Economics of Education Review 36 (2013) 12–25

Contents lists available at SciVerse ScienceDirect

Economics of Education Review

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

Educational evaluation schemes and gender gaps in student achievement☆

Torberg Falch [a,b,*], Linn Renée Naper [c,d]

[a] Department of Economics, Norwegian University of Science and Technology (NTNU), Dragvoll, N-7491 Trondheim, Norway
[b] CESifo, Germany
[c] EC Group AS, Beddingen 8, N-7014 Trondheim, Norway
[d] Centre for Economic Research at NTNU, Norway

ARTICLE INFO

Article history:
Received 14 December 2011
Received in revised form 14 May 2013
Accepted 22 May 2013

JEL classification:
I21
J16

Keywords:
Educational evaluation schemes
Teacher grading
Gender gaps
Gender interactions

ABSTRACT

This paper investigates whether gender gaps in student achievement are related to evaluation schemes. We exploit different evaluations at the end of compulsory education in Norway in a difference-in-differences framework. Compared to the results at anonymously evaluated central exit exams, girls get significantly higher grades than boys when the same skills are assessed by their teacher. This gender grading gap in favor of the girls is found in both languages and mathematics. We find no evidence that the competitiveness of the environment can explain why boys do relatively better on the exam. We find some evidence that the gender grading gap is related to teacher characteristics, which indicates that the teacher–student interaction during coursework favors girls in the teacher grading.

1. Introduction

Gender gaps in student test scores are observed throughout the world, most notably in favor of girls in languages (Machin & Pekkarinen, 2008), while the results in mathematics are more mixed (Guiso, Monte, Sapienza, & Zingales, 2008; Hyde, Lindberg, Linn, Ellis, & Williams, 2008). In addition, girls have recently improved their position relative to boys (Hyde et al., 2008; Machin & McNally, 2005). Since literacy and numeracy skills are important determinants of success later in life, e.g., Murnane, Willett, and Levy (1995), Leuven, Oosterbeek, and Van Ophem (2004), and Heckman, Stixrud, and Urzua (2006), the gender achievement gaps might have important economic implications.

In this paper we analyze whether the observed gender gaps in student achievement are related to evaluation schemes by exploiting achievement scores for Norwegian students at the end of compulsory schooling. We find that girls get better grades than boys when assessed by their teacher compared to results at anonymously evaluated central exit exams. We investigate whether this gender grading gap in favor of girls is related to different competitiveness of the environment at the two evaluations and whether it is related to teacher characteristics.

Mechtenberg (2009) presents a game theoretical model in which different kinds of gender gaps are a result of teacher and student behavior in school. In equilibrium, the

☆ Linn Renée Naper thanks the Ministry of Local Government and Regional Development for financial support. Comments from two anonymous referees, Hans Bonesrønning, Lars-Erik Borge, Julie Cullen, Astrid Kunze, and Sandra McNally are gratefully acknowledged. The authors bear the full responsibility for the analysis and the conclusions that are drawn.

* Corresponding author at: Department of Economics, Norwegian University of Science and Technology (NTNU), Dragvoll, N-7491 Trondheim, Norway. Tel.: +47 7359 6757.
E-mail addresses: Torberg.Falch@svt.ntnu.no (T. Falch), linnreneenaper@gmail.com (L.R. Naper).

0272-7757/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.econedurev.2013.05.002

gender gaps are similar to observed gender differences in school achievement, university enrollment, and wages. In her model, there are two subjects at school – mathematics and humanities – and students' beliefs about own abilities depend on teacher grading. The crucial assumption for the equilibrium is that girls do not fully trust bad grades in humanities and good grades in mathematics, while boys do not fully trust good grades in humanities. Teachers respond to these beliefs by easy grading of boys in humanities and of girls in mathematics, and hard grading of girls in humanities. Thus, the central theorem in Mechtenberg (2009) is the existence of a significant gender grading bias against girls in humanities and a smaller gender grading bias against boys in mathematics.

The observed gender gap in student achievement in favor of girls is often explained by increased share of female teachers. For example Dee (2005, 2007) and Ammermueller and Dolton (2006) find evidence that students profit from having a same-sex teacher. Steel (1997) discusses a phenomenon referred to as "stereotype threats" as an explanation of how demographic matches between students and teachers may influence educational outcomes. The idea is that students' academic self-confidence, and therefore their performance, is limited by possible and perceived stereotypes in the classroom. Another potential explanation, often referred to as "role-model" effects, is that the presence of a demographically similar teacher may raise students' academic motivation and expectations, and thus positively affects performance.

Both stereotype threats and role-model effects are "passive" teacher effects in that they are not related to intentional behavior of teachers. Thus, passive teacher effects cannot explain systematic differences in performance across evaluation schemes as far as they test the same skills.

The hypothesis in Lavy (2008) is that schools and teachers are sources of stereotypes that harm girls. The hypothesis is tested by exploiting that the matriculation exam in the academic track at Israeli high schools consists of both a state exam, which is anonymously graded, and an internal school exam. Contrary to the hypothesis, Lavy (2008) finds that the bias on the non-blind test is in favor of girls in all subjects.

Compared to the exam system in Israel, the potential for discrimination is higher in countries where teacher grading is based on more than a single test. In a review of the literature on gender differences in economic experiments, Croson and Gneezy (2009) argue that women's behavior is more context-dependent than men's behavior. If the way people treat others depends on their gender, the teacher–student interaction in coursework might induce statistical discrimination. The findings of Emanuelsson and Fischbein (1986), Stobart, Elwood, and Quinlan (1992), Lindahl (2007a), and Bonesrønning (2008) indicate, however, that placing greater weight on coursework elements in the evaluation improves the relative performance of girls. Machin and McNally (2005) present similar evidence. They show that when the importance of coursework in the examination system in the UK increased in 1988, the girls started to outperform the boys in the assessments.

In the Norwegian case, teacher set grades are based on written tests throughout the school year, and all students conduct a written central exit examination which evaluates the same skills and are graded anonymously. The students are randomly selected to an exit examination in either mathematics, English, or Norwegian language. All grades matter for admission to upper secondary schools and they are in this respect high-stake tests. We find that girls obtain better scores than boys in teacher grading relative to the central exit exam in all subjects in the empirical period 2002–2005. Thus, our results are not in accordance with Mechtenberg's (2009) central theorem. The gender grading gaps estimated are, however, similar to those found by Lavy (2008), Bonesrønning (2008), and Lindahl (2007a).

We investigate whether the finding in Gneezy, Niederle, and Rustichini (2003) that males perform relatively better in competitive environments can explain the estimated gender grading gaps. We exploit the variation across counties in the extent to which grades matter for admission to upper secondary schools. We also exploit the fact that one cohort conducted a separate low-stakes test. The results indicate that the competitiveness of the environment cannot explain the gaps. In addition, the results for the low-stakes test indicate that the gaps are not simply related to the anonymous vs. non-anonymous dimension. However, we find some evidence that the gender of the teacher and teacher experience matter for the gender grading gaps.

The next section offers a more detailed description of the Norwegian educational system and student evaluation schemes. Section 3 presents the data. Section 4 includes the main results on the gender grading gap in teacher assessments, while Section 5 investigates some possible explanations of the observed gender gap. Section 6 discusses the results and concludes.

2. Institutional setting

Norway has 10 years of compulsory schooling (from the year children turn six to the year they turn 16). None repeat grades, which implies that every student graduates on-time after 10 years. Multi-purpose municipalities are responsible for the schools and assign students to schools according to neighborhood rules. In 2005, 1164 public schools provided education at the lower secondary level (8–10th grade).

At the end of lower secondary education, students are evaluated both non-anonymously by their teachers (grades given in all curricula-based subjects) and anonymously in central exit exams. Each student takes one central written exit exam of 5 h, which take place at the end of the final year. The Norwegian Directorate for Education and Training prepares the written central exams, while local authorities are responsible for a random assignment of examination subjects to schools and individual students. The Directorate determines the share of students in each examination subject. The schools and the teachers have no influence in the assignment of examination subject. The students, as well as the schools, are informed about their exam subject on the same day all

over the country, and the exam is 2–7 days later depending on exam subject. About 20 percent of the students are examined in Norwegian, about 40 percent are examined in mathematics, and about 40 percent in English.[1] The exam result is determined by two external examiners assigned to each student.

