WORKING PAPER SERIES

No. 4/2011

Educational Evaluation Schemes and Gender Gaps in Student Achievement

Torberg Falch
Department of Economics, Norwegian University of Science
and Technology (NTNU), and CESifo

and

Linn Renée Naper
ECgroup AS and Centre for Economic Research at NTNU

Department of Economics

Norwegian University of Science and Technology

N-7491 Trondheim, Norway
www.svt.ntnu.no/iso/wp/wp.htm

Educational Evaluation Schemes and Gender Gaps in Student Achievement

Torberg Falch

Department of Economics, Norwegian University of Science
and Technology (NTNU), and CESifo

and

Linn Renée Naper

ECgroup AS and Centre for Economic Research at NTNU

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-difference framework. Compared to scores at anonymously evaluated central exit exams, girls get significantly higher grades than boys when assessed by their teacher. We find no evidence that the competitiveness of the environment can explain why boys do relatively better at the exam. The gender grading gap is related to teacher characteristics. The results indicate that the teacher-student interaction during coursework favor girls in the teacher grading.

Keywords: Educational evaluation schemes; Teacher grading; Gender gaps; Gender interactions

JEL-classifications: I21

1

1. Introduction

Mechtenberg (2009) presents a game theoretical model in which gender gaps in equilibrium are similar to observed gender differences in school achievement, university enrolment, and wages. In her model, there are two subjects at school - mathematics and humanities – and students' beliefs about own ability depend on teacher grading. The crucial assumption for the fascinating 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 for boys in humanities and for girls in mathematics, and hard grading for 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. This theorem is testable by comparing teacher grading with external evaluation of achievement.[1]

The present paper exploits achievement scores for Norwegian students at the end of compulsory schooling. Teachers set grades based on tests given throughout the whole school year, and all students conduct a central exit exam that is graded anonymously. All grades matter for admission to upper secondary schools and are in this respect high-stake.

The observed gender gap in student achievement in favor of girls is often explained by increased share of female teachers. For example Dee (2005a, 2005b) and Ammermueller and Dolton (2006) find evidence that students profit from having 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 since they are not related to intentional behavior of teachers. Thus, passive teacher effects cannot explain

---

[1] 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).

2

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) find that the bias on the non-blind test is against boys 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. The findings of Emanuelsson and Fischbein (1986), Stobart et al. (1992), Lindahl (2007a), and Bonesrønning (2008) indicate that greater weight on coursework elements improves the relative performance of girls. Machin and McNally (2005) show that the gender gap in the UK aroused in the afterwards of the change in examination system in 1988. The importance of coursework increased in the new system.

We find that girls obtain better scores than boys in teacher grading relative to the central exit exam in both mathematics, English, and Norwegian language in the period 2002–2005. Thus, our results are not in accordance with Mechtenberg's (2009) central theorem. The gender grading gaps estimated, however, are of the same magnitude as found by Lavy (2008). Other mechanisms than those put forward by Mechtenberg (2009) must explain the grading behavior. We investigate whether the finding in Gneezy et al. (2003) that males perform relative better in competitive environments can explain the gender gap. We explore that the extent to which grades matter for admission to upper secondary education varies across counties, and, in addition, that one specific cohort conducted a separate low-stake test. The results indicate that the competitiveness of the environments cannot explain the gender grading gaps. In addition, the results for the low-stake test indicate that the gaps are not 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 teacher-student relationship seems important, although with a different kind of interaction than assumed by Mechtenberg (2009). We conclude that the most reasonable explanation for the gender grading gaps is that teacher grades are based on performance over a longer period than single tests. Teacher grades might include coursework elements that favor girls.

3

The next section offers a more closely description of the Norwegian educational system and student evaluation schemes. Section 3 presents the data. Section 4 includes the main results on gender grading gap in teacher assessments, while in Section 5 some investigations on possible explanations of the observed gender gap are discussed. Section 6 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 reach the 10th grade. 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 (8th to 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 conducts one central written exam of five hours, 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 the assignment of examination subjects to schools and individual students given clear instructions from the Directorate. The teachers or schools have no influence in this respect. 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 40 percent in English.[2] The exam result is determined by two external examiners assigned to each student.

