Cheap Talk in the Classroom: How biased grading at school explains gender differences in achievements, career choices, and wages Revised Version

Lydia Mechtenberg[0]
Technische Universität Berlin
l.mechtenberg@ww.tu-berlin.de

November 6, 2008

Abstract

In this paper, I provide a theoretical explanation for the gender differences in education and on the labor market that are observed empirically in most OECD countries, including the U.S. Within a cheap talk model of grading, I show that biased grading in schools results in (1) boys outperforming girls in math and sciences, (2) boys having more top and more bottom achievers in math and sciences than girls, (3) girls outperforming boys in reading literacy, (4) female graduates enrolling in university studies more often than male graduates, (5) the predominance of female students in arts and humanities at the university, (6) the predominance of male students in math and sciences at the university, and (7) the gender wage gap on the labor market for the highly educated.

Keywords: Cheap talk, Education, Discrimination, Gender, Wage gap.

JEL Classification: D82, I21, J16.

1 Introduction

Persisting gender differences in education and on the labor market constitute a three-fold puzzle to economists: Achievements in school, enrollments at university, and wages show gender-specific patterns that are very similar across different OECD countries. Concerning achievements in school, boys and girls differ both in mathematical and non-mathematical subjects. With regard to mathematical subjects, nearly all existing data on cognitive achievement of school children reveal that boys outperform girls in math and sciences.[1] Besides, significantly more males than females exceed the magnitude of the highest proficiency level in mathematics.[2] In about half of the OECD countries, including the U.S., boys perform also more frequently at the lowest proficiency level in the math and science tests of TIMSS and PISA. At moderate proficiency levels, females are more strongly represented.[3] With regard to reading abilities, by contrast, PISA data prove that girls are on average better and have more top achievers and less bottom achievers than boys in almost all OECD countries.[4] With regard to university enrollments, much more females than males enroll in university studies. But while female students are predominant in arts and humanities, male students occupy the

[0] I thank Helmut Bester, Annette Boom, Paul Heidhues, Kai A. Konrad, Daniel Krähmer, Dorothea Kübler, Johannes Münster, Kerstin Puschke, Roland Strausz, Elmar Wolfstetter, the audience at the ESWC 2005, two editors and two anonymous referees for valuable comments on earlier versions of this paper. The usual caveat applies. This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk". I also thank the Econometric Society for a travel grant.

[1] See the TIMSS 2000 Report and the PISA-USA 2003 Report. In almost all OECD countries, average PISA scale scores in mathematics are higher for males than for females. See, for instance, the PISA 2003 Report, Figure 2.18.

[2] In the U.S., for example, 2.8 percent of 15-years-old males and only 1.2 percent of 15-years-old females perform at Level 6 in mathematics, the highest possible proficiency level in PISA 2003.

[3] For the comparison of achievement distributions for males and females, compare Table 2.5b in the PISA 2000 and 2003 Reports.

[4] In the U.S., 9.3 percent of 15-years-old males but only 3.7 percent of 15-years old females perform below proficiency Level 1 in the reading-literacy test of PISA 2000. But 11 percent of 15-years-old males and 13.4 percent of 15-years old females perform at the highest level, Level 5. For the achievement distributions in reading (by gender), compare Table 5.2a in the PISA 2000 report. The PISA 2003 report claims that results did not change much from 2000 to 2003. The gender differences in mean achievement in reading are reported both for 2000 and 2003 in the PISA 2003 report, Figure 6.6.

math and science fields.[5] Finally, there is a significant wage gap between men and women on the labor market.[6]

While the economic literature has provided some possible theoretical explanations for the gender wage gap, there are so far no attempts to explain theoretically the gender differences in education. More importantly, to my knowledge there is no unified model providing a single explanation for all gender differences mentioned above. The current paper aims at closing this gap. I provide an explanation of gender differences in achievements at TIMSS and PISA tests, gender differences in choices of whether and what to study, and gender differences in earnings. The explanation I suggest locates distortions at school.

I analyze a symmetric cheap talk model of teachers and students. In this model, teachers in humanities and math get signals about the talent of their students. Then, they send messages (grades) to their students, either transmitting their signals or lying. Then, students choose the effort they would invest into the subject at university. I assume that these efforts are measured by the PISA and TIMSS data. Afterwards, the students choose whether or not to go to university. When they decide to go to university, they choose the field - math or humanities - they want to enroll in. They invest their corresponding effort. At the end of university, their human capital, the product of talent and effort, becomes publicly observable. They enter the labor market and earn a wage that is determined by their human capital.

The perfect Bayesian equilibria of this model are characterized by the equilibrium grading behaviour of teachers and the corresponding equilibrium beliefs that the students have about the meaning of the grades they get. Ignoring babbling equilibria, the existence of equilibria without full information transmission is due to the assumption that some teachers are biased. This assumption is justified in section 2 of this paper in which I discuss empirical studies on grading behaviour of teachers and other related literature. From this literature, one can draw two inferences. First, biased teachers often use grades as messages about their liking or disliking the student's attitude. But second, biased teachers are somewhat scrupulous; they do not want to bias their grades too much. Therefore, I assume that teachers do not want

[5] In the U.S., for example, 81 percent of all tertiary degrees rewarded in type-B programs of humanities, arts, and education are allocated to females. But females get only 31 percent of all such degrees rewarded in mathematics and computer sciences. Compare Table A3.4, OECD, www.oecd.org/edu/eag2006.

[6] See Weichselbaumer and Winter-Ebmer (2005) for a meta-study.

to distort the beliefs that their students have about their own talents too much. They give biased grades if their students do not take these grades too seriously; but if they anticipate a student to internalize a grade fully as message about his (her) talent, they react with using the grade as an unbiased message about talent only.

In the current paper, I concentrate on the asymmetric equilibrium that is characterized by the gender differences observable in schools, at universities, and on the labor market. I call this equilibrium the "U.S. equilibrium", because the U.S. is one of the most important OECD countries in which the gender differences described above are observed. The asymmetry of the U.S. equilibrium is due to asymmetric equilibrium beliefs of boys and girls about the meaning of the grades they get. These asymmetric equilibrium beliefs lead to asymmetric equilibrium grading behaviour of teachers toward boys and girls, or vice versa.

More specifically, in humanities, girls do not fully internalize bad grades as bad messages about their talent in humanities. But they believe good grades to be good news about their talent. Boys, by contrast, have the reverse beliefs. Teachers in humanities anticipate that bad grades for a girl are not fully internalized as messages about the girl's talent in humanities. They also anticipate that good grades for a boy are not fully internalized as messages about the boy's talent in humanities. Thus, biased teachers in humanities start to misuse these grades as messages about their liking or disliking the student. This, in turn, justifies the students in their unwillingness to fully internalize these grades. As a result, in humanities, lowly talented girls partly discard their bad grades and become overconfident, exhibiting higher achievement than lowly talented boys. Highly talented boys partly discard their good grades and become underconfident, achieving less than highly talented girls. The resulting gender difference in achievement distributions in humanities is (in qualitative terms) exactly the one that can be observed from the PISA data on reading literacy.

In math, the equilibrium situation is different because all students internalize bad grades as bad messages about their talent in math. This is intuitive because in math, it is difficult for teachers to distort their grades downwards without being detected to do so. Grades in math can only be biased easily when they include oral contributions; and it is much easier for teachers to bias grades upwards by referring to oral contributions, than downwards. In the U.S. equilibrium, girls do not fully internalize good grades as good messages about their talent in math. However, boys do. Thus, biased teachers who anticipate these beliefs of their students often use good grades to express their liking of a female student, thereby justifying the reluctance of girls to internalize good grades in math. The teachers' patronizing of girls in math has two consequences. On the one hand, more lowly talented girls than lowly talented boys receive praise in math; and as these girls *partially*, though not fully, internalize the praise, they achieve more on average than the lowly talented boys. But on the other hand, the highly talented girls in math mistrust the praise they get and achieve therefore less on average than the highly talented boys. In sum, boys outperform girls in math. Again, the resulting gender difference in achievement distributions in math is (in qualitative terms) exactly the one that can be observed from the PISA data.

The driving force of these effects is a loss of information due to potential biases in grades. Grades are signals that students receive about their own talent, and potentially biased grades are are noisier signals than unbiased grades. In the equilibrium that will be analyzed in the current paper, the loss of information is generally larger for girls than for boys.

The above-mentioned self-assessments of boys and girls which are produced at school can explain the gender difference in university enrollment decisions and in earnings, too. To grasp the intuition of this result, consider first the group of students with good grades both in humanities and math. The girls in this group fully trust their praise in humanities, but not their praise in math. Thus, they decide to enroll in humanities. The boys, however, exhibit the reverse beliefs. They therefore choose to enroll in math. Next, consider the group of students with bad grades in both subject fields. Because boys believe both the bad grade in humanities and the bad grade in math to reflect low abilities, they decide not to go to university at all. Girls, however, partly attribute the bad grade in humanities to dislike of the teacher. Thus, they enroll in humanities *in spite of* their mediocre success in this subject at school. As a result, females enroll more often at university than males. Second, they choose humanities instead of math more often. Third, female university students are less talented on average than male university students, because too many females who are lowly talented in humanities enroll at university. Because this gender difference in average talents of university students is also reflected in wages, the gender wage gap can be explained without assuming any distortions on the labor market.

At this point, a remark on the rationale of analyzing asymmetric equilibria in a symmetric model is in order. In general, an asymmetric equilibrium in a symmetric model should be viewed as an equilibrium that persists even when the (unmodeled) asymmetry that originally has triggered it vanishes.[7] There are reasons to believe that the persisting gender differences observable in reality should be modeled in this way. Beliefs about talents of males and females and about the right way to educate boys and girls have changed considerably in all industrialized countries; but gender differences in schools, at universities and on the labor market have not vanished yet.

The remaining part of this paper is organized as follows. Section II gives a short overview of related literature. Section III presents the model. In section IV, equilibrium results are derived and discussed. Section V concludes.

2 Related literature

The current paper bridges the gap between two separate branches of the literature, namely the literature on cheap talk games and the literature on discrimination. The stream of literature on cheap talk games that started with Crawford and Sobel (1982) contains a small but important sidearm of papers on the manipulation of beliefs about one's own self. Benabou and Tirole (2000) and (2003) have applied a cheap talk game to the interaction between an agent and a principal who is more informed about the agent's abilities than the agent himself.[8] The authors consider situations in which the principal has an incentive to manipulate the agent's self-confidence either upwards or downwards. The message space of the principal allows him to evoke, according to his preferences, either underconfidence or overconfidence of the agent. For example, the principal can choose between transmitting information about the agent's abilities or not sending any message to the agent at all. The agent cannot observe whether the principal has received any information. Then, by transmitting only bad news to the agent and keeping

[7] One way to think about the unmodelled asymmetry that might have triggered the U.S. equilibrium originally is to include parents into the model. Suppose that originally, parents had different priors about the talents of their sons and daughters. These (irrational) priors influenced the way in which they interpreted the grades their children got at school. Suppose that girls and boys originally accepted the way in which their parents interpreted grades and therefore ended up with exactly the asymmetric beliefs that are part of the U.S. equilibrium. Then, the U.S. equilibrium will persist even if parents stop having irrational priors about the talents of their children or if children stop listening to their parents' interpretation of grades.

[8] Compare also Fang and Moscarini (2005). In their 2002 paper, Benabou and Tirole analyze the rational management of one's own self-confidence.

silent about good news, the principal can make the agent underconfident.

The core intuition behind the current paper is that mechanisms like those depicted in Benabou and Tirole (2000) and (2003) are the reason of gender differences in education and on the labor market. Nevertheless, the model in the current paper diverges in important respects from Benabou and Tirole (2000) and (2003). First, I consider not only one cheap talk game but two simultaneous and interrelated cheap talk games. The message that a teacher (principal 1) transmits to a given student eventually also depends on the signal that the teacher of the other subject field (principal 2) receives. This is because a signal received by a teacher does not only, as in Benabou and Tirole (2000) and (2003), reflect the agent's (i.e., the student's) abilities. It also encodes information about the student's personality that is liked or disliked by the teachers.

Second, the message space of the principal is more standard in my model. The teacher cannot choose to send no message at all; instead, he is able to lie. Third, I consider not only one type, but several types of principals, i.e. teachers. Thus, my paper is related to the literature on cheap talk with two types of senders. As in Benabou and Laroque (1992), Morris (2001), and Sobel (1985), I consider a cheap talk game where the receiver of the message (the student, in my model) does not observe the type of the sender. Fourth, I also consider several types of receivers, namely male and female lowly and highly talented students. This setting allows for underconfidence and overconfidence to occur simultaneously at the two different points of the talent distribution, resulting in achievement distributions as observed from the data.

My paper is also related to the literature on discrimination. The basic ideas of both taste discrimination and statistical disrimination are combined. Taste discrimination has been defined for the labor market by Gary Becker in his 1957 book. According to Becker, if "good behaviour" toward a given group of agents costs a principal more than this same good behaviour toward another group of agents, then the principal is likely to discriminate against the first group whenever the two groups are distinguishable. In my model, teachers sometimes have higher utility from dishonest grading than from honest grading; and the amount of this difference in utility depends on the personality of the students receiving the grades. Thus, my model incorporates the basic element of taste discrimination.