Teacher grading is the responsibility of individual teachers. In Norwegian language and English, there are separate marks for written and oral skills, where the former is based on written tests and the latter on performance in class. We compare the grade on written skills with the central exit exam results because they shall measure the same skills. According to the school law, the teacher grades on written skills must reflect the students' competence and skill, and not student effort such as, e.g., punctuality in turning in assignments. It is the case also in mathematics that the grade must only reflect the performance on written tests. Performance in class, which to some extent reflects student effort, is not a part of the foundation of these grades. Teachers typically use questions from former central exit exams in their tests. Most important for their evaluation are one-day tests structured to be identical to the exam and taking the same length of time (5 h). Although the performance throughout the whole school year matters, the performance in the latest one-day test is given the highest weight. This test is typically conducted 3–4 weeks before the central exit exam.[2] In exam subjects, teacher grades should be given at least one day before the notification of exam results, a rule that is followed without exception.

Overall, the relevant teacher grading and the central exit exams are based on the same curricula and should evaluate the same type of skills. The main difference is that the teacher grade reflects performance somewhat earlier in time than the exam. Only the very last one-day test in class is based on the full curriculum that is tested in the central exit exam. The timing difference is larger than in Lavy's (2008) study, in which the two tests are spaced only 1–3 weeks apart.[3]

Teacher grades and central exit exam results are equally important for students' final grade point average (GPA). GPA matters for the prospect of admission to upper secondary study tracks and schools.[4] There is a legal right to upper secondary schooling. Over 95 percent of the cohort enrolls the year they finish compulsory education. Upper secondary education is the responsibility of the counties, which determine location of schools and the composition of study tracks at each school. About 45 percent enroll in the academic study track that qualifies for higher education. In addition, during the empirical period of this paper, there were 12 vocational study tracks, which at graduation certify for work as an electrician, carpenter, practical nurse, etc. Most schools have several study tracks.

In their application for upper secondary education, students have to rank three different study tracks. They have a legal right to be enrolled into one of these three tracks, but whether they are enrolled in the first, second, or third preferred track depends on GPA. No other factors matter. Teacher grades and the result on the exit exam is high-stakes in this respect. In addition, some counties have free school choice. Students have to rank schools in addition to study tracks in their application, and admission to oversubscribed schools is solely based on GPA. Other counties rely on school catchment areas; the students are enrolled in the closest school with the preferred study track.[5] Thus, GPA is high-stakes to a larger degree in counties with free school choice than in counties using well-defined school catchment areas, a feature that we exploit in the analysis below.[6]

A national student evaluation scheme was implemented in the spring of 2004.[7] All students in the final grade had to take tests in all three exam subjects. The tests were designed to evaluate and monitor performance and to provide feedback to municipalities, schools, and teachers. The tests were evaluated by the student's teacher, but the teachers were not allowed to take the results into consideration when they decided on final grades. The tests had thus no consequences for the students.[8] The

[1] Students with the exam in Norwegian language had the exam over two days in the empirical period. There are two formal written Norwegian languages, one "main" Norwegian language and one "second-choice" form of Norwegian language. The students have exams in both and receive separate grades from their teachers in both. Because almost 10 percent of the students are exempted from the second-choice form, including a majority of the immigrants, we only consider the results for the main Norwegian language in this paper.

[2] The Norwegian Directorate for Education and Training has guidelines for the determination of teacher set grades. They are on purpose imprecise because the teachers can take individual circumstances into account. The main instruction is that the grades shall provide information about the competence of the student at the end of the course. This implies that the latest test is of highest importance, which clearly is in accordance with the casual evidence.

[3] Lindahl (2007a,b) also uses teacher set grades when estimating gender grading gaps. While we compare the teacher set grade with a separate grade from the central exit exam, Lindahl compares the teacher set grades with the result on a national test that is not reported on the students' diploma. The result on this test is, however, important for teachers when they set the final grade. Our analysis below using a Norwegian national test is closely related to Lindahl's studies.

[4] The students' diploma consists of 12 teacher set grades and the grade from the written central exit examination. In addition, about 2/3 of the students have one oral examination. The average of these 13 (14) grades is used to rank students for admission to upper secondary education.

[5] A closer description of one system of free school choice is given in Machin and Salvanes (2010). They study the effect on house prices of increased school choice from 1997 in the Oslo county.

[6] Note that the legal right is related to enrollment in study tracks, not in specific schools. In a system without school choice, the catchment area will in reality be different for different study tracks within the same school. A student with a low GPA faces the risk of not being offered a study place in the preferred study track at the nearest school. Then the student will have the opportunity to travel to another school with this specific study track or to enroll in another study track. This is very different from free school choice, because students with low GPAs have no control over the choice of school in which they will be offered a study place or whether the offer will be at a school-study track with few applicants.

[7] The tests were conducted in February–March, while the central exit exam was conducted in late May. Thus, there is also a relatively long spacing between these evaluations.

[8] The official guidelines related to this test clearly stated that the test results should not be taken into account by the teachers when setting final grades. Because this was public knowledge, it is hard to imagine that some teachers told their students the opposite. No evidence exists, however, on the actual behavior of the teachers.

Table 1
A classification of evaluation schemes.

	High-stakes	Low-stakes
Anonymous one-day test	**Central exit exam**	(Monitoring)
Non-anonymous one-day test	(Part of matriculation)	**National test**
Non-anonymous assessment over time	**Teacher grading**	(Make diagnoses)

testing time was short, about 1 h, and the content of the test could differ from the skills tested in the high-stakes assessments. In particular, the test in Norwegian did not include writing an essay, but was solely a test of reading skills.

According to Borghans, Meijers, and ter Wel (2006), individual effort and achievement depend on the reward related to the result. Student evaluation schemes in general differ along three dimensions. They can be anonymous or non-anonymous; they can be based on a single test or on performance over a longer period; and they can influence individual students' prospects (high-stakes tests) or not (low-stakes tests).[9] Table 1 classifies possible evaluation schemes into six different types.[10] In this paper, we exploit three of the evaluation types, as indicated in bold in the table. Possible tests of other kinds are indicated in the table.

3. Data and descriptive statistics

Information on students and teachers in lower secondary schools is provided by Statistics Norway. Data on teacher grades and results from the central exit exam are available for the cohorts graduating in the spring in 2002–2005, while results for the national test are available only for 2004. The data are merged with extensive information on individual student background, such as gender, immigration status, and the income, marital status and education of parents. Information on teachers includes gender, teaching experience, marital status, and number of children. The teacher information is aggregated to the school level and merged with student level data using a school identifier.

There are several mixed schools with students in 1–10th grade that typically are small and located in rural areas. Because information is not available on whether teachers work at the primary or lower secondary level, we exclude mixed schools from the sample to avoid linking primary school teachers to students at the lower secondary level. This reduces the sample by 24 percent.[11] Since our identification is based on within-student variation in achievement, the estimation sample includes only students with both a teacher grade and a central exit exam result in a given subject.[12] Each student is observed only in one subject because each student has only one central exit exam.

The grading scale is from one to six, where score six is best and one is very weak. Fig. 1 presents the distribution of grades across assessment schemes, subjects, and gender. A score of three or four is most common, each including 20–40 percent of the students in the different groups. The distributions are close to normal, although there are two distinct patterns. First, the scores are better in teacher grading than on the exam. Several students get a lower grade on the exam than when they are assessed by their teacher. Second, female students perform better than male students in languages, and in particular in Norwegian.

Tables 2 and 3 cross-tabulate the percentages with the different combinations of scores on the exam and the

Table 2
Teacher grades and central exit exam results in mathematics, female students.

Teacher assessment	Exam result						
	1	2	3	4	5	6	Sum
1	1.0	0.7	0.0	0.0	0.0	0.0	1.7
2	2.3	14.6	2.1	0.1	0.0	0.0	19.1
3	0.2	9.3	16.6	3.0	0.0	0.0	29.1
4	0.0	0.8	10.0	16.3	2.0	0.0	29.1
5	0.0	0.0	0.7	7.7	9.4	0.7	18.5
6	0.0	0.0	0.0	0.1	1.5	1.0	2.6
Sum	3.5	25.4	29.4	27.2	13.0	1.7	100.0

Table 3
Teacher grades and central exit exam results in mathematics, male students.