Teacher grading is the responsibility of individual teachers. According to the school law, teacher grade should be based on the achievements throughout the school year, and should express the students' competence and skills. Grades shall not reflect student effort. In Norwegian language and English, there are separate marks for written and oral skills, where the former is based on tests and the latter on performance in class. We compare the former with the exam results because they shall measure the same skills. In fact, teachers often use questions from former exams in their tests, and as a part of the basis for their evaluation is a

---

[2] Students with exam in the Norwegian language had the exam over two days in the empirical period. There are two formal written Norwegian languages, and the students conducted exam in both, and got separate grades from their teachers in both. In this paper we only consider the results for the main Norwegian language.

4

one-day test of five hours. Although the performance throughout the whole school year matter, the performance in the latest tests is most important. In central exam subjects, teacher grades should be given at least one day before the notification of exam results.

Teacher grades and central exam results are equally important for students' final grade point average (GPA). GPA matters for the prospects of admission to upper secondary study tracks and schools. There is a legal right of 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 determines location of schools and the composition of study tracks at each school. About 45 percent enroll in the academic study track which 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 electrician, carpenter, practical nurse, etc. Most schools have several study tracks.

In their application, 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-stake in this respect. In addition, in some counties there is free school choice. Students have to rank schools in addition to study tracks in their application, and admission to over-subscribed 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.[3] Thus, GPA is high-stake 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.

A national student evaluation scheme was implemented in the spring 2004. All final grade students had to conduct tests in all three exam subjects. The tests were designed as instruments 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 results should not be taken into consideration when the teachers decided on final grades. The tests had thus no consequences for the students. The testing time was short, about one hour, and the content of the test could differ from the skills tested in the high-stake assessments. In particular, the test in Norwegian did not include writing an essay. It was a test of reading skills.

---

[3] 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.

5

According to Borghans et al. (2006), individual effort and achievement depend on the reward related to the result. Student evaluation schemes can in general differ along three dimensions. They can be anonymous or non-anonymous, based on a single test or the performance over a longer period, or influence individual students' prospects (high-stake tests) or not (low-stake tests).[4] Table 1 classifies possible evaluation schemes into six different types.[5] 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 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 and immigration status, and parents' income, marital status and education. 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 at 1st to 10th grade that typically are small and located in rural areas. Since information on whether teachers work at the primary or lower secondary level is not available, 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.[6] Since our identification is based on within-student variation in achievement, the

---

[4] Low-stake 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-stake tests by nature, which to a large extent also is the case for the grades in US high school since admission to colleges mainly rely on the SAT test conducted after graduating from high school. 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-sake tests for the schools, while, at the same time, low-stake 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).

[5] Our classification into three dimensions indicates that there are eight different types of evaluation schemes (4 x 2 table). However, it is hard to imagine anonymous evaluations based on observations over a longer time period.

[6] In models that do not include information of 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.

estimation sample only includes students with both a teacher grade and an exam result in a given subject.[7] Each student is only observed in one subject.

Both teacher grading and exams use a grading scale from 1 to 6, where score 6 is best and 1 is very weak. Figure 1 presents the distribution of scores across assessment schemes, subjects, and gender. A score of 3 or 4 is most common, each including 20 to 40 percent of the students in the different groups. The distributions are close to normal in all cases, although there are two distinct patterns. First, the scores are always better in teacher grading than at exam. Obviously, several students get a lower score at 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.

Table 2 and 3 cross-tabulates the percentages with the different combinations of scores at the exam and the teacher grading in mathematics for female and male students, respectively. It is most common to get the same score in both evaluation schemes. However, several students get one grade lower at 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 at the exam (that is 10.0 percent of the total sample of girls). The figures are similar for boys, but with a tendency that fewer students get a lower score at 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 lower score at the exam than in the teacher assessment.

Table 4 compares mean scores in teacher grading and exam across gender. For each subject, the table report average teacher grades, exam results, and the test-statistic from a mean comparison test across gender and evaluation schemes. 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 small 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 (2009). Guiso et al. (2008) argue that the gender achievement gap in mathematics is related to gender equality in general, and in the most gender-equal societies girls perform at least as well as boys.