Statistical discrimination, by contrast, originates from prior beliefs instead of preferences.[9] Consider, for example, a principal who believes women to be less likely to be productive than men. Consequently, if the agent's payoff depends on the principal's belief, a women has to invest more effort than a man to earn the same payoff. Put differently, her marginal benefit of effort is often lower, compared with a man. Thus, women end up being less productive than men on average. As it is easy to see, statistical discrimination is a self-fulfilling prophecy.[10]

In my model, girls and boys internalize the same kind of feedback differently because they have different expectations about how honest the teacher behaves to them. As a result, a teacher who wants to use grades as messages about his liking or disliking the student's personality anticipates different reactions from boys and girls. This, in turn, induces the teacher to behave in a discriminatory way that fulfils exactly the expectations of boys and girls. This kind of self-fulfilling prophecy is close to the mechanism underlying statistical discrimination.

[9] The literature on statistical discrimination started with Arrow (1972), Phelps (1972), and Arrow (1973). See also Coate and Loury (1993).

[10] Evidently, the assumption of discrimination, regardless of its being taste discrimination or statistical discrimination, necessitates research on the effects of affirmative action. Coate and Loury (1993) discuss affirmative action with regard to statistical discrimination. They prove that the effect need not be positive. Defining affirmative action as the commitment to equal representation of the advantaged and the disadvantaged group in the labor market, they show the following: Although there is a "good" equilibrium in which affirmative action leads to hamogeneous beliefs about the two groups, there is also a "bad" equilibrium. In the bad equilibrium, negative beliefs about the disadvantaged group, combined with affirmative action, lead to patronizing of this group. The patronized group faces easier test standards than the other group. Consequently, the disadvanteged group invests less effort than the other group and thereby confirms the negative beliefs held about them. A similar effect is demonstrated in Fryer, Loury, and Yuret (2003). In their paper, they show that color-blind affirmative action (defined as a flattening of the function that relates worker productivity to the probability of being employed) weakens the incentives for all kinds of workers to invest effort. An argument in favor of affirmative action has been put forward by Benoit (1999). Benoit assumes that the disadvanteged group of workers scores lower in assesment tests than the other group although both groups are equally productive. In such a situation, affirmative action could be an efficient (temporary) solution. In my paper, I abstract from affirmative action, since affirmative action has not yet become as prevalent in the education system as in the labor market, especially not with regard to grading. However, I shortly discuss the potential effects of affirmative action with regard to grading in the Conclusion of this paper. There, I argue that such a kind of affirmative action would lead to biased grading and would produce exactly the gender differences that it would be aimed against.

Since the way in which teachers influence the beliefs of their students is at the heart of the model, I relate my assumptions about teacher preferences closely to empirical studies of grading behaviour. Researchers on grading practices broadly agree that a large percentage of teachers consciously use grades not only to inform their students about their talents or achievements, but also to signal to them how their "attitude" or "motivation" is evaluated. For example, Bursuck et al. (1996) report on page 308 that approximately "*50% of all teachers [use] certain specific grading adaptions for their students (...) including basing grades on improvement, giving multiple grades (e.g., grades for tests and effort), and making individual adjustments to grading weights (e.g., counting projects more than tests for some students).*" Non-achievement factors included into grades are pure effort, improvement, and compliance or attitude of the students.[11] Friedman and Manley (1992) find that according to teachers, 72 percent of a final grade should reflect the pupil's achievement. Thus, teachers give biased grades, but do not want to bias their grades too much.

Jussim (1989) and Jussim and Harber (2005) show empirically that self-fulfilling prophecies and biased grading do indeed occur in the classroom. In his 1989 paper, Jussim analyses longitudinal data obtained from 27 teachers and 429 6th-graders in math classes. He uses a path analyses, controlling for past achievements, grades and motivation of students in order to separate the effects that teachers' perceptions and grades have on the students' achievements from the reverse effects. With regard to biased grading, he reports the following result: "In comparison to students whom teachers believed to be lazy, those whom teachers believed to try hard received higher grades, but not higher standardized test scores. (...) Perhaps teachers intentionally used grades as a way of rewarding hard-working students or as a way of punishing lazy students." (p. 473) Moreover, Jussim (1989) finds that the correlation between actual effort of the students and the effort that teachers perceive them to invest is very weak and sometimes even negative. Thus, teachers' perceptions are biased. With regard to self-fulfilling prophecies due to biased grading, Jussim (1989) reports: "Teacher perceptions of performance had no direct effects on achievement, but they did have indirect effects mediated by students' self-concept of ability and teacher perceptions of effort and tal-

[11] Compare Cross and Frary (1996), Hills (1991), Stiggins et al. (1989), and Zeidner (1992). Measurement experts agree, though, that grades should be based solely on achievements. See for example Waltman and Frisbie (1994).

ent (...)." (p. 475) Thus, the results "were consistent with the occurence of modest-sized self-fulfilling prophecies and perceptual biases (...)." (p. 475) Jussim and Harber (2005) give an overview over these and similar results from other studies.[12]

As Howley, Kusimo, and Parrot (2000) claim, the grading practices of teachers reported above make a class or race bias of grading more probable. This, of course, is also true with regard to a gender bias.[13] Yet, this topic is still not widely explored; and there is so far not enough evidence on the relation between the students' grades and gender. What is known, though, is that both boys and girls receive confounded grades that reflect both their abilities and their perceived attitude. My model accounts for this.

3 The Model

There are male students of mass one and female students of mass one. Both the group of male students and the group of female students are independently and identically distributed on a unit square. Each student i has talent θ_{ij} in the subject field j, whith $j \in \{H,\ M\}$. The location of student i on the respective unit square is given by $\left(x_i^H, x_i^M\right)$. When $x_i^H \leq \frac{1}{2}$, the student is lowly talented in humanities, $\theta_{iH} = k$; and when $x_i^H \geq \frac{1}{2}$, the student is highly talented in humanities, $\theta_{iH} = 1$. Correspondingly, when $x_i^M \leq \frac{1}{2}$, the student is lowly talented in math, $\theta_{iM} = k$; and when $x_i^M \geq \frac{1}{2}$, the student is highly talented in math, $\theta_{iM} = 1$. Students do not know their talents; both boys and girls have the priors $\Pr\left\{\theta_H = 1\right\} = \Pr\left\{\theta_M = 1\right\} = \frac{1}{2}$. Besides, each student i has either a pleasing or displeasing personality, $a_i = 1$ or $a_i = 0$. The two types of personality have equal probability and cannot

[12] For another general description of self-fulfilling prophecies in the classroom, see Trouilloud et al. (2002), p. 591. Many psychological studies prove that teacher expectations are strongly correlated with the effort choice of students. See for example Alvidrez and Weinstein (1999), Clifford and Walster (1973), Clifton et al. (1986), Hoge and Butcher (1984), the seminal paper of Rosenthal and Jacobson (1968) and, for an economic paper, Lavy (2004).

[13] Most interestingly, Dee (2005) proves that perceptual biases of teachers occur with increasing probability when the teacher's gender and / or race differs from that of the student. For a similar study, compare Ouazad (2008). However, Ouazad does not find effects of gender differences between teachers and students. Thus, I abstract from this point since the overall empirical evidence on effects of gender differences between teachers and their students is still not strong enough.

be observed directly; also students themselves do not know if their own personality is pleasing or not. Hereafter, the two types of personality are called "attitudes". From the perspective of teachers, a pupil's good attitude might in general consist in a bundle of traits that teachers happen to believe to be most important for a good "learning climate", such as manner and cleanliness in one school, and curiosity and liveliness in another school. Attitudes and talents are independent.[14]

Each student is taught both in humanities and math. In each field, he or she is matched with probability α with a biased, i.e. potentially dishonest, teacher.[15] (For simplicity, I assume a uniform α for both fields. But this does not affect the main results.) The teacher r gets an imperfect signal $s_{rij} \in \{1, k\}$ on the talent of his student i in field j, with expected signal quality $\sigma \in \left(\frac{1}{2}, 1\right)$.

Then, the grades for student i in the fields j and j' are determined in the faculty room. The grading continuation game in the faculty room has the following structure. First, both teachers, r and r', are forced to reveal publicly the signals s_{rij} and $s_{r'ij'}$ they have got. They are not allowed to lie in the faculty room. Then, one of them is drawn randomly to be the first mover in grading. Without loss of generality, let teacher r (i.e. the teacher of field j) be the first mover. He decides on the grade $m_{rij} \in \{1, k\}$ he wants to give. Afterwards, the other teacher, r', decides on his grade $m_{r'ij'} \in \{1, k\}$. Both are allowed to opt for messages that are not in line with their signals.

This is the simplest possible structure of the grading continuation game between teachers that achieves the equilibrium results reported in section 4. Two assumptions contribute to simplifying the analysis. First, the assumption that teachers cannot lie in the faculty room, although they can do so in the classroom, guarantees that the teachers' updating of their beliefs about their pupil's attitude be straightforward. Second, the sequential structure of

[14] Independence is assumed for tractability reasons. Generally, the fundamental assumption is that attitude and talent are not perfectly correlated, i.e. that there is a possibility of biased grading when teachers react to attitudes. However, the assumption of (imperfect) correlation between attitude and talent would complicate the updating of beliefs in the model too much, without contributing much to the general idea.

[15] The model would be robust to the assumption that each pupil is taught by one teacher in both fields who is potentially dishonest in field j with probability α, with $j \in \{M, H\}$ and independent probabilities. However, the assumption of two teachers, one per field, is more natural in my view, since potential dishonesty seems to be a trait of character rather than a subject-dependent property. I thank an anonymous referee for directing my attention toward this point.

the game between teachers leads to a unique pure-stragey equilibrium of the continuation game starting with a given signal-vector $s = (s_j, s_{j'})$. Apart from simplifying the analysis, both assumptions are also intuitive. First, teachers want to learn more about their pupils' attitudes, so they appreciate honesty in the faculty room, and it is easier for a teacher than for a pupil to find out that some fellow teacher has misreported his signal. A teacher, for example, could check whether his collegue empoyed equal grading standards across all his pupils' written tests while a pupil, by contrast, could not do that. Thus, the social norm of honesty could be enforced in the faculty room, but not in the classroom. Second, the assumption of a sequential structure of the game is also natural, since communication is sequential by nature.

The continuation game that starts with the end of the faculty meeting proceeds as follows: Back in the classroom, teachers send their messages $m_{rij}, m_{r'ij'} \in \{1, k\}$ to their student i. Students date up their beliefs about their talents according to Bayes rule, taking into account the probability that the teacher was not honest about his signal. Then, students decide between a university program in math (field M), a university program in humanities (field H), and the outside option, i.e. entering the labor market directly after school and earning wage w_0.

A university student i invests effort e_{ij} into the chosen program, with effort costs e_{ij}^2. Then, his or her talent in the chosen field j, θ_{ij}, becomes publicly observable. At the end of his or her university career, student i has human capital $\theta_{ij}e_{ij}$. At the labor market, he or she has full bargaining power and earns $w_{ij} = \theta_{ij}e_i$. Then, the game ends.

The equilibrium concept which applies to the game is that of perfect Bayesian equilibrium: Given their beliefs, teachers and students must make optimal decisions at all information sets; and they must update their beliefs according to Bayes rule whenever that is defined.

3.1 Preferences of students

The expected utility of student i at the end of school is[16]

$$\max\left\{U_H\left(m\right), U_M\left(m\right), w_0\right\}$$

[16] I drop subscripts when they do not contribute to clarity.

with

$$m = (m_M, m_H),$$

$$\begin{aligned} U_H(m) &= \pi_H(m)\left(e_H - e_H^2\right) + (1 - \pi_H(m))\left(ke_H - e_H^2\right), \text{ and} \\ U_M(m) &= \pi_M(m)\left(e_M - e_M^2\right) + (1 - \pi_M(m))\left(ke_M - e_M^2\right) \end{aligned} \tag{1}$$

where $\pi_H(m)$ (or $\pi_M(m)$) represents the student's posterior subjective probability of being highly talented in humanities (or math), given the messages m. Accordingly, the student's optimal effort choice with respect to field j is given by

$$e_j^* = e_j(m) = \frac{1}{2}\left[\pi_j(m)(1-k) + k\right]. \tag{2}$$

From (1) and (2), one can easily see that the expected utility of choosing field j for university studies increases monotonously in the student's confidence in his abilities in field j:

$$U_j\left(m, e_j^*\right) = \frac{1}{4}\left[\pi_j(m)(1-k) + k\right]^2 \tag{3}$$

I assume that w_0, the wage of non-academics, lies marginally above the expected utility of studying a field j in which one has received a fully credible bad message about one's talent, so that $\pi_j(m) = \pi_j(k) = (1-\sigma)$. Thus,

$$w_0 = \frac{1}{4}\left[1 - \sigma(1-k)\right]^2 + \varepsilon, \text{ with } \varepsilon > 0. \tag{4}$$

This assumption guarantees that a fully credible bad grade in field j prevents students from studying this field at the university. Nevertheless, a bad grade that is not fully credible (but might be due to the teacher's dislike of the student's attitude) does not necessarily close the door to university studies in the respective field. Thus, students' decisions between studying humanities, studying math, and entering the labor market are fully determined by their beliefs about their respective abilities, as reported in

Lemma 1 (1) When $\pi_j(m) = (1-\sigma)\ \forall j \in \{H, M\}$, the student enters the labor market directly after school and earns w_0. (2) Otherwise, the student decides to study field j at the university if $\pi_j(m) > \pi_{j'}(m)$,

$j \neq j'$, $j, j' \in \{H, M\}$, and to study field j' if $\pi_j(m) < \pi_{j'}(m)$. (When $\pi_j(m) = \pi_{j'}(m)$, the student is assumed to toss a coin.)