Teacher assessment	Exam result						
	1	2	3	4	5	6	Sum
1	1.7	0.9	0.0	0.0	0.0	0.0	2.5
2	3.1	16.6	2.6	0.1	0.0	0.0	22.5
3	0.1	9.0	16.8	3.6	0.0	0.0	29.5
4	0.0	0.6	8.1	15.5	2.3	0.0	26.6
5	0.0	0.0	0.5	6.1	8.9	0.7	16.2
6	0.0	0.0	0.0	0.1	1.5	1.1	2.7
Sum	4.9	27.2	28.0	25.4	12.7	1.8	100.0

[9] Low-stakes tests include several instruments to monitor school, school district, or country performance. International comparative achievement tests (like PISA and TIMMS) that are widely used in empirical work are low-stakes tests by nature. Grades in US high schools are only one criterion for college admission, since many colleges also rely on the SAT test. In contrast, in most European countries, admission to higher education institutions is based on grades set by teachers. Test results may also involve economic incentives for the schools and school owners and in some sense be high-stake tests for the schools, while, at the same time, low-stakes tests for the students. Evaluations of the reliability of tests used in accountability systems include Kane and Staiger (2002), Jacob and Levitt (2003), and Jacob (2007).

[10] Our classification into three dimensions indicates that there are eight different types of evaluation schemes. However, it is hard to imagine anonymous evaluations based on observations over a longer time period.

[11] In models that do not include information on teachers, the results for the main parameter of interest are very similar in the full sample and in our regression sample. The estimate on the full sample is 1–12 percent larger (depending on subject) than the results for our regression sample reported below.

[12] Some students are exempted from the central exit exam because of illness on examination day, disabilities, etc. The written exam absence rate is close to three percent each year.

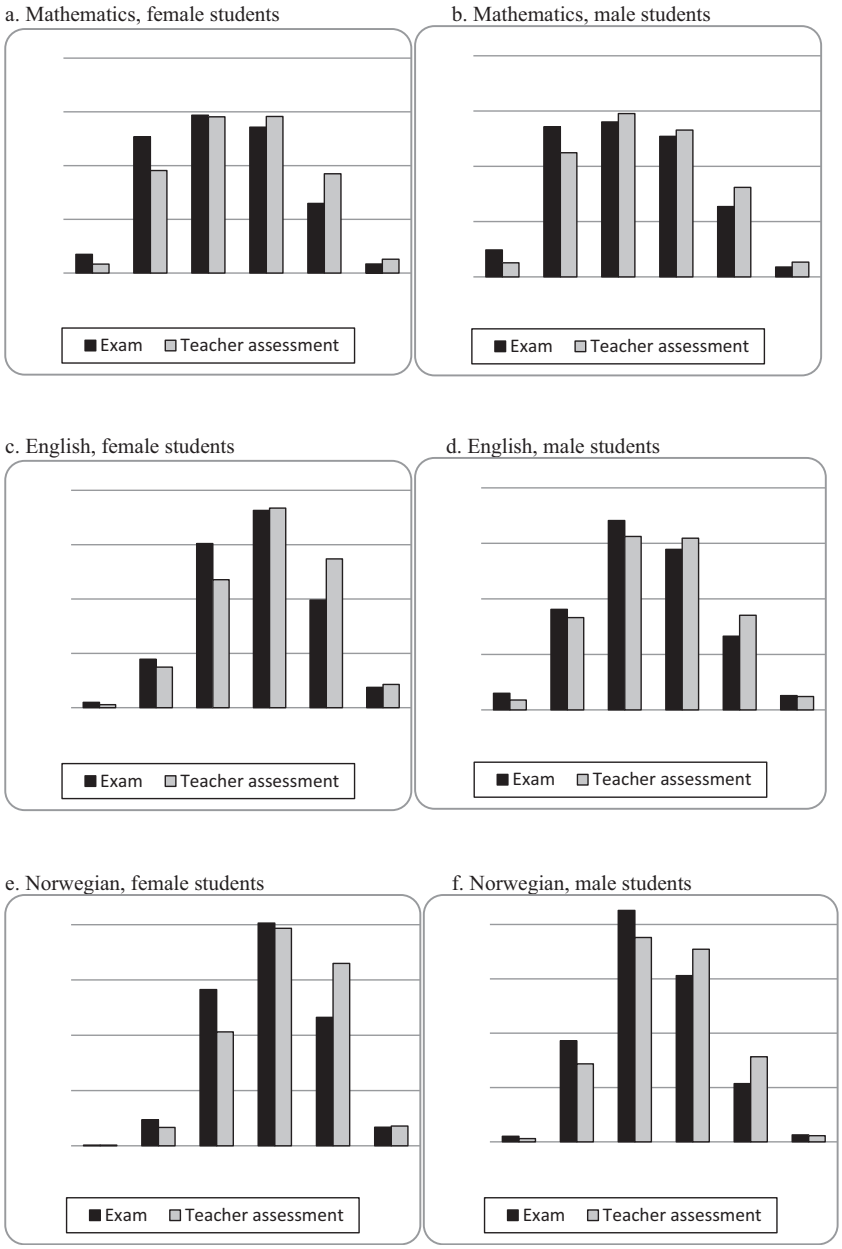

Fig. 1. Percentage distribution of scores across gender, subject, and evaluation scheme.

teacher grading in mathematics for female and male students, respectively. It is most common to get the same score in both evaluations. However, several students get one grade lower on the exam than in the teacher assessment. For example, out of the 29.1 percent of girls with teacher grade equal to 4, 34.4 percent got score 3 on the exam (that is 10.0 percent of the total sample of girls). The figures are similar for boys, but with the tendency that fewer students get a lower score on the exam. Overall, 58.9 percent of the girls and 60.6 percent of the boys get the same result in the two evaluation schemes, and 32.6 and 29.2 percent, respectively, get a lower score on the exam than in the teacher assessment.

Table 4 compares mean scores in teacher grading and central exit exams across gender. For each subject, the table reports average teacher grades, exam results, and the test-statistic from a mean comparison test across gender and evaluation schemes. The average score is higher for female students than for male students in all cases, and the differences are statistically significant. The gender gap is largest in Norwegian and smallest in mathematics. In contrast to most countries, girls outperform boys even in mathematics. This is in line with the findings in, for example, the international comparative student test of eight graders TIMSS 2003, see for example Fryer and Levitt (2010). Guiso et al. (2008) argue that the gender

Table 4
Mean comparison tests by gender and evaluation scheme, 2002–2005.

	Mathematics			English			Norwegian		
	Teacher assessment	Exam result	Difference	Teacher assessment	Exam result	Difference	Teacher assessment	Exam result	Difference
All	3.45	3.22	0.23	3.74	3.57	0.17	3.82	3.62	0.20
	[1.14]	[1.15]	(33.4)	[1.07]	[1.08]	(25.0)	[0.98]	[0.98]	(25.1)
Females	3.51	3.26	0.25	3.96	3.76	0.20	4.13	3.92	0.21
	[1.12]	[1.13]	(26.6)	[1.01]	[1.02]	(22.2)	[0.89]	[0.92]	(19.9)
Males	3.39	3.19	0.20	3.52	3.39	0.13	3.52	3.33	0.19
	[1.15]	[1.16]	(20.9)	[1.08]	[1.09]	(14.2)	[0.96]	[0.96]	(17.4)
Difference	0.12	0.07	0.05	0.44	0.37	0.07	0.61	0.58	0.02
	(12.4)	(6.70)	(10.26)	(48.4)	(40.9)	(10.71)	(57.4)	(55.1)	(2.38)

Note. Standard deviations in brackets and *t*-values in parentheses.

achievement gap in mathematics is related to gender equality in general, and that, in the most gender-equal societies, girls perform at least as well as boys.

Table 4 also shows that the average scores in teacher grading are higher than the exam scores. The last column for each subject shows that the differences are of about the same size in all subjects and statistically significant. In addition, the score differences between the assessment schemes are higher for girls than for boys. The simple difference-in-differences estimator is equal to 0.05, 0.07, and 0.02 in mathematics, English, and Norwegian, respectively, and significant at five percent level in all cases.[13]

Appendix Table A1 reports descriptive statistics. The first column includes all students. Regarding student characteristics, 69 percent are living with both parents, and about 30 percent of the parents have some college or university education. Teacher characteristics are only available as year specific averages at the school level. 54 percent of the teachers are women at the lower secondary level,[14] 64 percent are married, and 18 percent do not have children.