---

[7] Some students are exempted from the external exam because of illness on examination day, disabilities, etc. The written exam absence rate is close to 4 percent each year.

Table 4 also shows that average scores in teacher grading are higher than exam scores in all cases. 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-difference estimator, corresponding to parameter γ in equation (1) below, is equal to 0.05, 0.07, and 0.02 in mathematics, English, and Norwegian, respectively, and significant at 5 percent level in all cases.

Appendix Table A1 reports descriptive statistics. The first column includes all students, and the statistics for the "score"-variable merge all subjects. Regarding student characteristics, 69 percent are living with both parents, 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,[8] 64 percent are married and 18 percent do not have children.

In the last three columns in Appendix Table A1, only the individuals conducted the relevant exam is included. Each student only takes one exam. Overall, there are very small differences in background characteristics across the subjects, even though at many schools it is the case that all students have the same exam subject. This clearly support 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-difference model.

$$(1) \qquad A_{Eijt} = \alpha + \lambda G_{ijt} + \delta E_{ijt} + \gamma \left(E_{ijt} \text{ x } G_{ijt} \right) + \beta X_{ijt} + \phi_j + \mu_t + \sigma_{Eijt},$$

where the score A_{Eijt} at evaluation E ($E=1$ for teacher grading and $E=0$ for central exam) of student i at school j at time t is assumed to be a function of gender G 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

---

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

effects, ϕ_j and μ_t, respectively. σ_{Eijt} is an i.i.d. error term. Since 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-difference 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 at 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

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

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

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

9

4.2 Results

Table 5 presents results using the model specification described in equation (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 the 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 tests 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 achievement gap (as measured by the exam) is 0.30 score points at 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.[9] 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 compare two single day tests, we

---

[9] 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.

10

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.

The results discussed here are not in line with the equilibrium conditions in Mechtenberg's (2009) cheap-talk-in-the-classroom 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.

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

We consider two possible explanations of the observed grading gap against boys. First, it is arguably a more competitive environment at an exit exam than at tests taken throughout the school year, and we will investigate whether this can explain why males do relatively better at 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. In addition, Niederle and Vesterlund (2007) find that females are less willing than men to enter a mixed-sex competition. 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. (2008) 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 that GPA matters more in some counties than in other counties. Secondly we to compare the exam result to the one-day low-stake national tests in 2004.

If the observed gender grading gap is due to a more competitive environment at 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 at 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), 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 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 are few changes over the last 20 years.[10] In 11 municipalities, mainly medium sized cities, there was a mixed system. We skip these observations from the analysis below (six percent of the observations). Appendix Table A1 shows that there where free school choice for 55 percent of the observations.

Table 6 presents models estimating different version of equation (2) above. The model in column (1) is identical to the model in column (2) in Table 5, except that the model includes the interaction term between gender and free school choice.[11] 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, relative to male students, 0.059 score points better in teacher grading than at the exam, while with free school choice the gender grading gap is 0.040 score points. Female students thus perform relatively better at the exam in areas with free school choice. The difference is significant at five percent level. The model in column (2) in Table 6 includes time fixed effects and student and teacher characteristics, without affecting the estimated gender grading gaps.

One can argue that school choice is always limited in rural areas. Thus, in column (3) in Table 6, we exclude observations in small municipalities classified as having free school choice. 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

---

[10] Some changes occur. Oslo changed from a mixed system to a system of free 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.

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

school choice. The grading gap in Oslo is estimated to 0.035 score points, close to the average of counties with free school choice.

The last columns in Table 6 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. The latter yields the low average gender grading gap revealed in Table 5.

Table 7 compares the exam results to the one-day low-stake national tests.[12] It turns out that female students have a relatively lower score at the low-stake test than at the exam. 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. This result must be interpreted with caution since the low-stake test in Norwegian focused on reading and thus tested somewhat different skills than the exam. The results for Norwegian can be interpreted as boys are relatively better in reading than in writing. However, there is also a large gender gap in mathematics which is clearly not in accordance with the competitiveness of the environment hypothesis.