Proof Lemma 1 follows directly from (3) and (4). □

3.2 The signal technology and types of teachers

Consider a signal s_{rij} received by teacher r about the talent θ_{ij} of student i in field j. I assume that the quality of this signal depends on the true attitude a_i of the student. Consider first the case when the student's attitude is such as teachers want it to be, i.e. $a_i = 1$. Then, if the student is highly talented in field j, the teacher r will be very likely to detect this:

$$\Pr\{s_{rij} = 1 \mid a_i = 1, \theta_{ij} = 1\} = \sigma_{rH} > \sigma.$$

But if the student is lowly talented in field j, the teacher r will be less likely to find this out:

$$\Pr\{s_{rij} = k \mid a_i = 1, \theta_{ij} = k\} = \sigma_{rL} < \sigma,$$

with

$$\frac{1}{2}(\sigma_{rH} + \sigma_{rL}) = \sigma, \text{ and } \sigma_{rL}, \sigma_{rH} \in \{k, 1\}.$$

Thus, if the student i has a convenient attitude, a teacher's perception of this student's talent will be biased toward a positive judgment.

Consider now the case when the student's attitude is rather inconvenient, i.e. $a_i = 0$. Then, if the student is highly talented, the teacher r will not be very likely to find this out:

$$\Pr\{s_{rij} = 1 \mid a_i = 0, \theta_{ij} = 1\} = \sigma_{rL}.$$

But if the student is lowly talented, the teacher will be quite likely to detect this:

$$\Pr\{s_{rij} = k \mid a_i = 0, \theta_{ij} = k\} = \sigma_{rH}.$$

Thus, if the student i has an inconvenient attitude, a teacher's perception of the student's abilities will be biased toward a negative judgment.

Consequently, the signal s_{rij} received by the teacher r is not only a signal about the student's talent θ_{ij}, but also a signal about his (or her) attitude

a_i. More specifically, a positive signal $s_{rij} = 1$ about the student's talent is also a positive signal about his (or her) attitude; and a negative signal $s_{rij} = k$ about the student's talent is also a negative signal about his (or her) attitude.

The teachers' perception of their students' talents can be *weakly* or *strongly attitude-sensitive*, $\tau_r \in \{\tau_w, \tau_s\}$. A fraction $\alpha \in (0,1)$ of teachers is weakly attitude-sensitive ($\tau_r = \tau_w$), i.e. their signal quality depends only weakly on the student's attitude:

$$\sigma_{rH} = \sigma_H \text{ and } \sigma_{rL} = \sigma_L \text{ if and only if } \tau_r = \tau_w.$$

The remaining fraction $(1-\alpha)$, however, is strongly attitude-sensitive ($\tau_r = \tau_s$), i.e. their signal quality depends strongly on the student's attitude:

$$\sigma_{rH} = \sigma_H + \Delta \text{ and } \sigma_{rL} = \sigma_L - \Delta, \text{ with } \Delta > 0, \text{ if and only if } \tau_r = \tau_s.$$

A given teacher himself knows whether he is weakly or strongly attitude-sensitive. But his type cannot be observed directly by others.

Consider now the first move within the grading continuation game in the faculty room. Both the teacher r in field j and the teacher r' in field j' of student i disclose their respective signals. Consequently, both teachers have *two* pieces of information about the student's attitude a_i, namely s_{rij} and $s_{r'ij'}$. Consider first the case where r is strongly attitude-sensitive, i.e. $\tau_r = \tau_s$. Applying Bayes Rule, it is easy to see that his posterior belief about a_i is

$$\begin{aligned}\Pr_r \{a_i = 1 \mid s_{rij}, s_{r'ij'}\} &> \frac{1}{2} \text{ iff } s_{rij} = 1, \text{ and} \\ \Pr_r \{a_i = 0 \mid s_{rij}, s_{r'ij'}\} &< \frac{1}{2} \text{ iff } s_{rij} = k, \ \forall s_{r'ij'} \in \{k, 1\}.\end{aligned}$$

Intuitively, the signal that a strongly attitude-sensitive teacher received is, from his own perspective, more informative about the student's attitude than the signal of his collegue. This is because the strongly attitude-sensitive teacher does not know whether or not his collegue is strongly attitude-sensitive, too. Consequently, the posterior belief of a strongly attitude-sensitive teacher r about a student's attitude moves into the same direction as his posterior belief about this student's talent, regardless of the signal that his collegue r' received.

Consider now the case where r is weakly attitude-sensitive. It is straightforward to show that his posterior belief about a_i is

$$\begin{aligned}\Pr_r\{a_i = 1 \mid s_{rij}, s_{r'ij'}\} &> \frac{1}{2} \text{ iff } s_{r'ij'} = 1, \text{ and} \\ \Pr_r\{a_i = 0 \mid s_{rij}, s_{r'ij'}\} &< \frac{1}{2} \text{ iff } s_{r'ij'} = k, \ \forall s_{rij} \in \{k, 1\}.\end{aligned}$$

Thus, from the perspective of a weakly attitude-sensitive teacher r, the signal that his collegue r' received is always more informative about the student's attitude a_i than the signal that he himself received. Thus, the weakly attitude-sensitive teacher follows his collegue in his belief about the pupil's attitude. This is because a weakly attitude-sensitive teacher must take into account that his collegue might be strongly attitude-sensitive. As a result, the posterior belief of a weakly attitude-sensitive teacher r about a student's attitude runs contrary to his posterior belief about this student's talent if and only if $s_{rij} \neq s_{r'ij'}$. Put differently, his belief about the pupil's attitude diverges from his belief about the pupil's talent if and only if the pupil seems to be talented in one field but not the other. This is intuitive, since teachers discuss their pupils in the faculty room, and very likely some influence the others in their opinion.

3.3 Preferences of teachers

Define the perceived attitude $\widehat{a}_{ri}$ that teacher r suspects student i to have as follows:

$\widehat{a}_{ri} = 1$ iff the posterior belief of teacher r about a_i is
$\Pr_r\{a_i = 1 \mid s_{rij}, s_{r'ij'}\} > \frac{1}{2}$;
$\widehat{a}_{ri} = \frac{1}{2}$ iff the posterior belief of teacher r about a_i is
$\Pr_r\{a_i = 1 \mid s_{rij}, s_{r'ij'}\} = \frac{1}{2}$; and
$\widehat{a}_{ri} = 0$ otherwise.

Besides, define $d_i(m)$ as the decision of student i, given his or her grades $m = (m_{ij}, m_{ij'})$ in fields j and j', with $j \neq j'$. Let $d_i(m) = 1$ when student i decides to enroll in subject field j at university, and $d_i(m) = 0$ otherwise.

Then, the expected utility that any teacher r gets from giving student i the grade m_{rij} in field j amounts to

$$V_r\left(m_{rij}\right)=\left(\widehat{a}_{ri}-\frac{1}{2}\right)\left(m_{rij}-s_{rij}\right)-c\left[d_i\left(s_{rij},m_{r'ij'}\right)-d_i\left(m_{rij},m_{r'ij'}\right)\right]^2,\ j\neq j', \tag{5}$$

with

$$c>\frac{1}{2}\left(1-k\right). \tag{6}$$

The expected utility of the other teacher r' is defined analogously. It is easy to deduce from (5) and (6) that a weakly attitude-sensitive teacher will exhibit the following behaviour: If the student has most probably a likable attitude ($\widehat{a}_{ri}=1$) and the teacher's signal has been positive ($s_{rij}=1$), then the teacher will report his signal honestly ($m_{rij}=1$). Similarly, if the student has most probably a less likable attitude ($\widehat{a}_{ri}=0$) and the teacher's signal has been negative ($s_{rij}=k$), the teacher will transmit his information ($m_{rij}=0$). But if the teacher's signal runs contrary to the perceived attitude of the student, the teacher has an incentive to lie and use the grade as a positive or negative reaction to the student's attitude alone. He will only do so, however, if distorting the grade in this way does not change the student's decision about his or her future career.

To understand the nature of the teachers' preferences and their resulting behaviour more deeply, consider the two expressions in (5) separately. The first expression represents the teacher's emotional payoff from grading. If the teacher suspects the pupil of having an attitude that the teacher does not like, he has a purely emotional incentive to give his pupil a bad grade. However, if the teacher believes his pupil to have most probably a likable attitude, he is emotionally inclined to give a good grade. Thus, the first expression in the teacher's utility function represents the *direct utility* the teacher has from giving a good or bad grade.[17]

By contrast, the second expression in (5) stands for the teacher's *indirect costs* of biased grading, i.e. costs that he might incure due to his pupil's reaction to the distorted grade. The intuition behind this second term is that teachers feel a (limited) responsability for their pupils; they do not want to

[17] I assume that a teacher does not react to the intesity of his belief about the pupil's attitude, but only to the direction of this belief, i.e. whether a good or a bad attitude is more probable. In the simple model of this paper with only two possible grades, the assumption that teachers react to the intesity of their belief would not contribute much to realism, although it would of course make sense in a more complicated model with a broader message-space. I thank an anonymous referee to point that out to me.

do them too much harm by biased grading. Thus, teachers partly internalize their pupils' welfare. To see this, note that a weakly attitude-sensitive teacher will have an incentive to distort his grade only if the signals about his pupil's talents are either $s = (1, k)$ or $s = (k, 1)$, i.e. if the pupil seems to be talented in one field but not the other. In this case, biased grading can in general have two consequences for the pupil's future career: Either the pupil chooses the field in which he or she is most probably talented but does not invest the optimal effort. Or the pupil chooses the wrong career path *and* fails therefore to invest the optimal effort, too. For certain paramters, e.g. for a low k and w_0, the second consequence (choosing the wrong career path) is worse than the first one (choosing the correct career path but suboptimal effort) in terms of the pupil's welfare. Thus, the assumption that teachers incure costs from making their pupils choose suboptimal career paths but not from making them choose suboptimal efforts can be rationalized by the assumption that teachers partly internalize their pupils' welfare: Teachers like to express their emotions by biased grading, but not when the welfare losses that the pupil would incure in consequence would become too pronounced.[18]

The following Lemma reports when biased grading actually occurs.

Lemma 2 (1) If $s_{rij} = k$, a *weakly attitude-sensitive* teacher r who is the *first mover* in the grading continuation game distorts his grade of student i in field j upwards when $s_{r'ij'} = 1$ and $d_i\left(1, m_{r'ij'}\right) = d_i\left(k, m_{r'ij'}\right)$for $j \neq j'$. (2) If $s_{rij} = 1$, a weakly attitude-sensitive teacher r who is the first

[18] Indeed, the model could be extended so as to relate the teachers' preferences more directly and explicitly to the welfare of their pupils. First, one could include a direct "consumption" utility from getting good grades (and a disutility from getting bad ones) into the pupils' preferences. Then, the first expression in (5) would describe how teachers internalize this utility, depending on the perceived attitude of the pupil. Second, one could replace the second expression in (5) by a weighted function of the pupil's welfare from his/her career decisions. Then, one would have to specify the paramters for which the equilibrium results of the current paper still hold. However, this would make the calculations (and the paper itself) much more complicated and lengthy. Among other things, the teachers' updating of their beliefs would become more complicated. Besides, it is an open question whether teachers really fully internalize their puplis' welfare (in the positive or in the negative) or whether they just follow a thumb rule. For a related model in which the teachers internalize their puplis' welfare (but often in a biased way), compare Mechtenberg (2006). I thank two anonymous referees and one editor for directing my attention to the relation between the teachers' preferences and the pupils' welfare in the current model.

mover in the grading continuation game distorts his grade of student i in field j downwards when $s_{r'ij'} = k$, $j \neq j'$, and $d_i\left(1, m_{r'ij'}\right) = d_i\left(k, m_{r'ij'}\right)$. (3) strongly attitude-sensitive teachers who are first movers in the grading continuation game always report their signal, $m_{rij} = s_{rij}$. (4) The second mover r' in the grading continuation game behaves as follows. If the first mover r did not distort his grade, r' acts as if he was a first mover. In all other cases, r' reports his signal, $m_{r'ij'} = s_{r'ij'}$.