In the last three columns in Appendix Table A1, only the individuals that took the relevant central exit exam are included. Overall, there are very small differences in background characteristics across the subjects, even though in some small schools all students have the same exam subject. This clearly supports that allocation of exam subject across students is random.

4. Gender gap in teacher grading

4.1. Empirical strategy

We follow Lavy (2008) and estimate the following linear difference-in-differences model.

$$A_{Eijt} = \alpha + \lambda G_{ijt} + \delta E_{ijt} + \gamma(E_{ijt} \times G_{ijt}) + \beta X_{ijt} + \phi_j + \mu_t + \sigma_{Eijt}, \quad (1)$$

where the score A_{Eijt} at evaluation E ($E=1$ for teacher grading and $E=0$ for central exit exam) of student i at school j at time t is assumed to be a function of gender G ($G=1$ for females and $G=0$ for males) and the type of evaluation E. Each student is observed at one point in time, at the end of 10th grade. The model includes co-variates X_{ijt} (as reported in Appendix Table A1), and school and time fixed effects, ϕ_j and μ_t, respectively. σ_{Eijt} is an i.i.d. error term. Because the data set is stacked, including both the teacher grade and the exam result, the number of observations in the regression will be twice the number of students. We estimate the model separately for each subject, and, in addition, we estimate a model including all subjects. The latter will return the average effects across the three different subjects.

The difference-in-differences parameter γ identifies the mean gender difference in score gaps. A positive γ indicates that female students, conditional on the individual exam result, receive higher grades from their teachers than male students. The parameters λ and δ identify the gender achievement gap on the exam and the "grade inflation" for male students in the teacher grading, respectively.

In this model, all individual and school fixed effects are implicitly assumed away with regard to the parameter γ, as long as these effects are homogenous across evaluation schemes. In essence, γ is identified on the difference between the teacher grade and the exam result. Estimating γ from (1) is algebraically identical to estimating γ from the equation

$$A_{E=1,ijt} - A_{E=0,ijt} = \Delta A_{ijt} = \delta + \gamma G_{ijt} + \Delta\sigma_{ijt} \quad (2)$$

Eq. (2) highlights that including co-variates and time fixed effects in Eq. (1) does not influence the estimate of γ because the basic specification saturates all these effects. However, one advantage in estimating (1) is that more coefficients are revealed.

Consistency of the difference-in-differences parameter γ requires that assignment of female students to schools is not systematically related to teacher grading practices. Systematic assignment of students with respect to gender is very unlikely in the Norwegian system with fixed school catchment areas. However, we will take into account that schools may be heterogeneous with respect to teacher grading practices, peer effects due to different student

[13] These difference-in-differences parameters correspond to the parameter γ in Eq. (1) below.

[14] In contrast, at the primary level, there is clearly a majority of female teachers. For mixed schools (1–10th grade), there are 65 percent female teachers.

Table 5
Gender gap in teacher assessment. Dependent variable is student score.

	All subjects		Mathematics		English		Norwegian	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Female	0.302	0.302	0.059	0.060	0.375	0.375	0.600	0.600
	(38.3)	(38.4)	(5.61)	(5.70)	(33.2)	(33.0)	(46.5)	(46.6)
Teacher assessment	0.172	–	0.197	–	0.130	–	0.191	–
	(22.8)		(18.6)		(11.6)		(14.5)	
Female × (teacher assessment)	0.051	0.051	0.058	0.057	0.066	0.067	0.016	0.016
	(11.0)	(11.1)	(8.84)	(8.98)	(8.89)	(9.05)	(1.67)	(1.65)
Subject specific effects	Yes	Yes	–	–	–	–	–	–
(School fixed effects) × (teacher assessment)	No	Yes	No	Yes	No	Yes	No	Yes
Observations	260,928	260,928	103,090	103,090	100,528	100,528	58,208	58,208
Standard error of equation	0.977	0.975	1.023	1.020	0.967	0.964	0.854	0.850

Note: *t*-values in parentheses are heteroskedasticity robust and clustered at the school level. All models include year fixed effects, school fixed effects, student characteristics, and teacher characteristics. Full models for columns (1), (3), (5), and (7) are reported in Appendix Table A2.

composition, unobserved teacher quality, etc., by including school fixed effects interacted with the assessment scheme. This is identical to including school fixed effects in Eq. (2). When we estimate the model at the differenced form as in Eq. (2), we will also present results from model specifications including student and teacher characteristics.

4.2. Results

Table 5 presents results using the model specification described in Eq. (1). The first model includes all students and thus merges data across subjects, while the rest of the table presents results separately for the three subjects. The table only reports the parameters of main interest. Full results are provided in Appendix Table A2. The appendix table shows, as expected, that immigrants have lower scores than native students in all subjects, and that scores are highest for students with highly educated parents living together. The effects of teacher characteristics are imprecisely estimated, presumably because they are measured at the school level and the models include school fixed effects.

The results in Table 5 are very similar to the mean comparison tests in Table 4. The differences between the first columns for each subject and the results in Table 4 are related to some missing observations of student characteristics. The differences in standard errors are mainly related to clustering of errors at the school level. The mean gender achievement gap (as measured by the exam) is 0.30 score points on average across all subjects. The average grade inflation in teacher grading is 0.17 score points.

The effect of main interest, the interaction effect between the dummy variables for female student and teacher grading, is positive in all regressions. Female students are on average rewarded significantly better by their teachers, relative to the exam, than male students. The average gender grading gap across all subjects is 0.05 score points, and highly significant. The gap is largest in mathematics and English, and barely significant in Norwegian. The fact that the gap is not sensitive to the inclusion of interactions with school fixed effects indicates that teacher grading is not related to student or teacher sorting across schools.

The size of the interaction effects can be evaluated relative to the standard deviation of the distribution of the score difference between the assessment schemes.[15] The estimated effects in mathematics and English correspond to about 0.09 standard deviations, while in Norwegian the effect is about 0.02 standard deviations. The former effects are in line with Lavy's (2008) results, while the latter is smaller. Given the differences in assessment schemes analyzed in this paper compared to the schemes analyzed by Lavy (2008), we would expect that the gender gap would be larger in our case. While Lavy compares two single day tests, we compare the externally graded test with assessment based on performance over the whole school year, leaving more room for teacher–student interactions to have an impact.

These results are not consistent with the equilibrium conditions in Mechtenberg's (2009) model. We do not observe a grading gap against girls in humanities. We have two language subjects in our analysis, and the grading gap is against boys in both cases. The grading gap against boys in mathematics is in accordance with Mechtenberg's model, but the driving force in her model is that treatment and responses to treatment differ across subjects.

Interpreting our empirical results as a test of the Mechtenberg (2009) theorem is not straightforward because teacher grading practices influence student effort in her model. As students' expectations adjust to the grading signals, easy grading has a negative effect on achievement in equilibrium. This feature of the model is in line with the empirical evidence on easy grading, see for example, Figlio and Lucas (2004) and Bonesrønning (2004, 2008). Latent ability is not observed empirically. One would expect, however, that grading gaps in terms of central exit exams have the same sign as grading gaps in terms of latent ability since easy grading is expected to reduce the performance on the exam in all subjects.

[15] The mean score differences [standard deviation] between teacher grades and the central exit exam results are 0.197 [0.693] across all subjects, 0.228 [0.636] in mathematics, 0.162 [0.712] in English, and 0.195 [0.765] in Norwegian.

Table 6
School choice or not, descriptive statistics.

