al. (2010).

[42] On student impatience, see for instance Bettinger and Slonim (2007). We mitigated the impact of differing student discount factors, for instance across genders, by giving monetary rewards at the same time for all students of a given school.

[43] Although students seem to be motivated by teacher praise & feedback in the short-run (see for instance in psychology Ames (1992)), as long-run monetary returns may be hard to assess at this age.

and Fryer, 2005). Several elements indicate that such non-monetary incentives are unlikely to be driving the results.

First, the experiment' provides substantial monetary incentives for 13 year old students. Students can earn up to £8, which represents 1.25 times students' average weekly pocket money (around £6).[44] From our personal experience and the feedback we received from students, the prospect to win monetary rewards was a key motivator for students and it prompted them to think carefully about the best option to maximize their payoffs.

Second, non-monetary incentives would not necessarily bias results. If the desire to please the presenter or teacher varies across students but not across anonymous or nonanonymous conditions, random assignment to treatment and control, together with the within-student estimation, captures this confounding factor. In other words, non-monetary incentives are averaged out in the difference in differences estimation. Non-monetary incentives could naturally be stronger in the second session when students are marked by their teachers, but even in this situation, non-monetary incentives bias our results only if the desire to please the experimenter is systematically different across subsamples.

Third, we use the answers from the post experiment survey to check whether non-monetary incentives seem to have a significant influence on students' choices. The survey includes a question about the desire of the student to value the relationship with the teacher independent of the monetary incentives of the experiment: “A good relationship with the teacher matters (Strongly Disagree… to Strongly Agree).” We focused on the sample of female students in schools where the teacher was male, and on the sample of male students in schools where the teacher was female. 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). If non-monetary incentives play a strong role in students choice we could expect to see different treatment effects depending on the answer to these questions. In practice, the treatment effect for those two subgroups does not significantly differ (Table 9). This is most visible for female students for whom treatment effects are significant and positive for each subgroup. This suggests that desires to please the teacher may not drive female students' behavior when assessed by a male teacher. For male students, treatment effects are negative in each subgroup and point estimates are close, but only significant for the subgroup signaling a higher desire to value the relationship with the teacher. This could possibly indicate that non-monetary incentives may explain some dispersion in the treatment effect for male students assessed by female teachers. However the difference between the two subgroups is in itself nowhere near significant. Overall, these results suggest that the pattern of behavior we observed does not seem strongly driven by students' willingness to please their teachers.

5.4. Stereotype threats vs students' perceptions of teacher biases

An important strand of the psychology literature (Steele and Aronson, 1995; Steele, 1997; Aronson et al., 1998) describes the phenomenon of stereotype threats whereby students' performance at a test is lower when primed with a gender or ethnic stereotype — e.g. a stereotype that male students underperform in English. Because, in England, male students have on average lower test scores than female students in English (Machin and McNally, 2005), male students' fear of confirming the gender stereotype associated with boys may impair their test performance. Although the initial stereotype threat literature was concerned with effects on performance, stereotype threats may also affect students' *expectations* of test performance.

Table 9
Treatment effect for female students graded by a male teacher — By answer to the survey questionnaire.

	Treatment effect of male teachers for female students	Treatment effect of female teachers for male students
Good relationship with the teacher matters		
More than the median student	0.801	−0.594
	(0.442)*	(0.303)**
Less than the median student	0.892	−0.612
	(0.467)*	(0.383)
The advice and help of my teacher have played an important role in my progress		
More than the median student	0.849	−0.628
	(0.432)**	(0.325)*
Less than the median student	0.834	−0.558
	(0.482)*	(0.393)
Number of observations	278	801

**: Significant at 5%. *: Significant at 10%.

Stereotype threats could be an interesting alternative interpretation of our findings if male students' fear of confirming the gender stereotype when assessed by a female teacher specifically leads to lower investment. This paper's experiment is indeed about students' expectations of performance rather than about performance per se. For stereotype threat to affect the number of questions chosen, male students would need to anticipate ex-ante that their test performance will be lower with a female teacher — perhaps because the teacher has different grading practices, but also because of expectations of stereotype threats.