Proof of Lemma 2 See Appendix. □

To understand this result intuitively, note that only weakly attitude-sensitive teachers have the incentive to distort their grade. The reason is as follows. Strongly attitude-sensitive teachers follow only their own signal when updating their beliefs about their pupil's attitude. Consequently, the belief of a strongly attitude-sensitive teacher about his pupil's attitude is always in line with his belief about his pupil's talent in his field, and he does never have an incentive to distort his grade. By contrast, weakly attitude-sensitive teachers follow their collegue's signal when updating their beliefs about their pupil's attitude. The weakly attitude-sensitive teacher has therefore divergent beliefs about his pupil's talent and attitude when his pupil seems to be talented in one field but not the other. In this case, the weakly attitude-sensitive teacher has an incentive to distort his grade. He will do so, however, only if his grade will not distort the pupil's choice of a future career, given the grade of the other teacher. Thus, a teacher gives a biased grade only if he is not pivotal for the career choice of his pupil. The condition that is necessary and sufficient for non-pivotality of ateacher is given in a

Remark For w_0 low enough, a teacher in field j will not be pivotal for his pupil's career choice if and only if the grade that he wants to give is credible in field j but would not be credible in field j'.

To see this, consider (w.l.o.g.) the grades (a) $m = (1, m_2)$ and (b) $m = (k, m_2)$. In (a), the teacher in field 2 will not be pivotal if and only if the grade 1 is credible in field 1 but not in field 2, since only in this case, the pupil chooses field 1 independenly of m_2. In (b), the teacher in field 2 will not be pivotal if and only if the grade k is credible in field 1 but not in field 2, since only then, the pupil chooses field 2 independently of m_2, given w_0 is low enough.

4 The U.S. equilibrium and gender differences in enrollment decisions

The model has an equilibrium in which all teachers report their signals honestly, so that there are no differences between the decisions of boys and girls. To see this, note that any teacher will be honest if he is pivotal for the pupil's career descision, and suppose c to be high enough so that this holds true even if career decisions are mixed strategies. Suppose now that both boys and girls always believe their teachers to be honest. Then, a pupil with message $m = (1, k)$ or $m = (k, 1)$ will choose the field that he or she is better in for a future career; a pupil with message $m = (1, 1)$ will (supposedly) flip a coin; and a pupil with message $m = (k, k)$ will enter the labor market directly. Consequently, both the math teacher and the teacher in humanities are always pivotal for their pupils' career decisions and will therefore be always honest, thereby justifying the trusting beliefs of their pupils. This equilibrium is efficient, since all the information is transmitted.

However, the empirical literature on grading behaviour discussed in section 2 implies that *not* all teachers report their signals honestly to all pupils; and there *are* differences between the careers of men and women. Thus, I concentrate on equilibria where at least some pupils suffer from a loss of information due to biased grading.

There are more than one such equilibrium, but they are limited in number. First, there is no babbling equilibrium. To see this, suppose for the moment that teacher r in field j babbles to his pupil i. Then, the pupil does not believe the grade m_{rij} to be in the least informative. Hence, i chooses his or her career path independently of m_{rij}, and teacher r is therefore not pivotal. Consequently, r will want to use m_{rij} as a reaction to his belief about i's attitude alone. If r is strongly attitude-sensitive, this incentive will make him reporting his signal honestly. But this, in equilibrium, must be anticipated by i, so that i uses m_{rij} to update his or her belief about his or her talent in j. Since this is a contradiction, there is no babbling equilibrium. Second, given that an indifferent pupil flips a coin to decide about his or her future career path, there are no equilibria in which a given pupil can be discriminated (patronized) in both fields, M and H, at once. This follows from the condition for non-pivotality that is stated in the above Remark. Thus, an equilibrium with biased grading in this model will always be asymmetric in one sense or the other.

In the remainder of the paper, I will show that the game described in the sections above has a pure strategy equilibrium that is characterized by all the gender differences observable in half of the OECD countries, including the U.S. This is true even though boys and girls are ex ante identical. For simplicity, the respective equilibrium shall be referred to as the *U.S. equilibrium*.[19]

In order to formally characterize the U.S. equilibrium, define $D_G\left(m_M^G, m_H^G\right) \in \{H, M, w_0\}$ as the career decision variable of a female student, given her grade m_M^G in math and her grade m_H^G in humanities; and define $D_B\left(m_M^B, m_H^B\right) \in \{H, M, w_0\}$ as the corresponding career decision variable of a male student. Let r_H be a teacher in humanities who is weakly attitude-sensitive; and let r_M be a teacher in math who is weakly attitude-sensitive, too. The message that r_H sends to a boy i_B when he would be the first to distort his message to i_B shall be denoted by $m_{riH}^B\left(s_{r'iM}^B, s_{riH}^B\right)$, with $s_{r'iM}^B$ and s_{riH}^B denoting the two signals on i_B's abilities in math and humanities, respectively. The corresponding message that r_H sends to a girl i_G shall be denoted by $m_{riH}^G\left(s_{r'iM}^G, s_{riH}^G\right)$. Similarly, the message that r_M sends to a boy i_B when he would be the first to distort his message to i_B is $m_{riM}^B\left(s_{riM}^B, s_{r'iH}^B\right)$; and the corresponding message that r_M sends to a girl i_G is $m_{riH}^G\left(s_{riM}^G, s_{r'iH}^G\right)$. Note that the equilibrium strategies of the teachers are fully described by giving m_{riH}^B, m_{riH}^G, m_{riM}^B, and m_{riH}^G, because Lemma 2 already implies that all other possible messages, including those of other types of teachers, must be honest anyway.

Theorem 1 There exists a pure strategy equilibrium ("U.S. equilibrium") of the game that is characterized by the following strategies. (1) Career decsions of female students are: $D_G(1,1) = H$, $D_G(1,k) = M$, $D_G(k,1) = H$, $D_G(k,k) = H$. (2) Career decsions of male students are: $D_B(1,1) = M$, $D_B(1,k) = M$, $D_B(k,1) = H$, $D_B(k,k) = w_0$. (3) r_H grades boys as follows: $m_{riH}^B\left(1, s_{riH}^B\right) = 1\ \forall s_{riH}^B$, and $m_{riH}^B\left(k, s_{riH}^B\right) = s_{riH}^B\ \forall s_{riH}^B$. (4) r_H grades girls as follows: $m_{riH}^G\left(1, s_{riH}^G\right) = s_{riH}^G\ \forall s_{riH}^G$, and $m_{riH}^G\left(k, s_{riH}^G\right) = k\ \forall s_{riH}^G$. (5) r_M grades boys as follows: $m_{riM}^B\left(s_{riM}^B, s_{r'iH}^B\right) = s_{riM}^B\ \forall s_{riM}^B, s_{r'iH}^B$. (6) r_M grades girls as follows: $m_{riM}^G\left(s_{riM}^G, 1\right) = 1\ \forall s_{riM}^G$, and $m_{riM}^G\left(s_{riM}^G, k\right) = s_{riM}^G\ \forall s_{riM}^G$.

[19] This does not mean that the phenomena described are more prominent in the U.S. than in other OECD countries that are decribed by the current model. On the contrary, this conclusion would be wrong. I choose the label "U.S. equilibrium" only because the U.S. is the most important country among those to which the model applies.

Proof See Appendix. □

For a better understanding, the directions of the equilibrium grading biases and the equilibrium career decisions of the students in reaction to their grades are depicted in the table below. The meanings of the arrows are as follows. The upward-pointing arrow in the second cell in the first column of the girls' table indicates that by patronizing girls in math, teachers "shift" some girls from the second to the first cell in the column. The arrow pointing to the right means that by being too strict toward girls in humanities, teachers "shift" some girls from the second cell in the first column to the second cell in the second column. Finally, the arrow pointing to the left in the first cell of the second column in the boys' table indicates that by patronizing boys in humanities, teachers "shift" some boys from the first cell of the second column to the first cell of the first column.

Girls	$m_H = 1$	$m_H = k$	**Boys**	$m_H = 1$	$m_H = k$
$m_M = 1$	H	M	$m_M = 1$	M	$\Longleftarrow M$
$m_M = k$	$\Uparrow$ $H \Longrightarrow$	H	$m_M = k$	H	w_0

Intuitively, the equilibrium can be described like this. Girls trust a good grade in humanities, but they do not trust a good grade in math. Thus, girls who get a good grade in humanities always choose humanities for their future career, independently of their grade in math. A math teacher is therefore never pivotal for the career choice of a girl who is good in humanities. Consequently, a weakly attitude-sensitive math teacher indulges in patronizing a girl who seems to be good in humanities. Thereby, he justifies her distrust in good math-grades.

A similar argument holds with regard to girls who get a bad grade in math. Note that a bad grade in math must come from an honest teacher. Therefore, girls are justified in taking it seriously. However, they do not believe a bad grade in humanities to be very informative. Therefore, they choose humanities for their future career even if their grade in humanities is bad. Thus, the teacher in humanities is never pivotal for the career choice of girls who get a bad grade in math. A weakly attitude-sensitive teacher in humanities feels therefore free to give downward-biased grades to girls who seem to be bad in math. Again, his behaviour justifies the girls' beliefs: Good grades in humanities are to be trusted, bad ones are not.

Consider now the boys. They always fully believe their math teacher. Thus, the math teacher is always pivotal for a boy's career choice and grades him therefore honestly, thereby justifying the boy's beliefs about his grades in math. However, boys do not trust a good grade in humanities. Consequently, they will never choose humanities for their future career if they get a good grade in math; they will choose math instead. Thus, the humanity teacher is never pivotal for the career choice of a boy who seems to be good at math; and the teacher can therefore feel free to patronize this boy in humanities. Again, the teacher's behaviour justifies the beliefs of the pupil.

In general, both boys and girls suffer from a loss of information at school. However, this loss is greater for the girls. For girls, the bad grades in humanities are noisier than they would be if all teachers were honest, whereas in math, the good grades are too noisy. For boys, only the good grades in humanities are noisier than the other grades.

To analyze the equilibrium more formally, consider first the girls. Define

$$\pi_M^G\left(m_{iM}, m_{iH}\right) \equiv \Pr\left\{\theta_{iM}=1 \mid m_i=\left(m_{iM}, m_{iH}\right), i=i_G\right\}$$

to be the posterior confidence of a girl in math after having received the grades $m=(m_M, m_H)$ in math and humanities, respectively. Accordingly, the girl's confidence in humanities and the boys' confidence in math and humanities are denoted as

$$\begin{aligned}
\pi_H^G\left(m_{iM}, m_{iH}\right) &\equiv \Pr\left\{\theta_{iH}=1 \mid m_i=\left(m_{iM}, m_{iH}\right), i=i_G\right\}, \\
\pi_M^B\left(m_{iM}, m_{iH}\right) &\equiv \Pr\left\{\theta_{iM}=1 \mid m_i=\left(m_{iM}, m_{iH}\right), i=i_B\right\}, \text{ and} \\
\pi_H^B\left(m_{iM}, m_{iH}\right) &\equiv \Pr\left\{\theta_{iH}=1 \mid m_i=\left(m_{iM}, m_{iH}\right), i=i_B\right\}.
\end{aligned}$$

Theorem 2 reports the precise distribution of female students over all four possible tuples of grades and their respective posterior beliefs about their talents.

Theorem 2 In the U.S. equilibrium, (1) there are $\left[\frac{1}{4}(1+\alpha)-\frac{1}{8}\alpha^2\right]$ girls with $m=(1,1)$, $\pi_M^G(1,1)=\frac{\sigma+(1-\sigma)\left(\alpha-\frac{1}{2}\alpha^2\right)}{1+\left(\alpha-\frac{1}{2}\alpha^2\right)}<\sigma$, and $\pi_H^G(1,1)=\sigma$. (2) There are $\frac{1}{4}$ girls with $m=(1,k)$, $\pi_M^G(1,k)=\sigma$, and $\pi_H^G(1,k)=(1-\sigma)$. (3) There are $\frac{1}{4}(1-\alpha)^2$ girls with $m=(k,1)$, $\pi_M^G(k,1)=(1-\sigma)$, and $\pi_H^G(k,1) = \sigma$. (4) There are $\left[\frac{1}{4}(1+\alpha) - \frac{1}{8}\alpha^2\right]$ girls with $m = (k,k)$, $\pi_M^G(k,k) = (1-\sigma)$, and $\pi_H^G(k,k) = \frac{(1-\sigma)+\sigma\left(\alpha-\frac{1}{2}\alpha^2\right)}{1+\left(\alpha-\frac{1}{2}\alpha^2\right)} > (1-\sigma)$. (5) Girls with grades $m = (1,1)$ are underconfident in math; and girls with with grades $m = (k,k)$ are overconfident in humanities.

Proof See Appendix. □

Consider now the boys. Theorem 3 reports their distribution over tuples of grades and their respective posterior beliefs formally.

Theorem 3 In the U.S. equilibrium, (1) there are $\frac{1}{4}(1+\alpha)$ boys with $m = (1,1)$, $\pi_M^B(1,1) = \sigma$, and $\pi_H^B(1,1) = \frac{\sigma+(1-\sigma)\alpha}{1+\alpha} < \sigma$. (2) There are $\frac{1}{4}(1-\alpha)$ boys with $m = (1,k)$, $\pi_M^B(1,k) = \sigma$, and $\pi_H^B(1,k) = (1-\sigma)$. (3) There are $\frac{1}{4}$ boys with $m = (k,1)$, $\pi_M^B(k,1) = (1-\sigma)$, and $\pi_H^B(k,1) = \sigma$. (4) There are $\frac{1}{4}$ boys with $m = (k,k)$, $\pi_M^B(k,1) = (1-\sigma)$, and $\pi_H^B(k,k) = (1-\sigma)$. (5) Boys with grades $m = (1,1)$ are underconfident in humanities.