	Mathematics		English		Norwegian		Municipal population in 2004 (unweighted)	
	Teacher assessment	Exam result	Teacher assessment	Exam result	Teacher assessment	Exam result	Total	Share in urban areas
No school choice	3.47	3.21	3.74	3.56	3.83	3.60	10,643	0.55
	[1.13]	[1.13]	[1.07]	[1.07]	[0.98]	[0.96]	[20,643]	[0.22]
Free school choice	3.50	3.30	3.78	3.64	3.85	3.67	19,232	0.63
	[1.13]	[1.15]	[1.05]	[1.06]	[0.97]	[0.99]	[50,817]	[0.26]
No school choice, excluding small municipalities	3.47	3.20	3.75	3.60	3.83	3.61	34,473	0.78
	[1.14]	[1.13]	[1.07]	[1.08]	[0.97]	[0.95]	[44,216]	[0.14]
Free school choice, excluding small municipalities	3.53	3.36	3.83	3.71	3.88	3.71	54,519	0.89
	[1.13]	[1.15]	[1.04]	[1.06]	[0.96]	[0.97]	[90,602]	[0.09]
Bergen (no school choice)	3.57	3.28	3.92	3.82	3.97	3.72	237,430	0.97
	[1.14]	[1.15]	[1.04]	[1.03]	[0.93]	[0.92]	[–]	[–]
Oslo (free school choice)	3.66	3.46	3.86	3.72	3.90	3.73	521,886	0.99
	[1.10]	[1.16]	[1.00]	[1.00]	[0.93]	[0.93]	[–]	[–]

Note: Standard deviations in brackets. Small municipalities are defined as population below 15,000 in 2004.

Another feature of the Mechtenberg (2009) model is that boys outperform girls in mathematics when innate ability is independent of gender. This feature is not in accordance with our data. One can argue that the underlying reason why girls do relatively well in mathematics in Norway is differenced out of our model. But since the mechanisms might be different in countries with a different relative performance of girls, it would be interesting to see evidence from other countries.

5. What can explain the gender gap in teacher grading?

We consider two possible explanations of the observed grading gap against boys. First, there is arguably a more competitive environment at a central exit exam than at tests taken throughout the school year, and we will investigate whether this can explain why males do relatively better on the exam than in teacher grading. Second, even though the specific story of the teacher–student interaction of Mechtenberg (2009) is not supported by the data, the interaction may take other forms.

5.1. Gender grading gap and competitiveness of the environment

Gneezy et al. (2003) designed an experiment to investigate performance under different incentive schemes. Their findings suggest that women are less effective than men in competitive environments. Some evidence also exists from real life data. Paserman (2007) studies Grand Slam tennis tournaments and finds that women are significantly more likely than men to hit unforced errors at the crucial stages of the match. Örs et al. (in press) examine an entry exam to a very selective French business school, and find that males do relatively better than females on the exam compared to prior achievement.

We will investigate whether gender differences in response to competition can explain the observed gender grading gap in our data in two ways. Firstly, we will exploit the fact that GPA matters more in some counties than in other counties. Secondly, we compare the exam result to the one-day low-stakes national tests in 2004.

If the observed gender grading gap is due to a more competitive environment on the exam than at the events relevant for the teacher grade, we would expect the grading gap to be larger in counties with free school choice than in counties with only free choice related to study track, i.e., girls perform relatively worse on the exam under free school choice. According to the classification of Haraldsvik (2004), seven counties had free school choice in the empirical period (including Oslo), and 10 counties had fixed school catchment areas (including Bergen, the second largest city in the country), while two counties had free school choice in the cities and not outside the cities. The systems are represented all over the country. The casual evidence indicates that the variation in school choice across counties is mostly historical. Even though school choice has an ideological bias, there have been few changes over the last 20 years.[16] In 11 municipalities, mainly medium sized cities, there was a mixed system. We skip these observations in the analysis below (six percent of the observations). Appendix Table A1 shows that there was free school choice for 55 percent of the observations.

Table 6 presents descriptive statistics separately for municipalities with and without school choice. Students might respond to increased competition with increased effort throughout the year. Average teacher grades are higher in municipalities with free school choice, but the differences are very small in all subjects. For the exams, the difference is slightly larger. Notice that our difference-in-differences approach allows girls and boys to respond differently to increased competition with respect to effort during the school year.

One can argue that school choice is always limited in rural areas. Thus, Table 6 also presents statistics when municipalities with population below 15,000 are excluded (that is, 79 percent of the municipalities and 38 percent of

[16] Some changes have occurred. Oslo changed from a mixed system to a system of free school choice in 1997, see Machin and Salvanes (2010). In the city of Trondheim, the exact opposite change was implemented after the empirical period of this paper, while more choice has been introduced in Bergen. The arguments for changes are typically ideological, and are indeed not related to gender differences.

Table 7
Degree of school choice and gender gap in teacher assessment. Dependent variable is the *difference* in student score.

	All subjects				Mathematics		English		Norwegian	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Female	0.059	0.060	0.062	0.103	0.068	0.078	0.060	0.052	0.051	0.056
	(9.01)	(8.87)	(6.63)	(7.16)	(6.81)	(5.91)	(5.46)	(3.46)	(3.41)	(2.90)
Female × (free school choice)	−0.019	−0.020	−0.026	−0.068	−0.026	−0.050	0.009	0.017	−0.061	−0.061
	(2.08)	(2.15)	(2.14)	(2.68)	(2.04)	(3.06)	(0.58)	(0.85)	(3.12)	(2.39)
Sample	All	All	Restricted	Oslo and Bergen	All	Restricted	All	Restricted	All	Restricted
School fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Subject fixed effects	Yes	Yes	Yes	Yes	–	–	–	–	–	–
Year fixed effects	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Student and teacher char.	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	131,850	122,406	75,380	13,446	48,459	29,803	47,424	28,907	26,983	16,994
Standard error of equation	0.678	0.675	0.678	0.683	0.603	0.604	0.685	0.688	0.733	0.741

Note: The dependent variable is the difference between the teacher set grade and the result on the central exit exam. *t*-values in parentheses are heteroskedasticity robust and clustered at the school level. The restricted sample excludes municipalities with population below 15,000 in 2004.

the observations). Compared to the sample averages, the scores in the large municipalities are similar without school choice and slightly higher with school choice. The last two rows restrict the sample to the two largest cities: Oslo, with school choice, and Bergen, without school choice in the empirical period. The pattern is the same as for the whole country in mathematics, but Bergen performs well in the languages.

Finally, Table 6 presents average population sizes and urbanization rates. Municipalities with school choice are on average larger and have a higher degree of urbanization than municipalities without school choice. In the sample excluding small municipalities, this is mainly related to the fact that Oslo is the largest municipality in the country.

Table 7 presents models estimating different versions of Eq. (2) above. The model specification in column (1) is identical to the model specification in column (2) in Table 5, except that the model includes the interaction term between gender and free school choice.[17] Contrary to the hypothesis, the gender grading gap is smaller when there is free school choice. In the case of catchment areas, female students achieve on average 0.059 score points better than male students in teacher grading than on the exam, while with free school choice the gender grading gap is 0.040 score points. Female students thus perform relatively better on the exam in areas with school choice, and the difference is significant at five percent level. The model in column (2) in Table 7 includes time fixed effects and student and teacher characteristics, without affecting the estimated gender grading gaps.

Column (3) in Table 7 excludes observations in small municipalities. This does not change the results. Finally, in column (4) we restrict the sample to the two largest cities. The results indicate that the gender grading gap in Bergen, with well-defined catchment areas, is 0.10 score points in favor of girls, about twice the country average. But most interestingly, the grading gap is significantly smaller in Oslo in which there is free school choice. The grading gap in Oslo is estimated to be 0.035 score points, close to the average of counties with school choice.

The last columns in Table 7 show that the gender grading gap is related to school choice in mathematics and Norwegian, while the interaction term is small and insignificant in English. The subject specific regressions test whether there is a gender grading gap in six different cases: three subjects under two different degrees of the stakes. With choice only over study track, there is a grading gap against boys in all three subjects. With free choice both for study tracks and schools, there is a grading gap against boys in mathematics and English, but not in Norwegian.

Table 8 compares the exam results to the one-day low-stakes national tests.[18] It turns out that female students have a relatively lower score at the low-stakes tests than on the exams. This result is also contrary to the hypothesis that girls perform less well in competitive environments. The gender gap in absolute value is about twice the observed gender grading gap in Table 5. It turns out that the average gender gap is sizable particularly in Norwegian, and there is also a large gender gap in mathematics, which is clearly not in accordance with the competitiveness of the environment hypothesis. This result must, however, be interpreted with caution for several reasons. First, the low-stake test in Norwegian focused on reading and thus tested somewhat different skills than the central exit exam. The results for Norwegian can be interpreted as boys doing relatively better in reading than in writing. Second, these tests were shorter (1 h) than the exams (5 h), and boys and girls might perform differently on tests of different length. In addition, the space between the tests and the exams was relatively long (2–3 months).