Table 7 also includes models in which the female dummy variable is interacted with whether there is free school choice in the county. Since the difference in 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 estimate in column (3) in Table 7 does not support the hypothesis. However, restricting the sample by excluding small municipalities classified as having free school choice, changes the sign of the interaction term. In particular in Norwegian it seems like girls are doing relatively better on the low-stake test in counties with free school choice

---

[12] 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. Since data is only available for one year, teacher characteristics are collinear to the school fixed effects.

than in other counties.[13] However, giving the high standard errors, no conclusions can be drawn concerning the effect of free school choice.

Overall, conditional on the high-stake central exit exam, boys outperform girls on the low-stake national one-day test. Thus, the hypothesis that girls perform relatively worse when stakes are high is not supported.[14]

These results also indicate that the anonymous vs. non-anonymous dimension of the evaluations schemes is not important for 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.

5.2 Gender grading gap and teacher-student interaction

Inspired by the literature on gender stereotypes and student–teacher gender interactions (e.g., Steel, 1997; Dee, 2005b; Ammermueller and Dolton, 2006), 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 student–teacher 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 student–teacher 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 use teacher composition at school instead of matching teachers to

---

[13] It is the sign of the interaction effect in Norwegian which is sensitive to whether we restrict the sample or not. If we restrict the sample to the two largest cities, Oslo and Bergen, the average interaction effect across all subjects increases to 0.10. Also in this case the interaction effect is largest for Norwegian, but insignificant at 10 percent level.

[14] Our results are quantitatively larger than Lindahl's (2007a) finding for Sweden. Lindahl compares high-stake teacher grades based on whole year assessment with low-stake national tests. We have not estimated the same gender grading gap directly, but this gap follows by taking the difference between the estimated gap in teacher grading and the estimated gap for the national test. In mathematics, our estimate comparable to Lindahl is 0.168 (0.057–(-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.

14

students. On the other hand, we avoid biases related to strategic assignment of teachers to classes within schools.

Lavy (2008) discusses in 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 on teacher strategies and teachers react on 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, 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 8 we expand equation (2) by interaction terms between female student and both the share of female teachers at the school and average teacher experience. Column (1) indicates that the gender grading gap on average across subjects are 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) in Table 8 implies that female teachers on average, conditional on exam results, assess girls 0.19 points below male teachers. 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 8 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 opposite sign in Norwegian.

15

Column (3) in Table 8 includes both interaction effects, which indicates that there may be 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.[15]

6. Conclusion and discussion

Student achievement measures are important for both admission prospects in further education as well as for future job prospect. 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 outperform boys in all subjects considered both at high-stake teacher grading and central exit exams. In a difference-in-difference framework, we find gender gaps in teacher grading. In all subjects girls score relatively better than boys in the teacher grading than at the exams. This result contradicts the equilibrium theorem of Mechtenberg (2009). Our evidence indicates that teacher grading practices and teacher-student interactions cannot explain observed gender differences in university enrolment and wages the way she suggests. Indeed, our 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.

One cannot *a priori* say whether the observed gender grading gap is related to exams being considered by the students as the most high-stake test, the fact that exams are evaluated anonymously, or that exams are one-day tests. We have investigated the whether some of these three potential explanations are reasonable.

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 that one cohort conducted an additional low-stake test, we find evidence in the

---

[15] 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 7 with the interaction term related to school choice included in Table 6, all estimated parameters change only marginally.

opposite direction. In addition, the results regarding the low-stake test indicate that the anonymous vs. non-anonymous dimension is important. The gender grading gap has the opposite signs for the non-anonymous low-stake test and teacher grading.

The gender grading gap seems to be related to characteristics of the teachers. In Norwegian, girls receive highest scores when assessed by a male teacher, and in mathematics girls receive highest grades form low-experienced teachers. One interpretation of the results is that 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 more 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.

References

Ammermueller, A. and Dolton, P. (2006): Pupil–Teacher Interaction Effects on Scholastic Outcomes in England and the USA, ZEW Discussion Papers 06-60.

Bagues, M. F. and B. Esteve-Volart (2010). Can Gender Parity Break the Glass Ceiling? Evidence from a Repeated Randomized Experiment. Review of Economic Studies, 77, 130-1328.