Proof See Appendix. □

Given Theorem 1, the distributions of males and females over all possible career decisions can be inferred directly from Theorems 2 and 3. As one can easily see, the students' career decisions in the U.S. equilibrium reveal significant gender differences: More female than male students go to university after school, but more male students study math and sciences. The respective formal results are summarized in[20]

Corollary 1 (1) In the U.S. equilibrium, all girls but only $\frac{3}{4}$ of the boys go to university. (2) Only $\frac{1}{4}$ of the girls but $\frac{3}{4}$ of the boys study math and sciences. (3) Boys do not study arts and humanities, but $\frac{3}{4}$ of the girls do.

In qualitative terms, the career decisions of boys and girls in the U.S. equilibrium that are reported in Corollary 1 reflect the differences in enrollment decisions between male and female students within the broad majority of the OECD countries. For instance, OECD data reveal that the percentage of first tertiary degrees (degrees in type-B-programs) rewarded to females in

[20] Obviously, the quantitative results in Corollary 1 are due to the specific distribution of individuals over the type space. But the qualitative results can be generalized.

math and sciences lies below 35 percent in 16 of 23 OECD countries with available data, including the U.S. and the U.K.; and it lies below 50 percent in 22 of these 23 OECD countries. By contrast, in humanities, arts and education, the percentage of first tertiary degrees rewarded to females lies above or equals 70 percent in 14 of 28 OECD countries, including the U.S. with 81 percent; and it lies above or equals 60 percent in 26 of 28 OECD countries. Besides, the percentage of first tertiary degrees in all fields that are rewarded to females lies above or equals 60 percent in 12 of 28 OECD countries; and it lies above 50 percent in 23 of 28 OECD countries.[21]

In my model, the predominance of females at the university in general and in arts and humanities in particular, and the absence of females in math and sciences originate from biased grading solely. Thus, it is possible to explain these phenomena without referring to any possible differences in innate abilities or ex ante preferences of boys and girls.[22]

5 Why girls outperform boys in reading literacy

The predominance of female students in arts and humanities at the university suggests that girls perform better than boys when confronted with tasks that are typical for university programs in arts and humanities. A prevalent task in humanities is to read and interpret complex texts. Therefore, the PISA test on reading literacy necessitates abilities and efforts similar to those needed for a university program in humanities.[23] Not surprisingly, the high enrollment rates of females in humanities and the respective low rates of males are in line with the empirical fact that girls outperform boys in reading literacy, being better on average and having less bottom and more top achievers.

However, rash conclusions about underlying differences in abilities should be avoided. Obviously, PISA test scores do not measure talent but rather achievement, a combination of talent and effort. Because effort invested

[21] Compare Table A3.4, OECD, www.oecd.org/edu/eag2006.

[22] This, of course, does not imply that no such innate differences exist. It rather means that grading practices should become a more important topic in empirical research.

[23] During the PISA tests on reading literacy, 15-year-old pupils are asked to retrieve and interpret verbally coded information. Compare the PISA 2003 report, p. 268.

by students into the PISA test is supposedly already affected by their self-confidence, biased grading has an influence on the resulting achievement distributions of boys and girls.

How students perceive their (strategic) situation during the PISA test is not clear; but it is plausible to assume that their effort during the test is highly correlated with the effort they would invest if forced to study a subject field comprising similar tasks. Therefore, I assume the hypothetical effort that the boys and girls in my model would invest if forced to study humanities to be equal to the effort they would invest into a PISA test on reading literacy.

These efforts can be defined by substituting into equation (2) the equilibrium beliefs of boys and girls about their respective talents in humanities. The results are:

$$\begin{aligned}
e_{h1} &\equiv \frac{1}{2}\left[\sigma\left(1-k\right)+k\right], \\
e_{h2} &\equiv \frac{1}{2}\left[\frac{\sigma+\left(1-\sigma\right)\alpha}{1+\alpha}\left(1-k\right)+k\right], \\
e_{l1} &\equiv \frac{1}{2}\left[\frac{\left(1-\sigma\right)+\sigma\left(\alpha-\frac{1}{2}\alpha^2\right)}{1+\left(\alpha-\frac{1}{2}\alpha^2\right)}\left(1-k\right)+k\right], \text{ and} \\
e_{l2} &\equiv \frac{1}{2}\left[\left(1-\sigma\right)\left(1-k\right)+k\right], \text{ where} \\
e_{h1} &> e_{h2} > e_{l1} > e_{l2}.
\end{aligned}$$

A student's achievement in reading literacy that is measured by the respective PISA scores is assumed to be $\theta_{iH}e_{iH}$, with e_{iH} representing the effort that student i would invest if studying humanities. Thus, bottom achievers in reading literacy will be students lacking talent in humanities ($\theta_{iH} = k$) who choose the lowest possible equilibrium effort, $e_{iH} = e_{l2}$. By contrast, top achievers in reading literacy will be those students who are highly talented in humanities ($\theta_{iH} = 1$) and choose high effort, $e_{iH} = e_{h1}$.

In order to see how biased grading by teachers of humanities affects the achievement distributions of boys and girls in reading literacy, compare first the respective fractions of bottom and top achievers. The fraction n_{l2}^G of female bottom achievers in reading literacy in the U.S. equilibrium is

$$n_{l2}^G = \Pr\left\{\theta_{iH} = k \wedge m_i = (1,k) \mid i = i_G\right\} = \frac{1}{4}\sigma,$$

while the corresponding fraction of male bottom achievers is

$$\begin{aligned} n_{l2}^{B} &= \Pr\{\theta_{iH} = k \wedge (m_i = (1,k) \vee m_i = (k,k)) \mid i = i_B\} \\ &= \frac{1}{4}\sigma(2-a) > \frac{1}{4}\sigma. \end{aligned}$$

Thus, boys have more bottom achievers in reading literacy than girls. This effect is due to teachers of humanities using their grades partly as a reaction to incongruous attitudes. The corresponding carelessness with which girls consider bad grades in humanities leads to overconfidence among the lowly talented girls: They achieve more than the lowly talented boys who take their bad grades in the non-mathematical classes seriously.

Consider now the top achievers. There again, the difference between boys and girls is in favor of the girls. To see this, consider first the fraction n_{h1}^{G} of female top achievers in reading literacy:

$$\begin{aligned} n_{h1}^{G} &= \Pr\{\theta_{iH} = 1 \wedge (m_i = (1,1) \vee m_i = (k,1)) \mid i = i_G\} \\ &= \frac{1}{4}\sigma\left[2 - \left(\alpha - \frac{1}{2}\alpha^2\right)\right]. \end{aligned}$$

It is straightforward to show that the fraction n_{h1}^{B} of male top achievers is lower:

$$n_{h1}^{B} = \Pr\{\theta_{iH} = 1 \wedge m_i = (k,1) \mid i = i_B\} = \frac{1}{4}\sigma < \frac{1}{4}\sigma\left[2 - \left(\alpha - \frac{1}{2}\alpha^2\right)\right].$$

The reason for this gender difference among students gifted with reading literacy is that teachers in humanities often use their grades to "reward" boys for supposedly good attitudes. Boys react with mistrusting praise in humanities; and consequently, highly talented boys achieve less than highly talented girls in tests of reading literacy. These distributional results of biased grading in humanities in the U.S. equilibrium are summarized in

Proposition 1 In the U.S. equilibrium, biased grading in school shifts the achievement distribution of girls in the reading literacy test to the right of the corresponding achievement distribution of boys: There are more girls than boys with top achievement e_{h1} and less girls than boys with bottom achievement ke_{l2}.

This result is already of some interest, because the overconfidence of girls with low reading literacy and the underconfidence of boys with high reading literacy already suggest that men will specialize in mathematical subject fields or avoid the university at all, while women will fill the lecture rooms in humanities even if they are not highly talented. Besides, the distributional effect reported in Proposition 1 is exactly what can be observed from the PISA data in all OECD countries.[24]

Nevertheless, this first observation is not sufficient. In order to fully understand the distributional effects of biased grading in humanities at school, one has to compare the average achievements of boys and girls. Consider first the students with high talents in humanities ($\theta_{iH} = 1$). The average achievement A_{H1}^{G} in the reading literacy test of girls with $\theta_{iH} = 1$ amounts to

$$A_{H1}^{G} = \frac{1}{2}(1-\sigma)\,e_{l2} + \frac{1}{2}\left[(1-\sigma)+\sigma\eta\right]e_{l1} + \frac{1}{2}\left[2\sigma - \sigma\eta\right]e_{h1}, \text{ with} \tag{7}$$

$$\eta \equiv \alpha - \frac{1}{2}\alpha^{2},$$

while the corresponding average achievement A_{H1}^{B} of boys with $\theta_{iH} = 1$ is

$$A_{H1}^{B} = \frac{1}{2}(1-\sigma)(2-\alpha)\,e_{l2} + \frac{1}{2}\left[\sigma + (1-\sigma)\,\alpha\right]e_{h2} + \frac{1}{2}\sigma e_{h1}. \tag{8}$$

Let Δ_{H1} represent the difference between the highly talented girls' and the highly talented boys' average achievement in the reading literacy test. Subtracting (9) from (8), substituting for the efforts, and simplifying yields

$$\Delta_{H1} = \frac{1}{4}(1-k)\,\frac{\alpha^{2}\left[\frac{1}{2} - 2\sigma(1-\sigma)\right]}{(1+\alpha)\left[1+\left(\alpha - \frac{1}{2}\alpha^{2}\right)\right]} > 0. \tag{9}$$

[24] Compare the PISA 2003 report, p. 282 and table 6.4.

Thus, girls who are talented in humanities achieve more on average in the reading literacy test than boys with the same talent. This effect is due to the underconfidence of talented boys who are praised by their teacher of humanities but do not trust this praise.

Consider next the average achievements in the reading literacy test of boys and girls who are lowly talented in humanities ($\theta_{iH} = k$). Lowly talented girls achieve on average

$$A_{Hk}^{G} = \frac{1}{2}\sigma k e_{l2} + \frac{1}{2}k\left[(1-\sigma)\eta + \sigma\right]e_{l1} + \frac{1}{2}k\left(2-\eta\right)(1-\sigma)\,e_{h1}, \text{ with} \quad (10)$$

$$\eta \equiv \alpha - \frac{1}{2}\alpha^{2},$$

while the average achievement of lowly talented boys amounts to

$$A_{Hk}^{B} = \frac{1}{2}\sigma k\left(2-\alpha\right)e_{l2} + \frac{1}{2}k\left[(1-\sigma) + \sigma\alpha\right]e_{h2} + \frac{1}{2}k\left(1-\sigma\right)e_{h1}. \quad (11)$$

After subtracting (12) from (11), substituting for the efforts, and simplifying, the difference Δ_{Hk} between the average achievements of girls and boys who are lowly talented in humanities can be specified as follows:

$$\Delta_{Hk} = \frac{1}{4}k\left(1-k\right)\frac{\alpha^{2}\left[\sigma\left(1-\sigma\right) - \frac{1}{2}\sigma^{2} - \frac{1}{2}\left(1-\sigma\right)^{2}\right]}{\left(1+\alpha\right)\left[1+\left(\alpha-\frac{1}{2}\alpha^{2}\right)\right]} < 0. \quad (12)$$

As one can see from (13), the relation between average achievements of girls and boys in the reading literacy test is reversed at the lower end of the achievement distribution, compared with the upper end: Lowly talented boys achieve more on average than lowly talented girls. The reason for this effect is that the weakly attitude-sensitive teachers in humanities patronize boys. Thus, altough the lowly talented boys mistrust their good grades, they internalize their praise at least partly and invest and achieve more in consequence than the lowly talented girls who got an honest mark.

The question is now how these two effects together, i.e. the higher average achievement of highly talented girls and of lowly talented boys, affect the relation between the overall average achievements of boys and girls in the reading literacy test. Define $\Delta_H \equiv \Delta_{H1} + \Delta_{Hk}$. Adding (10) and (13) and simplifying the result yields

$$\Delta_H = \frac{1}{4}(1-k)\frac{\alpha^2\left[\frac{1}{2} - 2\sigma(1-\sigma)\right] + \alpha^2 k\left[\sigma(1-\sigma) - \frac{1}{2}\sigma^2 - \frac{1}{2}(1-\sigma)^2\right]}{(1+\alpha)\left[1 + \left(\alpha - \frac{1}{2}\alpha^2\right)\right]}. \tag{13}$$

This specification of the difference between average achievements of girls and boys in the reading literacy test implies

Proposition 2 In the U.S. equilibrium, biased grading in school has the effect that girls achieve more on average than boys in the reading literacy test: $\Delta_H > 0 \; \forall k < 1$.