Such an effect may be at play in our results. Indeed, while the first wave of papers on stereotype threats (Steele and Aronson, 1995; Steele, 1997; Aronson et al., 1998) initially focused on the impact of stereotypes on student performance, later papers focused on the impact of stereotype threats on students' expectations of performance (Spencer et al., 1999; Stone et al., 1999; Steele et al., 2002). Spencer et al. (1999) found that the stereotype threat manipulation did not affect women's performance expectations. Stone et al. (1999) found that, while stereotype manipulations affected the athletic performance of White and Black athletes, it did not affect performance expectations. From this evidence, in a literature review, Steele et al. (2002) conclude that there is no clear evidence of the role of stereotype threats in performance expectations.

6. Conclusion

Using a deception-free incentive-compatible experimental design in 29 English schools with 1200 students, we estimated the effect of students' perceptions of teacher biases on student investment in the classroom. Our results do not suggest that students from low-income families and minority ethnic backgrounds 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 shows that in some contexts teachers give better grades to students of their own gender (Dee, 2007). We find that students behave in ways which suggest that they may perceive biases as a function of their gender and of 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 can be explained by male students having 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

[44] Halifax Pocket Money Survey 2008.

by a male teacher. Interestingly, an analysis of teachers' grading practices suggests that these beliefs only partially match teachers' actual behavior. Male students' choices are in line with the fact that male teachers give them lower grades, but male teachers do not seem to favor female students. As we use the external examiner as a neutral benchmark, a limitation of our results is that we are not able to observe students' beliefs about the external examiner's characteristics, in particular the examiner's gender. For this reason, some caution is required when interpreting the results. The observed pattern of behavior reflects students' beliefs about teachers' biases only if students do not systematically differ in their beliefs of the examiner's characteristics.

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

These results are therefore interesting in light of the growing gender gap in education which has become a concern for policy makers (Weaver-Hightower, 2003). Further research is required to 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.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.jpubeco.2013.05.002.

References

Ames, C., 1992. Classrooms: goals, structures, and student motivation. Journal of Educational Psychology 84 (3), 261.

Andersen, S., Fountain, J., Harrison, G.W., Rutström, E.E., 2010. Estimating subjective probabilities. 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.

Arcidiacono, P., Hotz, V.J., Kang, S., 2010. Modeling college major choices using elicited measures of expectations and counterfactuals.

Aronson, J., Lustina, M.J., Good, C., Keough, K., 1998. When white men can't do math, necessary and sufficient factors in stereotype threat. Journal of Experimental Social Psychology 1–18.

Austen-Smith, D., Fryer, R.G., 2005. An economic analysis of “acting white”. Quarterly Journal of Economics 551–583.

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

Bettinger, E., 2012. Paying to learn: the effect of financial incentives on elementary school test scores. The Review of Economics and Statistics, MIT Press 94 (3), 686–698.

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.

Betts, J.R., 1995. Does school quality matter? Evidence from the national longitudinal survey of youth. The Review of Economics and Statistics 77 (2), 231–250.

Bishop, J., 2006. Drinking from the fountain of knowledge: student incentive to study and learn — externalities, information problems and peer pressure. Handbook of the Economics of Education, vol. 2 1–36.

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.

Camerer, C., Babcock, L., Loewenstein, G., Thaler, R., 1997. Labor supply of New York City cabdrivers: one day at a time. Quarterly Journal of Economics 112 (2), 407–441.

Card, D., Krueger, A.B., 1992. Does school quality matter? Returns to education and the characteristics of public schools in the United States. Journal of Political Economy 100 (1), 1–40.

Davis, D.D., Holt, C.A., 1993. Experimental Economics.

de Fraja, G., Oliveira, T., Zanchi, L., 2010. Must try harder: evaluating the role of effort in educational attainment. The Review of Economics and Statistics 92 (3), 577–597.

Dee, T., 2005. A teacher like me: does race, ethnicity, or gender matter? American Economic Review 95 (2), 158–165.

Dee, T., 2007. Teachers and the gender gaps in student achievement. Journal of Human Resources 42 (3), 528–554.

Epple, D., Romano, R.E., 2010. Peer effects in education: a survey of the theory and evidence. Handbook of Social Economics. 1–186.

Feldman, R., Theiss, A., 1982. The teacher and student as pygmalions: joint effects of teacher and student expectations. Journal of Educational Psychology 74 (2), 217.