Bonesrønning, H. (2004). 'Can effective teacher behavior be identified?', *Economics of Education Review*, 23(3), pp. 237–47.

Bonesrønning, H. (2008): The Effect of Grading Practices on Gender Differences in Academic Performance, Bulletin of Economic Research, 60 (3), pp. 245-264.

Borghans, L., Meijers, H. and ter Wel, B. (2006): The Role of Noncognitive Skills in Explaining Cognitive Test Scores, Economic Inquiry, 46 (1), pp. 2-12.

Dee, T. S., (2005a): A Teacher Like Me: Does Race, Ethnicity, or Gender Matter? American Economic Review 95(2), pp. 158–165.

Dee, T. S., (2005b): Teachers and the Gender Gaps in Student Achievement, Working Paper No. W11660, National Bureau of Economic Research.

Emanuelsson, I. and Fischbein, S. (1986): Vive la Difference? A Study on Sex and Schooling, Scandinavian Journal of Educational Research, 30 (2), pp. 71–84.

Figlio, D. N. and Lucas, M. E. (2004). 'Do high grading standards affect student performance?', *Journal of Public Economics*, 88(9–10), pp. 1815–34.

Fryer, R. G., and S. D. Levitt (2009): An Empirical Analysis of the Gender Gap in Mathematics, NBER Working Paper 15430.

Gneezy, U., Niederle, M. and Rustichini, A., (2003): Performance in Competitive Environments: Gender Differences, Quarterly Journal of Economics, 118(3), pp. 1049-1074.

Guiso, L., Monte, F., Sapienza, P. And Zingales, L. (2008): Culture, Gender, and Math, Science, 320(5880), pp. 1164-1165.

Haraldsvik, M. (2004):

Jacob, B. A. (2007): Test-based Accountability and Student Achievement: an Investigation of Differential Performance on NAEP and State Assessment, Working Paper 12817, National Bureau of Economic Research.

Jacob, B. A. and Levitt, S. D. (2003): Rotten Apples: and Investigation of the Prevalence and Predictors of Teacher Cheating, The Quarterly Journal of Economics, 118 (3), pp. 843–877.

Kane, T. J. and Staiger, D. O., (2002): The Promise and Pitfalls of Using Imprecise School Accountability Measures, Journal of Economic Perspectives, American Economic Association, vol. 16(4), pp. 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, pp. 2083-2105.

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. and S. McNally (2005): Gender and Student Achievement in English Schools, Oxford Review of Economic Policy, 21, 357-372.

Machin, S. and K. G. Salvanes (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.

Niederle, M. and L. Vesterlund (2007) Do Women Shy Away from Competition? Do Men Compete Too Much? Quarterly Journal of Economics, 122, 1067-1101.

Paserman, D. M. (2007): Gender Differences in Performance in Competitive Environments: Evidence from Professional Tennis Players, IZA Discussion Paper No. 2834 –June 2007.

Steel, C., M. (1997): A Threat in the Air – How Stereotypes Shape Intellectual Identity and Performance, American Psychologist, 65 (5), pp. 797–811.

Stobart, G., Elwood, J. and Quinlan, M. (1992): Gender Bias in Examination: How Equal are the Opportunities? British Educational Research Journal, 18(3), pp. 261–76.

Örs, E., F. Palomino, and E. Peyrache (2008). Performance Gender-Gap: Does Competition Matter? CEPR Working Paper 6891.