Table 8 also includes models in which the female dummy variable is interacted with whether there is free school choice in the municipality. Since the difference in

[17] Notice that the level effect of free school choice is not identified since the model includes school fixed effects.

[18] The grading scale was different for the national test. In order to facilitate comparability, we impose the same mean and standard deviation for the national test as for the exam. Because data are only available for one year, teacher characteristics are collinear to the school fixed effects.

Table 8
Gender gap and high-stakes vs. low-stakes tests. Dependent variable is *difference* in student score.

	All subjects				Mathematics		English		Norwegian	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Female	−0.099	−0.104	−0.096	−0.104	−0.111	−0.101	0.032	0.019	−0.348	−0.341
	(7.65)	(7.88)	(5.18)	(3.91)	(8.69)	(3.89)	(1.80)	(0.46)	(15.7)	(8.53)
Female × (free school choice)	–	–	−0.020	−0.008	–	0.014	–	−0.014	–	−0.017
			(0.77)	(0.22)		(0.42)		(0.28)		(0.30)
Sample	All	All	All	Restricted	All	Restricted	All	Restricted	All	Restricted
School fixed effects	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Student characteristics	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	28,688	26,664	25,452	15,510	9370	5317	11,262	6624	6053	3585
Standard error of eq.	0.773	0.735	0.734	0.744	0.633	0.640	0.748	0.753	0.830	0.842

Note: The dependent variable is the difference between the test score on the national test and the result on the central exit exam. *t*-values in parentheses are heteroskedasticity robust and clustered at the school level. The restricted sample excludes municipalities with population below 15,000 in 2004.

stakes is arguable higher under free school choice, the hypothesis above implies that the interaction effects are positive. The results indicate relatively low power for this test. Lower power in the model for the national test than in the model for teacher grading is probably related to smaller sample. Nevertheless, the point estimates in columns (3) and (4) in Table 8 do not support the hypothesis. In addition, the interaction effects are clearly insignificant in all subject specific models.

Overall, conditional on the high-stakes central exit exams, boys outperform girls on the low-stakes national one-day tests. Thus, the hypothesis that girls perform relatively worse when stakes are high is not supported.[19]

These results also indicate that the anonymous vs. non-anonymous dimension of the evaluations schemes cannot alone explain the gender grading gap. Both teacher grading and the national tests are non-anonymous, but, while the female grading gap is positive for the former, it is negative for the latter.

One possible explanation of these findings is that it is the teachers and not the students that behave differently on different types of tests. If teachers grade differentially across girls and boys, they also might react to the importance of the situation. If teachers want to promote some group of students, grading biases would seem more likely to appear in high-stakes assessments than on low-stakes assessments. This is consistent with findings in experimental studies on grading gaps. Hinnerich, Höglin, and Johannesson (2011) conduct blind grading on a Swedish national test in high school that previously had been graded by the students' teacher, and find no evidence of discrimination. Van Ewijk (2011) finds similar results for students' ethnicity in Dutch primary schools, where teachers are asked to grade essays with manipulated names of the writers.

5.2. Gender grading gap and teacher–student interaction

Inspired by the literature on gender stereotypes and teacher–student gender interactions (e.g., Ammermueller & Dolton, 2006; Dee, 2007; Steel, 1997), we investigate whether the observed gender grading gap is related to the gender distribution of teachers. With "passive teacher effects", as described above, there will be no teacher–student gender interaction effects on grading. Hence, an interaction effect in this setup indicates that teachers adjust their grades, intentionally or not, depending on the student's gender. Since evaluation in languages to a larger extent involves subjective elements, it may be argued that it is more reasonable to expect a form of assessment discrimination in languages than in mathematics.

A teacher–student gender interaction in this setup may be interpreted as a kind of teacher-initiated discrimination in assessment of students. Since it is reasonable to believe that teaching practices vary with experience, we also investigate whether the grading gap is related to the experience of the teachers.

Teacher characteristics are measured at the school level. One would expect more noisy estimates when one uses teacher composition at school instead of matching teachers to students. On the other hand, we avoid biases related to strategic assignment of teachers to classes within schools.

Lavy (2008) discusses at some length whether the grading gap in the Israeli case is due to student or teacher behavior. This is hard to investigate, however, if the teacher–student interaction can be described as a principal–agent relationship, as in, for example, Mechtenberg (2009). Then students react to teacher strategies and teachers react to observed student behavior. Lavy (2008) finds that the observed gender gap is sensitive to teacher characteristics, in particular teacher gender, and accordingly interprets the observed differences as a result of teacher behavior. Both Lavy (2008) and Lindahl (2007b) find that the gender grading gap is highest with male teachers. Teachers tend to assess same sex students more strictly than opposite sex students. In an interesting study,

[19] Our results are quantitatively larger than Lindahl's (2007a) finding for Sweden. Lindahl compares high-stakes teacher grades based on whole year assessment with low-stakes national tests at the end of compulsory education. We have not estimated the same gender grading gap directly, but this gap follows by taking the difference between our estimated gaps in teacher grading (Table 5) and the national test (Table 8). In mathematics, our estimate that is comparable to Lindahl's estimate is 0.168 (0.057 to −0.111), which is 0.26 standard deviations of the score difference. Lindahl's estimate is equal to 0.11 standard deviations of the score difference.

Table 9
Gender gap and gender interaction effects in teacher assessment. Dependent variable is difference in student score.

	All subjects			Mathematics			English			Norwegian		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Female	0.059	0.100	0.139	0.061	0.157	0.209	0.039	0.111	0.088	0.117	−0.043	0.097
	(2.40)	(3.60)	(3.19)	(1.67)	(3.80)	(2.93)	(1.04)	(2.64)	(1.33)	(2.14)	(0.69)	(1.13)
Female × (share of female teachers)	−0.015	–	−0.051	−0.007	–	−0.070	0.052	–	0.029	−0.187	–	0.177
	(0.33)		(1.08)	(0.11)		(0.96)	(0.78)		(0.42)	(1.91)		(−1.85)
Female × (mean teacher experience in years)	–	−0.0024	−0.0030	–	−0.0051	−0.0058	–	−0.0022	−0.0019	–	0.0030	0.00073
		(1.78)	(2.12)		(2.47)	(2.56)		(1.07)	(0.86)		(0.93)	(0.24)
Observations	130,464	130,464	130,464	51,545	51,545	51,545	50,264	50,264	50,264	29,104	29,104	29,104
Standard error of equation	0.677	0.677	0.677	0.603	0.603	0.603	0.686	0.686	0.686	0.737	0.737	0.737

Note: The dependent variable is the difference between the teacher set grade and the result on the central exit exam. All models include year and school fixed effects, student characteristics, and teacher characteristics. *t*-values in parentheses are heteroskedasticity robust and clustered at the school level. Full models for columns (1), (4), (7), and (10) are reported in Appendix Table A2.

Bagues and Esteve-Volart (2010) investigate whether the gender composition of recruiting committees matters for hiring decisions in the Spanish judiciary system. They find that male candidates are more likely to be hired when they are randomly assigned to a committee where the share of female evaluators is high.

In Table 9, we expand Eq. (2) above with interaction terms between female student and the share of female teachers at the school and average teacher experience at the school. Column (1) indicates that the gender grading gap on average across subjects is not related to the gender of the teachers. However, while that holds for mathematics, there is a positive interaction in English which is not negligible, although insignificant, and a negative interaction effect in Norwegian that is large in magnitude and significant at 10 percent level. Taken at face value, the result for Norwegian in column (10) implies that female teachers on average, conditional on exam results, assess girls 0.19 points below male teachers.[20] The effect is equal to 0.24 standard deviations of the score difference. For within-sample variation, an increase in the share of female teachers at the school from 0.3 to 0.8 changes the gender grading gap from 0.06 score points in favor of girls to 0.03 score points in favor of boys. This result is in accordance with the same-sex punishment found in the literature.