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

Fryer, R., 2010. Financial incentives and student achievement: evidence from randomized trials. NBER Working Paper Series.

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

Gruber, J., 2000. Risky behavior among youths: an economic analysis. Technical Report. National Bureau of Economic Research.

Hanushek, E.A., Rivkin, S.G., 2006. Teacher quality. Handbook of the Economics of Education, vol. 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. 1–36.

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

Jensen, R., 2010. The (perceived) returns to education and the demand for schooling. Quarterly Journal of Economics 125 (2), 515–548.

Jussim, L., Robustelli, S., Cain, T., 2009. Teacher expectations and self-fulfilling prophecies. Handbook of motivation at school. 349–380.

Kahneman, D., Tversky, A., 1979. Prospect theory: an analysis of decision under risk. Econometrica: Journal of the Econometric Society 263–291.

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

Lavy, V., 2008. Do gender stereotypes reduce girls' or boys' human capital outcomes? Evidence from a natural experiment. Journal of Public Economics 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.

Levitt, S., List, J., 2011. Was there really a Hawthorne effect at the Hawthorne plant? An analysis of the original illumination experiments. American Economic Journal: Applied Economics 3 (1), 224–238.

Levitt, S.D., List, J.A., Neckermann, S., Sadoff, S., 2012. The behavioralist goes to school: leveraging behavioral economics to improve educational performance. NBER Working Papers 18165. National Bureau of Economic Research, Inc.

List, J.A., 2006. Field experiments: a bridge between lab and naturally occurring data. Advances in Economic Analysis & Policy 6 (2).

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

Maehr, M., Midgley, C., 1991. Enhancing student motivation: a schoolwide approach. Educational Psychologist 26 (3–4), 399–427.

Manski, C., 1993. “Adolescent Econometricians:” How do Youth Infer the Returns to Schooling? Studies of Supply and Demand in Higher Education, National Bureau of Economic Research.

Mayo, E., 1949. Hawthorne and the Western Electric Company, The Social Problems of an Industrial Civilisation. Routledge.

McKown, C., Weinstein, R., 2002. Modeling the role of child ethnicity and gender in children's differential response to teacher expectations1. Journal of Applied Social Psychology 32 (1), 159–184.

Meece, J., Anderman, E., Anderman, L., 2006. Classroom goal structure, student motivation, and academic achievement. Annual Review of Psychology 57, 487–503.

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

Ronald, E., 1998. Can schools narrow the black-white test score gap? The black-white test score gap. 318.

Savage, L.J., 1954. The Foundations of Statistics. Wiley, New York.

Spencer, S., Steele, C., Quinn, D., 1999. Stereotype threat and women's math performance. Journal of Experimental Social Psychology 35, 4–28.

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 1–15.

Steele, C., Spencer, S., Aronson, J., 2002. Contending with group image: the psychology of stereotype and social identity threat. Advances in Experimental Social Psychology 34, 379–440.

Stone, J., Lynch, C., Sjomeling, M., Darley, J., 1999. Stereotype threat effects on black and white athletic performance. Journal of Personality and Social Psychology 77 (6), 1213.

Tajfel, H., 1982. Social psychology of intergroup relations. 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.

Urdan, T., Schoenfelder, E., 2006. Classroom effects on student motivation: goal structures, social relationships, and competence beliefs. Journal of School Psychology 44 (5), 331–349.

von Neumann, J., Morgenstern, O., 1944. Theory of Games and Economic Behavior. Princeton University Press.

Voors, M., Bulte, E., Kontoleon, A., List, J.A., Turley, T., 2011. Using artefactual field experiments to learn about the incentives for sustainable forest use in developing economies. American Economic Review 101 (3), 329–333.

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.

Wentzel, K., 1997. Student motivation in middle school: the role of perceived pedagogical caring. Journal of Educational Psychology 89 (3), 411.

Wentzel, K., Battle, A., Russell, S., Looney, L., 2010. Social supports from teachers and peers as predictors of academic and social motivation. Contemporary Educational Psychology 35 (3), 193–202.

Worrall, N., Worrall, C., Meldrum, C., 1988. Children's reciprocations of teacher evaluations. British Journal of Educational Psychology 58 (1), 78–88.