19

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

1a. Mathematics, female students
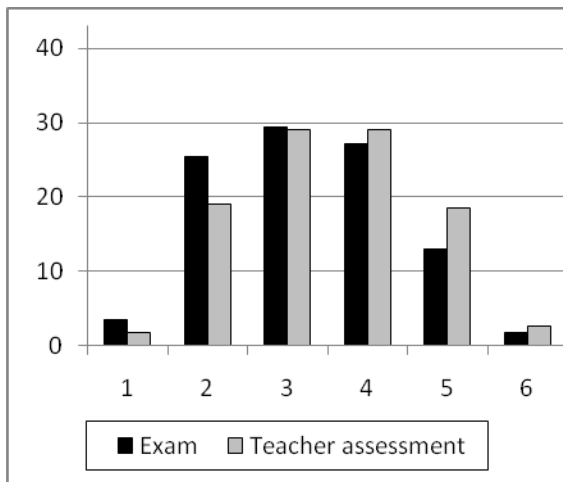
1b. Mathematics, male students
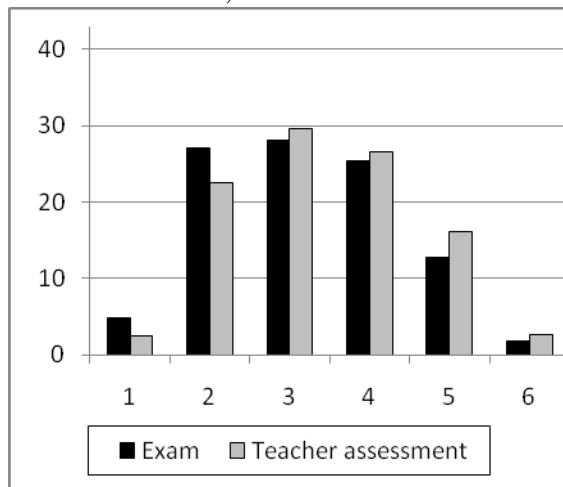
1c. English, female students
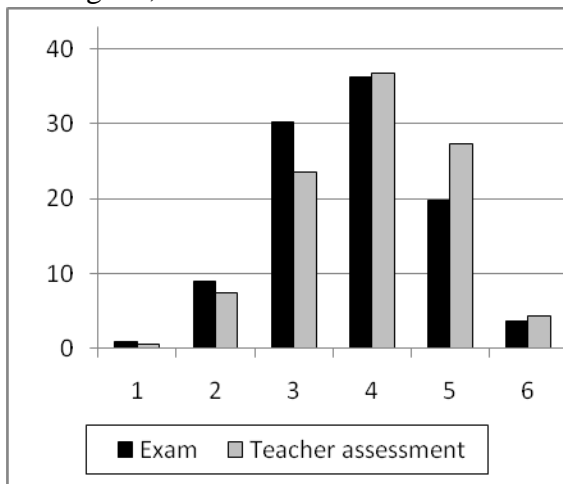
1d. English, male students
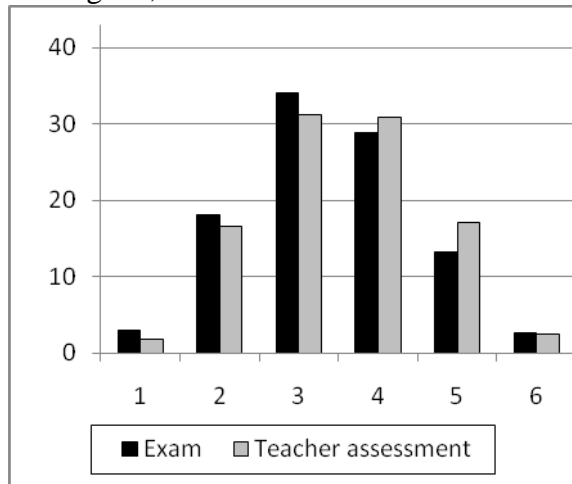
1e. Norwegian, female students
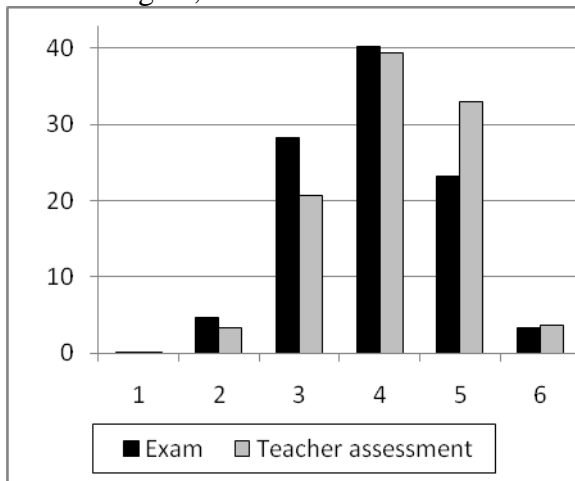
1f. Norwegian, male students
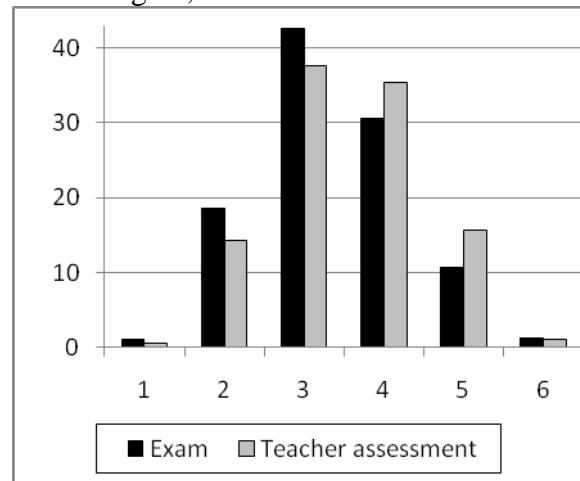
Table 1. A classification of evaluation schemes