Column (2) in Table 9 indicates that the gender grading gap is related to the experience of the teachers. Using the sample variation in average experience, the parameters imply that the gender grading gap across all subjects varies from 0.08 score points for minimum experience (10 years) to 0.03 score points for maximum experience (30 years). The impact of experience is larger in mathematics, but of the opposite sign in Norwegian.

Column (3) in Table 9 includes both interaction effects, and the results indicate that there are separate interaction effects of teacher gender and experience. The subject specific models indicate again that teacher gender is important for the gender grading gap in Norwegian and that teacher experience is important for the gap in mathematics.[21]

Since teacher characteristics are measured at the school level, they are noisy measures of the students' actual teacher. With classical measurement error, this works in the direction of downward biased estimates. Nevertheless, some of the point estimates of the interaction effects are non-negligible. One might wonder whether low precision of the estimates might be a result of the fact that the models include school fixed effects. School fixed effects can be important to include in these models in order to handle teacher sorting across schools. However, the results are insensitive to the exclusion of the school fixed effects in these difference-in-differences models. Teacher assessments, given student ability as measured by the results on the central exit exam, seem unrelated to school characteristics.[22]

6. Conclusion and discussion

Measures of student achievement are important for admission prospects in further education as well as for future job prospects. Test scores are also the preferred output indicator in studies of education production. Hence, the objectivity and reliability of available performance measures are important.

This paper has exploited information about individual student achievement for Norwegian students in their final year of compulsory education. On average, girls outper-

[20] Notice that since the interaction term is with the proportion of female teachers at the school, the coefficient reflects, strictly speaking, the effect of going from no female teachers to only female teachers at the school. This can only happen, of course, by replacing a male teacher with a female teacher in each classroom.

[21] One may wonder whether the interaction effects related to teacher characteristics are sensitive to the inclusion of interaction terms related to school choice. Expanding the models in Table 9 to include the interaction term related to school choice included in Table 7, all estimated parameters change only marginally.

[22] For example, in the model including both interaction terms an all subjects (column (3) in Table 9), the results when excluding the school fixed effects (*t*-values in parentheses) are 0.138 (3.15) for the indicator for female student, −0.052 (1.08) for the interaction with the share of female teachers, and −0.0030 (2.09) for the interaction with mean teacher experience.

form boys in all subjects considered both at high-stakes teacher grading and central exit exams. In a difference-in-differences framework, we find gender gaps in teacher grading. In all subjects, girls score relatively better than boys in the teacher grading than on the exams, even though the intention of these assessments clearly is to evaluate the same skills. The results indicate that the present evaluation system, which to a large extent relies on teacher grading, hurts boys more than the pure gender achievement gaps suggest. This evidence is not in accordance with the equilibrium theorem of Mechtenberg (2009), indicating that teacher behavior cannot explain the observed gender gaps in university enrollment and wages the way she suggests.

One cannot say a priori whether the observed gender grading gap is related to the students considering exams the most high-stakes test, the anonymous evaluation of exams, or the fact that exams are one-day tests. We have investigated whether some of these three potential explanations are more reasonable than others.

Boys may perform relatively better at the central exit exam because the exam is arguably the most competitive environment. Then the gender grading gap should increase in the importance of the grades. Exploiting both that the stakes are higher in some counties than in others, and an additional low-stakes test for one cohort, we find evidence in the opposite direction. In addition, the results regarding the low-stakes tests indicate that it is not simply the anonymous vs. non-anonymous dimension that matters. The gender grading gap has the opposite sign for the non-anonymous low-stakes test than for teacher grading.

We find indications that the gender grading gap is related to characteristics of the teachers. In Norwegian language, girls receive the highest grades when assessed by male teachers, and in mathematics girls receive the highest grades when assessed by inexperienced teachers. It might be that some teachers favor girls, either intentionally or not, and either in some complex interactions with student behavior or not, only when stakes are high. It seems reasonable to relate the gender grading gap to whether the assessment is based on one-day tests or performance over a longer period. For coursework elements in particular, one should expect teacher–student interactions to be important.

The results indicate that features of discrimination differ in schools and in the labor market. Some evidence for the labor market indicates statistical discrimination against females in male-dominated jobs and against males in female-dominated jobs (Riach & Rich, 2002). Education is female-dominated in the sense that girls in general perform better than boys and that a majority of the teachers are female, but the evidence indicates a grading gap in favor of girls.

Female behavior seems more context-dependent than male behavior (Croson & Gneezy, 2009). Thus, one might wonder to what extend the present evidence carries over to other countries. The Scandinavian countries are typically considered highly egalitarian. The employment rate of women is relatively high, and women have strong positions in politics. On the other hand, for a range of labor market and educational outcomes, Scandinavia is close to the European average (EU, 2008). That is the case for the female wage gap, the probability of being a manager or a senior civil servant, employment concentration in a few sectors, and enrollment in secondary and tertiary education. The age of the students might also be important for gender gaps since puberty can play an important role. However, similar findings for the gender grading gap for such diverse countries as Norway and Israel (Lavy, 2008), in our case at age 16 and in the Israeli case for high school graduates, indicates that girls in general tend to be graded more favorably than boys because of some important teacher–student interactions at school.

Appendix A

See Tables A1 and A2.

Table A1
Descriptive statistics independent variables.

	All subjects	Mathematics	English	Norwegian
Score	3.48	3.26	3.60	3.65
	(1.09)	(1.14)	(1.07)	(0.98)
Free school choice	0.55	0.56	0.54	0.57
Student characteristics				
Girl	0.49	0.49	0.49	0.49
First generation immigrant	0.019	0.019	0.018	0.019
Second generation immigrant	0.018	0.018	0.017	0.018
Student living with both parents	0.69	0.69	0.69	0.69
Higher education father	0.30	0.30	0.30	0.29
Higher education mother	0.32	0.32	0.31	0.32
Income father in 100,000 NOK	4.25	4.26	4.26	4.22
	(6.82)	(8.63)	(5.80)	(4.31)
Income mother in 100,000 NOK	2.31	2.32	2.31	2.30
	(1.94)	(2.23)	(1.72)	(1.72)
Teacher characteristics				
Mean experience in years	19.8	19.8	19.8	19.9
	(3.3)	(3.2)	(3.4)	(3.4)
Share female	0.54	0.54	0.55	0.54
	(0.10)	(0.10)	(0.11)	(0.11)

Table A1 (*Continued*)

	All subjects	Mathematics	English	Norwegian
Share without children	0.18	0.18	0.18	0.18
	(0.10)	(0.10)	(0.10)	(0.10)
Share married	0.64	0.64	0.64	0.65
	(0.12)	(0.12)	(0.12)	(0.13)
Observations	130,464	51,545	50,264	29,104

Table A2
Estimation results, full models.