Proof Define

$$N \equiv \frac{1}{4}(1-k)\left[\alpha^2\left[\frac{1}{2} - 2\sigma(1-\sigma)\right] + \alpha^2 k\left[\sigma(1-\sigma) - \frac{1}{2}\sigma^2 - \frac{1}{2}(1-\sigma)^2\right]\right].$$

Because $signum(\Delta_H) = signum(N)$, it suffices to analyze N. It is straightforward to show that for $k \in [0,1]$, both $N = 0$ and $\frac{dN}{dk} = 0$ if and only if $k = 1$, and $\frac{dN}{dk} < 0$ otherwise. Besides, $\frac{d^2N}{dk^2} > 0 \; \forall k \in [0,1]$. This implies that N has its minimum value at $k = 1$, this minimum value equals zero, and with decreasing k, N increases monotonously. Thus, $\Delta_H > 0$ $\forall k \in [0,1)$. □

Intuitively, the prevalence of the girls' advantage in the overall average achievement can be explained as follows. On the one hand, lowly talented boys who are made overconfident by undeserved praise invest more effort than they would have done if they were graded honestly. Besides, they invest also more than the lowly talented girls who are overconfident in spite of their bad grade. The reason is that the overconfidence of the lowly talented boys after undeserved praise is stronger than the overconfidence of the lowly talented girls after a criticism that they suspect to be undeserved. But because these boys are lowly talented, their excess effort does not fully translate into a corresponding excess achievement. On the other hand, however, many highly talented boys are made underconfident because they never fully internalize a good grade, while comparatively few highly talented girls become underconfident. The reduction of the boys' effort fully translates into a corresponding reduction of achievement. Therefore, the decrease in achievement of the highly talented boys outweighs the increase in achievement of the lowly talented boys. Thus, the achievement distributions of boys and girls in reading literacy in the U.S. equilibrium differ from each other exactly in the way that the PISA data show for the U.S. and all other OECD countries.

6 Why boys outperform girls in math and sciences

In math and sciences, the situation at school is almost, but not exactly, the opposite of the corresponding situation in humanities. According to the PISA data, boys have more top achievers and achieve more on average in the PISA math test than girls in all OECD countries. But in half of the OECD countries, including the U.S., boys also have more bottom achievers in the math test than girls.

Similar as before, the efforts in math can be defined by substituting into equation (2) the equilibrium beliefs of boys and girls about their respective talents in math and sciences:

$$\begin{aligned}
e_{h1} &\equiv \frac{1}{2}\left[\sigma\left(1-k\right)+k\right], \\
e_{h3} &\equiv \frac{1}{2}\left[\frac{\sigma+\left(1-\sigma\right)\left(\alpha-\frac{1}{2}\alpha^{2}\right)}{1+\left(\alpha-\frac{1}{2}\alpha^{2}\right)}\left(1-k\right)+k\right], \text{ and} \\
e_{l2} &\equiv \frac{1}{2}\left[\left(1-\sigma\right)\left(1-k\right)+k\right], \text{ where} \\
e_{h1} &> e_{h2}>e_{h3}>e_{l1}>e_{l2}.
\end{aligned}$$

A student's achievement in math measured by the PISA math test scores is therefore assumed to be $\theta_{iM}e_{iM}$, with e_{iM} representing the effort that student i would invest if studying math and sciences. Thus, bottom achievers in math will be students achieving only ke_{l2}, and top achievers in math will be students achieving e_{h1}.

The fraction ν_{l2}^{G} of female bottom achievers in math in the U.S. equilibrium is

$$\begin{aligned}\nu_{l2}^{G} &= \Pr\left\{\theta_{iM} = k \wedge (m_i = (k,1) \vee m_i = (k,k)) \mid i = i_G\right\} \\ &= \frac{1}{4}\sigma\left[2 - \left(\alpha - \frac{1}{2}\alpha^2\right)\right].\end{aligned}$$

It is easy to see that the corresponding fraction of male bottom achievers is higher:

$$\begin{aligned}\nu_{l2}^{B} &= \Pr\left\{\theta_{iM} = k \wedge (m_i = (k,1) \vee m_i = (k,k)) \mid i = i_B\right\} \\ &= \frac{1}{2}\sigma > \frac{1}{4}\sigma\left[2 - \left(\alpha - \frac{1}{2}\alpha^2\right)\right].\end{aligned}$$

The reason for this difference is that less lowly talented girls than lowly talented boys get a bad grade: Girls who are good in humanities are patronized in math.

Consider now the top achievers. There, the difference between boys and girls is in favor of the boys. To see this, consider first the fraction ν_{h1}^{G} of female top achievers in math:

$$\nu_{h1}^{G} = \Pr\left\{\theta_{iM} = 1 \wedge m_i = (1,k) \mid i = i_G\right\} = \frac{1}{4}\sigma$$

It is straightforward to show that the fraction ν_{h1}^{B} of male top achievers is higher:

$$\begin{aligned}\nu_{h1}^{B} &= \Pr\left\{\theta_{iM} = 1 \wedge (m_i = (1,1) \vee m_i = (1,k)) \mid i = i_B\right\} \\ &= \frac{1}{2}\sigma > \frac{1}{4}\sigma.\end{aligned}$$

Again, the reason for this gender difference lies in the math teachers patronizing girls who are good in humanities: While all boys with good marks in math can trust their praise, girls who are also good in humanities cannot do so. Thus, they are underconfident and achieve less. These distributional results of biased grading in humanities in the U.S. equilibrium are summarized in

Proposition 3 In the U.S. equilibrium, biased grading in mathematical classes at school creates mediocrity among girls: There are both less girls than boys with top achievement e_{h1} and less girls than boys with bottom achievement ke_{l2}.

This result does not only mirror the empirical gender differences that the PISA data show for about half of the OECD countries, including the U.S. But it also implies that due to having been patronized in math at school, talented women are on average less confident in their mathematical skills and therefore less productive in this area than men.[25] Consider now the average achievements, starting with students with high talents in math ($\theta_{iM} = 1$). The average achievement A_{M1}^{G} in the PISA math test of girls with $\theta_{iM} = 1$ amounts to

$$A_{M1}^{G} = \frac{1}{2}(1-\sigma)[2-\eta]e_{l2} + \frac{1}{2}[\sigma + (1-\sigma)\eta]e_{h3} + \frac{1}{2}\sigma e_{h1}, \text{ with} \tag{14}$$

$$\eta \equiv \alpha - \frac{1}{2}\alpha^{2},$$

while the corresponding average achievement A_{M1}^{B} of boys with $\theta_{iM} = 1$ is

$$A_{M1}^{B} = (1-\sigma)e_{l2} + \sigma e_{h1}. \tag{15}$$

Let Δ_{M1} represent the difference between the highly talented girls' and the highly talented boys' average achievement in the math test. Subtracting (16) from (15), substituting for the efforts, and simplifying yields

$$\Delta_{M1} = -\frac{\left(\frac{1}{2}-\sigma\right)^{2}\left(\alpha-\frac{1}{2}\alpha^{2}\right)(1-k)}{1+\left(\alpha-\frac{1}{2}\alpha^{2}\right)} < 0. \tag{16}$$

Thus, girls who are talented in math achieve less on average than boys with the same talent. This effect is due to the underconfidence of talented girls with good marks in both subject fields: In math, they do not trust their praise and invest less than boys with the same grades.

[25] To my knowledge, there is so far no economic paper on favoritism in schools. For a paper on favoritism in organizations in general, see Prendergast and Topel (1996).

Consider next the average achievements of boys and girls who are lowly talented in math and sciences ($\theta_{iH} = k$). Lowly talented girls achieve on average

$$A_{Mk}^{G} = k\left[\frac{1}{2}\left[\sigma\left(2-\eta\right)\right]e_{l2} + \frac{1}{2}\left[\left(1-\sigma\right)+\sigma\eta\right]e_{h3} + \frac{1}{2}\left(1-\sigma\right)e_{h1}\right], \text{ with} \tag{17}$$

$$\eta \equiv \alpha - \frac{1}{2}\alpha^{2}.$$

The average achievement of lowly talented boys, however, amounts to

$$A_{Mk}^{B} = k\left[\sigma e_{l2} + \left(1-\sigma\right)e_{h1}\right]. \tag{18}$$

Define $\Delta_{Mk} \equiv A_{Mk}^{G} - A_{Mk}^{B}$. Then,

$$\Delta_{Mk} = \frac{\left(\frac{1}{2}-\sigma\right)^{2}\left(\alpha - \frac{1}{2}\alpha^{2}\right)k\left(1-k\right)}{1+\left(\alpha-\frac{1}{2}\alpha^{2}\right)} > 0. \tag{19}$$

From (20), one can see that lowly talented girls achieve *more* on average than lowly talented boys. The reason is that the former receive an undeserved good grade in math more often than the latter: The weakly attitude-sensitive teachers in math patronize girls.

In order to see how the patronizing of girls in math affects the total gender difference in average achievement, define $\Delta_{M} \equiv \Delta_{M1} + \Delta_{Mk}$. It is easy to see that

$$\Delta_{M} = -\frac{\left(\frac{1}{2}-\sigma\right)^{2}\left(\alpha - \frac{1}{2}\alpha^{2}\right)\left(1-k\right)^{2}}{1+\left(\alpha-\frac{1}{2}\alpha^{2}\right)} < 0. \tag{20}$$

This implies

Proposition 4 In the U.S. equilibrium, biased grading in mathematical classes at school has the effect that girls achieve less on average than boys in the math test: $\Delta_{M} < 0 \;\forall k < 1$.

Proof Proposition 4 follows directly from (21). □

The reason why being patronized in math turns out to be, altogether, a disadvantage for the girls is as follows: The excess effort of the lowly talented girls who are made overconfident by undeserved praise does not fully translate into a corresponding excess achievement, because the lack of talent leads to a waste of effort. However, the lack of effort of highly talented girls distrusting their good grades does fully translate into a corresponding lack of achievement. Thus, the negative effect prevails. Accordingly, the gender differences between the achievement distributions of boys and girls in math in the U.S. equilibrium are exactly as can be observed from the PISA data for half of the OECD countries, including the U.S.

7 Why the majority of highly educated women earns less than highly educated men

Since the mid-sixties, gender wage gaps have been reported for almost all countries within and outside the OECD.[26] Although these gaps are mostly narrowing in the industrialized countries, this process has been slowing down significantly in the U.S. in the 1990s.[27] For some high-income jobs, the gender pay gap in the U.S. has even been growing again from the end of the 1990s until 2004.[28] Therefore, the question why women earn less than men remains challenging.

In the rich countries, the two main reasons why women earn less than men are job segregation and the so-called "glass ceiling", an unexplained difficulty for women to become promoted to the top jobs within their firm.[29] But even when differences between women and men in job characteristics, endowment such as education and job experience, and marital status have been controlled for, a significant residual wage gap has been reported for all countries.[30] Thus, a woman with the same university degree and the same job experience as her male colleague often earns sigificantly less within the

[26] Compare Weichselbaumer and Winter-Ebmer (2005) for a meta-analysis of these studies.

[27] Compare Blau and Kahn (2006).

[28] See Yurtoglu and Zulehner (2006), p.14.

[29] Compare for example Yurtoglu and Zulehner (2006).

[30] The empirical standard method to estimate this residual gender wage gap is the Blinder-Oaxaca approach. Compare Blinder (1973) and Oaxaca (1973).

same occupation. This is also true for high-income jobs.[31]

It is plausible to conjecture that discrimination on the labor market is the unobserved source of the residual gender pay gap; and this is indeed often found. But obviously, one could equally well suspect hidden productivity differences to be the cause. Indeed, performance related pay has become more prevalent in the last years; and the increasing gender wage gap in top corporate jobs in the U.S. could also be a result of this and a gap in productivities.[32] If this interpretation is true, further institutions against discrimination on the labor market would not be sufficient to eliminate the gender wage gap.

In my model, biased grading at school leads to productivity differences between men and women that contribute to the residual gender wage gap: A woman with a master's degree might be less confident and therefore less productive in her job than her male colleague with the same degree and in the same job.

More specifically, compare women and men who study field j at the university in the U.S. equilibrium. During their studies, they invest effort according to equation (2). At the end of their university career, their talents become publicly observable, and they enter the labor market. The wage of such a former student i is his (her) human capital gained at university, i.e. his (her) actual achievement $\theta_{ij}e_{ij}$.

It is easy to see that no gender wage gap exists for graduates in math. Both male students and the few female students in math are highly confident because they received honest praise at school for their mathematical skills. They all earn

$$w_M\left(\theta_{iM}, e_{iM}\right) = \theta_{iM}e_{h1}, \tag{21}$$

so that their average wage, too, amounts to

$$\overline{w}_M\left(\theta_M\right) = \theta_M e_{h1}. \tag{22}$$

But it is only a minority of women - only 25 percent within the model, according to Theorem 2 - who catch up in income with their male colleagues and former fellow students. These women are those who got a bad mark in humanities but were praised in math by a trustworthy teacher and studied

[31] For example, Yurtoglu and Zulehner (2006) find a significant gender wage gap for top corporate jobs in the U.S.

[32] See Yurtoglu and Zulehner (2006) for such an argument on page 16.

math in consequence. The vast majority of women, however, namely 75 percent in the model, study humanities. In order to see why, on average, they earn less than their male fellow students with the same abilities, average wages of male and female graduates have to be compared.