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

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

Exam result Teacher assessment	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

Exam result Teacher assessment	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

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

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

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

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 x (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) x (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 heteroscedasticity 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.

Table 6. 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	0.060	0.103	0.068	0.068	0.060	0.060	0.051	0.051
	(9.01)	(8.87)	(8.91)	(7.16)	(6.81)	(6.80)	(5.46)	(5.45)	(3.41)	(3.40)
Female x (Free school choice)	-0.019	-0.020	-0.024	-0.068	-0.026	-0.037	-0.009	0.015	-0.061	-0.074
	(2.08)	(2.15)	(1.82)	(2.68)	(2.04)	(2.30)	(0.58)	(0.75)	(3.12)	(2.77)
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 & teacher char.	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	131,868	122,422	78,173	13,447	48,463	30,441	47,424	31,208	26,983	16,838
Standard error of equation	0.678	0.675	0.678	0.683	0.603	0.602	0.685	0.688	0.733	0.739

Note: t-values in parentheses are heteroscedasticity robust and clustered at the school level. The restricted sample excludes municipalities with less than 250 students in the cohort on average during the empirical period in counties classified as having free school choice.

Table 7. Gender gap and high-stake vs. low-stake tests. Dependent variable is *difference* in student score

	All subjects				Mathematics		English		Norwegian	
	(1)	(2)	(3)	(4)	(4)	(5)	(6)	(7)	(8)	(9)
Female	-0.099	-0.104	-0.096	-0.096	-0.111	-0.101	0.032	0.022	-0.348	-0.310
	(7.65)	(7.88)	(5.18)	(5.14)	(8.69)	(5.12)	(1.80)	(0.81)	(15.7)	(11.1)
Female x (Free school choice)	-	-	-0.020	0.051	-	0.014	-	0.017	-	0.046
			(0.77)	(1.47)		(0.50)		(0.34)		(0.67)
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	14,829	9,370	8,948	11,262	10,688	6,053	5,839
Standard error of eq.	0.773	0.735	0.734	0.739	0.633	0.648	0.748	0.765	0.830	0.826

Note: Dependent variable is grade, stacking national test and central exam. t-values in parentheses are heteroscedasticity robust and clustered at the school level. The restricted sample excludes municipalities with less than 250 students in the cohort in counties classified as having free school choice.

Table 8. 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)	(2.927)	(1.04)	(2.64)	(1.33)	(2.14)	(0.69)	(1.13)
Female x (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 x (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: Dependent variable is the grade difference between teacher assessment and central exam. All models include year and school fixed effects, student characteristics, and teacher characteristics. t-values in parentheses are heteroscedasticity robust and clustered at the school level. Full models for columns (1), (4), (7), and (10) are reported in Appendix Table A2.

Appendix

Appendix 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 100,000NOK	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)
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)
Obervations	130,464	51,545	50,264	29,104

Table A2. Estimation results, full models

Dependent variable	Models in Table 5				Models in Table 8			
	(1)	(3)	(5)	(7)	(1)	(4)	(7)	(10)
	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 x (Teacher assessment)	0.051	0.058	0.066	0.016	-	-	-	-
	(11.0)	(8.84)	(8.89)	(1.67)				
Female x (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.