	Models in Table 5				Models in Table 9			
	(1)	(3)	(5)	(7)	(1)	(4)	(7)	(10)
Dependent variable	Student score				Difference in student score			
	All subjects	Math	English	Norwegian	All subjects	Math	English	Norwegian
Female, assessment, and choice								
Female	0.302	0.059	0.375	0.600	0.059	0.061	0.039	0.117
	(38.3)	(5.61)	(33.2)	(46.5)	(2.40)	(1.67)	(1.04)	(2.14)
Teacher assessment	0.172	0.197	0.130	0.191	–	–	–	–
	(22.8)	(18.6)	(11.6)	(14.5)				
Female × (teacher assessment)	0.051	0.058	0.066	0.016	–	–	–	–
	(11.0)	(8.84)	(8.89)	(1.67)				
Female × (share of female teachers)	–	–	–	–	−0.015	−0.007	0.053	−0.187
					(0.33)	(0.11)	(0.78)	(1.91)
English	0.373	–	–	–	−0.051	–	–	–
	(31.9)				(3.91)			
Norwegian	0.300	–	–	–	−0.024	–	–	–
	(29.3)				(1.67)			
Student characteristics								
First generation immigrant	−0.298	−0.406	−0.206	−0.258	0.006	0.049	0.003	−0.077
	(12.8)	(10.8)	(5.77)	(6.82)	(0.37)	(2.35)	(0.11)	(2.23)
Second generation immigrant	−0.122	−0.202	−0.055	−0.067	0.031	0.039	0.055	0.0001
	(4.12)	(5.13)	(1.14)	(1.47)	(1.96)	(2.00)	(2.19)	(0.001)
Student living with both parents	0.266	0.363	0.186	0.229	0.042	0.035	0.043	0.043
	(41.2)	(34.1)	(20.5)	(24.1)	(8.68)	(5.98)	(5.65)	(4.21)
Higher education father	0.438	0.496	0.414	0.362	0.016	−0.0002	0.020	0.032
	(54.7)	(41.7)	(37.3)	(29.1)	(3.29)	(0.03)	(2.36)	(2.88)
Higher education mother	0.405	0.452	0.384	0.353	0.024	−0.004	0.041	0.046
	(46.2)	(33.0)	(28.2)	(28.8)	(4.96)	(0.57)	(5.32)	(4.12)
Income father	0.004	0.003	0.006	0.008	0.0004	0.0001	0.0008	0.0018
	(1.82)	(1.21)	(3.58)	(3.78)	(1.55)	(1.13)	(1.09)	(2.03)
Income mother	0.025	0.021	0.034	0.024	0.003	0.0002	0.007	0.006
	(3.73)	(2.13)	(3.85)	(4.41)	(2.61)	(0.23)	(2.99)	(2.36)
Teacher characteristics								
Mean experience in years	−0.004	0.003	−0.013	−0.022	0.001	0.003	−0.001	−0.038
	(0.85)	(0.42)	(1.66)	(1.55)	(0.11)	(0.35)	(0.13)	(2.11)
Share female	0.019	−0.115	−0.045	0.047	0.153	0.063	0.149	−0.710
	(0.17)	(−0.55)	(−0.27)	(0.11)	(1.16)	(0.24)	(0.49)	(1.21)
Share without children	−0.006	0.260	−0.109	−0.520	0.266	0.216	0.268	0.917
	(0.05)	(1.37)	(0.52)	(1.54)	(2.01)	(0.81)	(0.99)	(2.01)
Share married	0.132	0.153	0.290	−0.275	0.065	−0.023	0.237	0.348
	(1.48)	(0.91)	(1.82)	(0.97)	(0.54)	(0.10)	(0.99)	(0.68)
Year								
2003	0.009	−0.006	0.003	0.055	0.029	0.040	0.013	−0.103
	(0.86)	(0.25)	(0.13)	(1.07)	(2.08)	(1.28)	(0.40)	(1.93)
2004	0.045	−0.004	0.055	0.082	0.027	0.045	0.034	0.010
	(3.65)	(0.16)	(2.40)	(1.54)	(1.62)	(1.37)	(1.00)	(0.17)
2005	0.031	−0.061	0.092	0.057	0.096	0.207	−0.025	0.072
	(2.08)	(2.29)	(3.34)	(1.06)	(5.02)	(6.27)	(0.60)	(1.15)
School fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	260,928	103,090	100,528	58,208	130,464	51,545	50,264	29,104
Standard error of equation	0.977	1.023	0.967	0.854	0.677	0.603	0.686	0.737

Note: *t*-values in parentheses are heteroscedasticity robust and clustered at the school level.

References

Ammermueller, A., & Dolton, P. (2006). *Pupil–teacher interaction effects on scholastic outcomes in England and the USA* (ZEW Discussion Papers 06-60).

Bagues, M. F., & Esteve-Volart, B. (2010). Can gender parity break the glass ceiling? Evidence from a repeated randomized experiment. *Review of Economic Studies, 77*, 1301–1328.

Bonesrønning, H. (2004). Can effective teacher behavior be identified? *Economics of Education Review, 23*, 237–247.

Bonesrønning, H. (2008). The effect of grading practices on gender differences in academic performance. *Bulletin of Economic Research, 60*, 245–264.

Borghans, L., Meijers, H., & ter Wel, B. (2006). The role of noncognitive skills in explaining cognitive test scores. *Economic Inquiry, 46*, 2–12.

Croson, R., & Gneezy, U. (2009). Gender differences in preferences. *Journal of Economic Literature, 47*, 448–474.

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

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

Emanuelsson, I., & Fischbein, S. (1986). Vive la difference? A study on sex and schooling. *Scandinavian Journal of Educational Research, 30*, 71–84.

EU. (2008). *The life of women and men in Europe. A statistical portrait*. Luxembourg: Eurostat Statistical Books.

Figlio, D. N., & Lucas, M. E. (2004). Do high grading standards affect student performance? *Journal of Public Economics, 88*, 1815–1834.

Fryer, R. G., & Levitt, S. D. (2010). An empirical analysis of the gender gap in mathematics. *American Economic Journal: Applied Economics, 2*, 210–240.

Gneezy, U., Niederle, M., & Rustichini, A. (2003). Performance in competitive environments: Gender differences. *Quarterly Journal of Economics, 118*, 1049–1074.

Guiso, L., Monte, F., Sapienza, P., & Zingales, L. (2008). Culture, gender, and math. *Science, 320*, 1164–1165.

Haraldsvik, M. (2004). *Elevprestasjoner og konkurranse i den videregående skolen: Gir konkurranse om skoleplassene elevene insentiver til å jobbe hardere på ungdomsskolen?* Norwegian University of Science and Technology (Master thesis).

Heckman, J., Stixrud, N., & Urzua, S. (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor Economics, 24*, 411–482.

Hinnerich, B. T., Höglin, E., & Johannesson, M. (2011). Are boys discriminated in Swedish high schools? *Economics of Education Review, 30*, 682–690.

Hyde, J. S., Lindberg, S. M., Linn, M. C., Ellis, A. B., & Williams, C. C. (2008). Gender similarities characterize math performance'. *Science, 321*, 494–495.

Jacob, B. A. (2007). *Test-based accountability and student achievement: An investigation of differential performance on NAEP and state assessment*. National Bureau of Economic Research (Working Paper 12817).

Jacob, B. A., & Levitt, S. D. (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *Quarterly Journal of Economics, 118*, 843–877.

Kane, T. J., & Staiger, D. O. (2002). The promise and pitfalls of using imprecise school accountability measures. *Journal of Economic Perspectives, 16*, 91–114.

Lavy, V. (2008). Do gender stereotypes reduce girls' or boys' human capital outcomes? Evidence from a natural experiment. *Journal of Public Economics, 92*, 2083–2105.

Leuven, E., Oosterbeek, H., & Van Ophem, H. (2004). Explaining international differences in male skill wage differentials by differences in demand and supply of skill. *Economic Journal, 114*, 466–486.

Lindahl, E. (2007a). *Comparing teachers' assessments and national test results—Evidence from Sweden*. Institute for Labour Market Policy Evaluation (Working Paper 2007:24).

Lindahl, E. (2007b). *Gender and ethnic interactions among teachers and students—Evidence from Sweden*. Institute for Labour Market Policy Evaluation (Working Paper 2007:25).

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

Machin, S., & Pekkarinen, T. (2008). Global sex differences in test scores variability. *Science, 322*, 1331–1332.

Machin, S., & Salvanes, K. G. (2010). *Valuing school quality via a school choice reform* (IZA Discussion Paper No. 4719).

Mechtenberg, L. (2009). Cheap talk in the classroom: How biased grading at school explains gender differences in achievement, career choices and Wages. *Review of Economic Studies, 76*, 1431–1459.

Murnane, R. J., Willett, J. B., & Levy, F. (1995). The growing importance of cognitive skills in wage determination. *Review of Economics and Statistics, 77*, 251–266.

Paserman, D. M. (2007). *Gender differences in performance in competitive environments: Evidence from professional tennis players* (IZA Discussion Paper No. 2834).

Riach, P. A., & Rich, J. (2002). Field experiments of discrimination in the market place. *Economic Journal, 112*, F480–F518.

Steel, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist, 65*(5), 797–811.

Stobart, G., Elwood, J., & Quinlan, M. (1992). Gender bias in examination: How equal are the opportunities? *British Educational Research Journal, 18*, 261–276.

Van Ewijk, R. (2011). Same work, lower grade? Student ethnicity and teachers' subjective assessments. *Economics of Education Review, 30*, 1045–1058.

Örs, E., Palomino, F., & Peyrache, E. (in press). Performance gender-gap: Does competition matter? *Journal of Labor Economics* (in press).