Consider first the male students. Only those of them who got a bad grade in math and a good grade in humanities study the latter subject field at the university. They trust their grade in humanities, invest high effort and earn

$$w_H(\theta_{iH}, e_{iH}) = w_H(\theta_{iH}, e_{h1}) = \theta_{iH} e_{h1}. \tag{23}$$

Obviously, the average wage $\overline{w}_H^B(\theta_H)$ of male graduates in humanities who are of type $\theta_H = 1$ amounts to

$$\overline{w}_H^B(1) = e_{h1}, \tag{24}$$

while the average wage of their lowly talented male fellow students is

$$\overline{w}_H^B(k) = k e_{h1}. \tag{25}$$

Consider now the female students of humanities. Define $g^H(\theta_H, 1, 1)$ to be the number of them who are of type θ_H and have received the message $m = (1, 1)$ at school. They all study humanities, and each of them earns $w_H(\theta_{iH}, e_{h1}) = \theta_{iH} e_{h1}$, like the men. Define now $g^H(\theta_H, k, 1)$ to be the number of females of type θ_H who have received the message $m = (k, 1)$ from their teachers. They, too, study humanities, and earn the wage $w_H(\theta_{iH}, e_{h1}) = \theta_{iH} e_{h1}$. But the remaining females gifted in humanities who study this field are those who got the bad message $m = (k, k)$ from their teachers. Let $g^H(\theta_H, k, k)$ represent their number. They are not as confident as their fellow-students, invest less effort and, consequently, become less productive and earn a lower wage:

$$w_H^G(\theta_{iH}, e_{iH}) = w_H(\theta_{iH}, e_{l1}) = \theta_{iH} e_{l1}\,.$$

Let $\overline{w}_H^G(\theta_H)$ represent the average wage of females of type θ_H and with a degree in humanities. Then, the average wage $\overline{w}_H^G(1)$ of women who are highly talented in humanities and who studied this field amounts to

$$\begin{aligned} \overline{w}_H^G(1) &= e_{h1} - \gamma(e_{h1} - e_{l1}) < e_{h1}, \text{ with} \qquad (26) \\ \gamma &\equiv \frac{g^H(1,k,k)}{g^H(1,1,1) + g^H(1,k,1) + g^H(1,k,k)} \end{aligned}$$

and the average wage $\overline{w}_H^G(k)$ of women who are lowly talented in humanities and who studied it amounts to

$$\overline{w}_H^G(k) = k\left[e_{h1} - \gamma\left(e_{h1} - e_{l1}\right)\right] < ke_{h1}. \tag{27}$$

Thus, there will be a gender wage gap among graduates in humanities, even if one controls for talents. Put differently, this gender wage gap is due to productivity differences but not to innate differences in abilities. Besides, this result already shows that it is impossible to separate between productivity differences and discrimination: The productivity differences that are reflected in the gender wage gap of former students of humanities are due to discriminatory behaviour (toward both sexes) at school. The gender differences in earnings of graduates are summarized in

Proposition 5 (1) Each male graduate i invests maximum effort e_{h1} and achieves $\theta_{ij}e_{h1}$, when he enters the labor market. (2) The female math graduates, i.e. $\frac{1}{4}$ of all female graduates, do so, too. (3) But $\frac{3}{4}$ of all female graduates earn less on average than their male colleagues with the same education and abilities. This is because among graduates in humanities, there is a gender wage gap: $\overline{w}_H^G(1) - \overline{w}_H^B(1) = -\gamma\left(e_{h1} - e_{l1}\right)$, and $\overline{w}_H^G(k) - \overline{w}_H^B(k) = -k\gamma\left(e_{h1} - e_{l1}\right)$. (4) This gender wage gap widens with increasing income: $\left|\overline{w}_H^G(1) - \overline{w}_H^B(1)\right| > \left|\overline{w}_H^G(k) - \overline{w}_H^B(k)\right|$.

Proof Parts (1) and (2) of Proposition 5 follow from (22), part (3) follows from (27) and (28), and part (4) is directly implied by part (3). □

Of course, there are most probably many different reasons for the residual gender wage gap of the highly educated that is observed empirically. This paper offers one possible explanation that differs from the standard explanations. Instead of holding discrimination on the labor market responsible for the gender wage gap, it shows that biased grading in school toward both sexes suffices to produce such a wage gap. In my model, there is no explicit or direct discrimination against women; on the labor market, highly educated women are, on average, simply less productive than highly educated men. Nevertheless, this has nothing to do with differences in innate characteristics between the sexes. The reason for their lack of productivity, compared with male graduates, is twofold. First, too many women study humanities without being sufficiently talented. Second, too many talented women who study humanities are underconfident. Thus, they invest less effort than their male fellow students.

This is the reason why the gender wage gap of the highly educated increases with increasing income, when abilities and effort are complementary: The more able the employee, i.e. the higher his or her income, the more impact does a reduction of effort have. Thus, low confidence of women produces more inequality among the highly talented wage earners than among the lowly talented ones. Actually, there is evidence that the gender wage gap at the higher end of the income distribution does indeed increase with increasing income and abilities: As Yurtoglu and Zulehner (2006) note at page 14, the gender pay gap in top corporate jobs in the U.S. is larger in higher positions.

8 Conclusion and Discussion

In the current paper, I offer a unified explanation for a large number of persistent gender differences that are observed empirically: the complex gender differences in achievements in reading literacy and math, as measured by the PISA test scores, the gender differences in enrollment decisions at the university, and finally the gender differences in earnings of the highly educated. I found that biased grading at school, which occurs quite often according to empirical research on grading policies, is sufficient to shift the girls' achievement distribution in reading literacy to the right of the boys' achievement distribution, to lower the girls' average, bottom and top achievement in math, compared with the corresponding achievements of the boys, to make women shy away from math at the university and to make them crowd in university programs in humanities, and to produce a gender wage gap among university graduates.

Although biased grading is discrimination in some sense of the word, because students with a supposedly negative attitude are treated differently than those with a supposedly positive one, I did not incorporate any kind of direct discrimination against women (or men) into my model. In particular, teachers in my model do not exhibit any taste for discrimination against girls (or boys). On the contrary, the preferences of teachers are the same toward boys and girls. It is the combination of the teachers' preferences - they sometimes like to use grades as expressions of sympathy or dislike - and the different expectations of boys and girls with regard to the meanings of their grades that creates the special situation in which biased grading leads to the gender differences described.

The behaviour of teachers and students in the model could also be derived in a different way. Note that the teachers praise girls too much in math and too seldom in humanities, while they praise boys too much in humanities. This grading policy could also result from naive counteractive measures against the gender differences described. Some teachers who are do-gooders, working against gender stereotypes, would want girls to become more confident in math and less obsessed with humanities. In consequence, they might well be tempted to be a bit too strict with girls in humanities and a bit too enthusiastic with them in math, not taking into account that the girls might start to suspect them of doing so. Analogously, these teachers would want to induce more zeal for humanities in their male students. Thus, they would patronize them in the non-mathematical classes, again not taking into account that the boys would become suspicious. Such naive do-gooders among teachers would produce exactly the gender differences that they would want to abolish, and the results would be the same as in the actual model. Thus, there is room for further research, in particular empirical research, on biased grading and its effects on achievements, career decisions and earnings of males and females.

9 Appendix

Proof of Lemma 2 Consider first a teacher who is the first mover in the grading continuation game. I now prove part (1) of the lemma. Assume that $d_i(1, m_{rij'}) = d_i(k, m_{rij'})$. An upward bias of a teacher r of field j will occur only if $\widehat{a}_{ij} = 1$ but $s_{rij} = k$. This, in turn, happens if and only if $\Pr_r\{a_{ij} = 1 \mid s_{rij} = k,\ s_{r'ij'}\} > \frac{1}{2}$. Remember that

$$\begin{aligned}\Pr_r\{s_{rij} = 1 \mid a_i = 1, \theta_{ij} = 1\} &= \sigma_{rH} \text{ and}\\ \Pr_{r'}\{s_{r'ij'} = 1 \mid a_i = 1, \theta_{ij'} = 1\} &= \sigma_{r'H};\end{aligned}$$

and remember that

$$\begin{aligned}\Pr_r\{s_{rij} = k \mid a_i = 1, \theta_{ij} = k\} &= \sigma_{rL} \text{ and}\\ \Pr_{r'}\{s_{r'ij'} = k \mid a_i = 1, \theta_{ij} = k\} &= \sigma_{r'L}.\end{aligned}$$

Correspondingly, it holds that

$$\begin{aligned}\Pr_r\{s_{rij} = 1 \mid a_i = 0, \theta_{ij} = 1\} &= \sigma_{rL} \text{ and}\\ \Pr_{r'}\{s_{r'ij'} = 1 \mid a_i = 0, \theta_{ij} = 1\} &= \sigma_{r'L};\end{aligned}$$

and

$$\begin{aligned}\Pr_r\{s_{rij} = k \mid a_i = 0, \theta_{ij} = k\} &= \sigma_{rH} \text{ and}\\ \Pr_{r'}\{s_{r'ij'} = k \mid a_i = 0, \theta_{ij} = k\} &= \sigma_{r'H}.\end{aligned}$$

Let λ denote the probability that $\sigma_{rH} = \sigma_{r'H}$ and $\sigma_{rL} = \sigma_{r'L}$ at the information set of teacher r; i.e. the probability that the teachers in j and j' are either both weakly attitude-sensitive or both strongly attitude-sensitive, given the information of teacher r. Then, the posterior belief of teacher r about the student's attitude, after having observed the other teacher's signal, is either

$$\Pr_r\{a_i = 1 \mid s_{rij} = k,\ s_{r'ij'} = k\}\,, \text{ or}$$

$$\Pr_r\{a_i = 1 \mid s_{rij} = k,\ s_{r'ij'} = 1\} = \frac{\nu}{\nu + \delta}, \text{ with}$$

$$\begin{aligned}\nu &= \lambda\left[\left(1-\sigma_{rH}\right)\sigma_{rH}+\left(1-\sigma_{rH}\right)\left(1-\sigma_{rL}\right)+\sigma_{rL}\sigma_{rH}+\sigma_{rL}\left(1-\sigma_{rL}\right)\right]+\\ &\quad \left(1-\lambda\right)\left[\left(1-\sigma_{rH}\right)\sigma_{r'H}+\left(1-\sigma_{rH}\right)\left(1-\sigma_{r'L}\right)+\sigma_{rL}\sigma_{r'H}+\sigma_{rL}\left(1-\sigma_{r'L}\right)\right],\end{aligned}$$

$$\begin{aligned}\delta &= \lambda\left[\left(1-\sigma_{rH}\right)\sigma_{rH}+\left(1-\sigma_{rH}\right)\left(1-\sigma_{rL}\right)+\sigma_{rL}\sigma_{rH}+\sigma_{rL}\left(1-\sigma_{rL}\right)\right]+\\ &\quad \left(1-\lambda\right)\left[\left(1-\sigma_{r'H}\right)\sigma_{rH}+\left(1-\sigma_{r'H}\right)\left(1-\sigma_{rL}\right)+\sigma_{r'L}\sigma_{rH}+\sigma_{r'L}\left(1-\sigma_{rL}\right)\right].\end{aligned}$$

It always holds that $\Pr_r\left\{a_i = 1 \mid s_{rij} = k,\ s_{r'ij'} = k\right\} < \frac{1}{2}$, regardless of the types of the teachers. This implies that teacher r wants to bias his grade upwards only if the other teacher's signal has been positive, $s_{r'ij'} = 1$. Then, teacher r wants to send the upward-biased message $m_{rij} = 1$ if and only if $\frac{\nu}{\nu+\delta} > \frac{1}{2}$. It holds that if $\sigma_{rH} \neq \sigma_{r'H}$ and $\sigma_{rL} \neq \sigma_{r'L}$, then either $\sigma_{r'L} < \sigma_{rL} \wedge \sigma_{r'H} > \sigma_{rH}$ or $\sigma_{r'L} > \sigma_{rL} \wedge \sigma_{r'H} < \sigma_{rH}$. It follows that $\frac{\nu}{\nu+\delta} > \frac{1}{2}$ if and only if $\lambda < 1$ and $\sigma_{r'L} < \sigma_{rL} \wedge \sigma_{r'H} > \sigma_{rH}$. Consequently, teacher r sends the upward-biased message if and only if (i) there is a positive probability $(1-\lambda)$ that the other teacher is strongly attitude-sensitive and he himself is not, and (ii) he himself is weakly attitude-sensitive. Because teacher r has the first move in the grading continuation game, $(1-\lambda)$ reduces to $(1-\alpha) > 0$. Thus, teacher r will always send the upward-biased message if he himself is weakly attitude-sensitive. This proves part (1) of Lemma 2.

The proof of part (2) is skipped because it is completely analoguous to the one of part (1). The proof of part (3) follows directly from the proofs of parts (1) and (2). It remains to prove part (4).

The second mover in the grading continuation game, teacher r', observes what the first mover did and dates up his beliefs about the first mover's type and the attitude of the student. Assume first that $d_i\left(1, m_{r'ij'}\right) \neq d_i\left(k, m_{r'ij'}\right)$. Then, the first mover never distorts his grade, regardless of his type. Accordingly, the second mover does not learn anything new about the first mover's type. He is in the same situation as the first mover has been. Thus, he behaves exactly alike. Consider now the case where $d_i\left(1, m_{r'ij'}\right) = d_i\left(k, m_{r'ij'}\right)$. Assume that nevertheless, the first mover did not distort his grade. Then, the second mover r' knows that the first mover is strongly attitude-sensitive. Thus, he himself will distort his grade if and only if he is weakly attitude-sensitive, $s_{rij} \neq s_{r'ij'}$, and if $d_i\left(m_{rij}, 1\right) = d_i\left(m_{rij}, k\right)$. Thus, the second mover r' will again behave as if he was the first mover.

Assume now that the first mover distorts his grade. Then, the second mover knows that the first mover is weakly attitude-sensitive. Thus, $\lambda = 1$, and the second mover r' does not want to distort his grade. This proves part (4) of the lemma.

Proof of Theorem 1 Suppose the career decisions of the students to be as described in parts (1) and (2) of the theorem. Consider now grading in humanities. Lemma 2 implies that only a teacher who is weakly attitude-sensitive and who would be the first to distort a grade of student i in the grading continuation game ever sends a message to i that is not in line with his signal. Thus, it is sufficient to define the strategy of teacher $r = r_H$ when characterizing equilibrium grading in humanities. Consider first the grading of a boy i_B. When $s^B_{r'iM} = m^B_{r'iM} = 1$, a change in m^B_{riH} does not influence the career decision $D_B\left(1, m^B_{riH}\right)$ of i_B. Hence, $d_i\left(1, m_{r'ij'}\right) = d_i\left(k, m_{r'ij'}\right)$, and biased grading is costless for r_H. Besides, because r_H is weakly attitude-sensitive and the other teacher's signal has been positive, r_H's belief about the attitude of i_B is $\Pr_r\left\{a_i = 1 \mid s^B_{riH},\ s^B_{r'iM} = 1\right\} > \frac{1}{2}$ $\forall s^B_{riH}$. From this and the definition of teacher types and preferences, it follows that $m^B_{riH}\left(1, s^B_{riH}\right) = 1\ \forall s^B_{riH}$. However, when $s^B_{r'iM} = m^B_{r'iM} = k$, a change in m^B_{riH} does influence the career decision $D_B\left(1, m^B_{riH}\right)$ of i_B. Hence, $d_i\left(1, m_{r'ij'}\right) \neq d_i\left(k, m_{r'ij'}\right)$, and biased grading is costly for r_H. Because $c > \frac{1}{2}\left(1-k\right)$, as assumed in inequality (6), these costs of biased grading imply that r_H grades honestly: $m^B_{riH}\left(s^B_{r'iM}, s^B_{riH}\right) = s^B_{riH}\ \forall s^B_{r'iM}, s^B_{riH}$. This proves that given the career decisions of boys in reaction to possible grades in math and humanities, grading of boys in humanities is as described in part (3) of the Theorem. The proofs that parts (1) and (2) of the theorem imply parts (4), (5), and (6) are analoguous and shall therefore be skipped. It remains to prove that grading practices as described in parts (3)-(6) of the theorem imply the student's career decisions given in parts (1) and (2). This follows from Lemma 1, Theorem 2, and Theorem 3. Therefore, the proofs of Theorems 2-3 below will complete the proof of Theorem1.

Proof of Theorem 2 Define the tuple of signals received by the math teacher and the humanities teacher on the respective talents of girl i to be $s_i = \left(s_{iM}, s_{iH}\right)$, and let $m_i\left(s_{iM}, s_{iH}\right) = \left(m_{iM}, m_{iH}\right)$ represent the tuple of messages. It is straightforward to show that s_i takes the values $(1,1)$, $(1,k)$, $(k,1)$, or (k,k) with equal probability $\frac{1}{4}$. Thus, there are four grading continuation games for each combination of teachers' types. Each teacher is weakly attitude-sensitive with probability α. Suppose the career decisions of the girls to be as described in part (1) of Theorem 1. Then, Lemma 2 implies that the three grading continuation games following the signals $(1,1)$, $(1,k)$, and (k,k) have the outcome $m_i\left(s_{iM}, s_{iH}\right) = \left(s_{iM}, s_{iH}\right)$ (both teachers honest), regardless of the types of the teachers. With regard to the grading continuation game subsequent to the signals $(k,1)$, Lemma 2 implies that the outcome will be $m_i\left(s_{iM}, s_{iH}\right) = \left(s_{iM}, s_{iH}\right)$ (both teachers honest) if and only if both teachers are not weakly attitude-sensitive, i.e. with probability $(1-\alpha)^2$. If both teachers are weakly attitude-sensitive, the first mover will distort his message and the second mover will not. Thus, the outcome will be $m_i\left(k,1\right) = (1,1)$ or $m_i\left(k,1\right) = (k,k)$, with probability $\frac{1}{2}\alpha^2$ each. If exactly one of the teachers is weakly attitude-sensitive, he will distort his message, and the other teacher will not. Thus, if only the math teacher is weakly attitude-sensitive, the outcome of this continuation game will be $m_i\left(k,1\right) = (1,1)$. This happens with probability $\alpha\left(1-\alpha\right)$. If, however, the humanities teacher is weakly attitude-sensitive, the outcome will be $m_i\left(k,1\right) = (k,k)$. This, too, happens with probability $\alpha\left(1-\alpha\right)$. For the girls receiving their grades, all grading continuation games with the same outcome belong to one information set. Taking this and the prior α about the teachers' types into account and calculating the girls' posterior beliefs about their talents with the help of Bayes Rule yields the values given in parts (1)-(4) of Theorem 2. Part (5) follows directly from parts (1) and (4). (Besides, given the posterior beliefs of the girls, Lemma 1 implies that their career decisions are as reported in part (1) of Theorem 1.)

Proof of Theorem 3 Again, s_i takes the values $(1,1)$, $(1,k)$, $(k,1)$, or (k,k) with equal probability $\frac{1}{4}$, leading to four grading continuation games for each combination of teachers' types; and again, each teacher is weakly attitude-sensitive with probability α. Suppose the career decisions of the boys to be as described in part (2) of Theorem 1. Then, Lemma 2 implies that the three grading continuation games following the signals $(1,1)$, $(k,1)$, and (k,k) have the outcome $m_i\left(s_{iM}, s_{iH}\right) = \left(s_{iM}, s_{iH}\right)$ (both teachers honest), regardless of the types of the teachers. With regard to the grading continuation game subsequent to the signals $(1,k)$, Lemma 2 implies the following: The outcome will be $m_i\left(1,k\right) = (1,k)$ (both teachers honest) if and only if either both teachers are not weakly attitude-sensitive or solely the math teacher is weakly attitude-sensitive. Thus, the outcome will be honesty of both teachers with probability $\left[(1-\alpha)^2 + \alpha\left(1-\alpha\right)\right] = (1-\alpha)$. If both teachers are weakly attitude-sensitive, the humanities teacher will distort his message, and the math teacher will not. This also happens if solely the humanities teacher is weakly attitude-sensitive. Thus, the outcome of the continuation game will be $m_i\left(1, k\right) = \left(1, 1\right)$ with probability $\left[\alpha^2 + \alpha\left(1 - \alpha\right)\right] = \alpha$. For the boys receiving their grades, all grading continuation games with the same outcome belong to one information set. Taking this and the prior α about the teachers' types into account and calculating the boys' posterior beliefs about their talents with the help of Bayes Rule yields the values given in parts (1)-(4) of Theorem 3. Part (5) follows directly from part (1). (Besides, given the posterior beliefs of the boys, Lemma 1 implies that the career decisions of boys are as reported in part (2) of Theorem 1.)

10 References

1. Alvidrez, J. and Weinstein, R.S., Early Teacher Perceptions and Later Student Academic Achievement, Journal of Educational Psychology 91 (1999), 731-746.

2. Arrow, K. J., Some Mathematical Models of Race in the Labor Market, in: Racial Discrimination in Economic Life, ed. by A. Pascal, pp. 187–204. Lexington Books, Lexington MA (1972).

3. Arrow, K. J., The Theory of Discrimination, in: Discrimination in Labor Markets, ed. by O. Ashenfelter, and A. Rees, pp. 3–33. Princeton University Press, Princeton NJ (1973).

4. Becker, G., The Economics of Discrimination. Chicago (1957).

5. Benabou, R. and Laroque, G., Using Privileged Information to Manipulate Markets: Insiders, Gurus, and Credibility, The Quarterly Journal of Economics 107, 3 (1992), 921-958.

6. Benabou, R. and Tirole, J., Self-Confidence and Social Interaction, NBER Working Paper (2000).

7. Benabou, R. and Tirole, J., Self-Confidence and Personal Motivation, The Quarterly Journal of Economics 117, 3 (2002), 871-916.

8. Benabou, R. and Tirole, Intrinsic and Extrinsic Motivation, The Review of Economic Studies 70, 3 (2003), 489-520.

9. Benoit, J.-P., Color Blind Is Not Color Neutral: Testing Differences and Affirmative Action. The Journal of Law, Economics, and Organization 15, 2 (1999), 378-400.

10. Blau, F.D. and Kahn, L.M., The US Gender Pay Gap in the 1990s: Slowing Convergence, Working Paper 508, Princeton University (2006).

11. Blinder, A.S., Wage Discrimination: Reduced Form and Structural Estimates, Journal of Human Ressources 8, 4 (1973), 436-455.

12. Cliffort, M. M. and Walster, E., The Effect of Physical Attractiveness on Teacher Expectations, Sociology of Education 46, 2 (1973), 248-258.

13. Clifton, R. A. et al., Effects of Ethnicity and Sex on Teachers' Expectations of Junior High School Students, Sociology of Education 59, 1 (1986), 58-67.

14. Coate, S., and Loury, G. C., Will Affirmative-Action Policies Eliminate Negative Stereotypes?, American Economic Review, 83(5) (1993), 1220–1240.

15. Crawford, V. P. and Sobel, J., Strategic Information Transmission, Econometrica 50, 6 (1982), 1431-1451.

16. Dee, T. S., A Teacher Like Me: Does Race, Ethnicity, or Gender Matter? The American Economic Review 95, 2, Papers and Proceedings (2005), 158-165.

17. Fang, H. and Moscarini, G., Moral Hazard, Journal of Monetary Economics 52 (2005), 749–777.

18. Fryer, R. G., Loury, G. C., and Yuret, T., Color-Blind Affirmative Action. NBER Working Paper No. 10103 (2003).

19. Hoge, H.D. and Butcher, R., Analysis of Teacher Judgments of Pupil Achievement Levels, Journal of Educational Psychology 76 (1984), 777-781.

20. Jussim, L., Teacher Expectations: Self-Fulfilling Prophecies, Perceptual Biases, and Accuracy, Journal of Personality and Social Psychology 57 (1989), 469-480.

21. Jussim, L. and Harber, K. D., Teacher Expectations and Self-Fulfilling Prophecies: Knowns and Unknowns, Resolved and Unresolved Controversies. Personality and Social Psychology Review 9, 2 (2005), 131-155.

22. Lavy, V., Do Gender Stereotypes Reduce Girls' Human Capital Outcomes? Evidence from a Natural Experiment. NBER Working Paper 10678, (2004).

23. Lemke, M. et al., International Outcomes of Learning in Mathematics Literacy and Problem Solving: PISA 2003 Results from the U.S. Perspective. (NCES 2005-003). Washington DC: U.S. Department of Education, National Center for Education Statistics (2004).

24. Morris, S., Political Correctness, Journal of Political Economy 109, 21 (2001), 231-265.

25. Mullis, I. V. S. et al., Gender Differences in Achievement: IEA's Third International Mathematics and Science Study (TIMSS), Chestnut Hill, MA: International Association for the Evaluation of Educational Achievement, TIMSS International Study Center, Boston College (2000).

26. Oaxaca, R., Male-Female Wage Differentials in Urban Labor Markets, International Economic Review 14, 3 (1973), 693-709.

27. OECD (2001), Knowledge and Skills for Life. First Results from the OECD Programme for International Student Assessment (PISA) 2000, Paris.

28. Ouazad, A., Assessed by a Teacher Like Me: Race, Gender and Subjective Evaluations. INSEAD Working Paper No. 2008/57/EPS.

29. Phelps, E. S., The Statistical Theory of Racism and Sexism, American Economic Review, 62 (1972), 659–61.

30. Prendergast, C. and Topel, R. H., Favoritism in Organizations, The Journal of Political Economy 104, 5 (1996), 958-978.

31. Rosenthal, R. and Jacobson, L., Pygmalion in the Classroom: Teacher Expectation and Pupils' Intellectual Development. New York (1968).

32. Sobel, J., A Theory of Credibility, The Review of Economic Studies 52, 4 (1985), 557-573.

33. Trouilloud, D.O. et al., The Influence of Teacher Expectations on Student Achievement in Physical Education Classes: Pygmalion Revisited, European Journal of Social Psychology 32 (2002), 591-607.

34. Weichselbaumer, R. and Winter-Ebmer, D.: A Meta-Analysis of the International Gender Wage Gap, Journal of Economic Surveys 19, 3 (2005), 479-511(33).

35. Yurtoglu, B.B. and Zulehner, C., The Gender Gap in Top Corporate Jobs is still there, unpublished manuscript, (2006).