Studies in Educational Evaluation 59 (2018) 141–149

Contents lists available at ScienceDirect

Studies in Educational Evaluation

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

Gender Bias in teachers' grading: What is in the grade☆

Tomáš Protivínský, Daniel Münich*

CERGE-EI, a joint workplace of Charles University in Prague and the Economics Institute of the Czech Academy of Sciences, Politickych veznu 7, 111 21 Prague, Czech Republic

ARTICLE INFO

Keywords:
Educational economics
Gender gap
Discrimination
Grading
The Czech Republic
Academic achievement
Educational assessment
Performance of educational system

ABSTRACT

It is a common empirical finding across countries that, on average, girls outperform boys in languages, whereas boys outperform girls in mathematics. We enrich the existing empirical evidence by combining admission test scores and teachers' grading of 15-year-old pupils' performance in mathematics and their native language in the Czech Republic, and furthermore, we investigate possible gender bias in teachers' grading. The directions of the gender differences in performance we estimate are consistent with international patterns and we document that teachers' grading is biased in favor of girls both in mathematics and in native language. The gender effect in grading is sizeable across the whole performance distribution and can be explained neither by the students' differing perceptions of stress at exams, nor by the students' attitudes toward the subject in question. The most plausible explanation is that the gender grading gap is due to gender difference in non-cognitive skills, such as in-class behavior and homework, confounding teachers' grades but not test scores. Since grades constitute the main feedback about students' academic performance and a crucial factor in decision-making about their future academic careers, biased grading may cause inefficiencies to the educational system and consequently can negatively affect future labor market careers.

1. Introduction

Various types of gender gap are the subject of intensifying empirical research efforts worldwide. Our work is focused on gender bias in teachers' grading, which confounds the quantitative indicators used to identify actual (real) gender gaps in pupils' and students' skills achievement. Two different types of gender gap can be found in the economics of education literature: gaps in achievement and relative gaps in grades, awarded by teachers at their discretion, in contrast to test scores. The term "gender gap" (sometimes also called "gender achievement gap"), usually referring to the gap in performance between girls and boys, is commonly measured by the difference in their test scores. Several stylized patterns have been observed: girls outperform boys in languages, whereas boys typically score better in mathematics, although the latter difference is not as distinct and pronounced (cf. Machin & Pekkarinen, 2008; Cornwell, Mustard, & Van Parys, 2013; Guiso, Monte, Sapienza, & Zingales, 2008; Angelo, 2014). Machin and McNally (2005) document that the position of girls relative to boys has improved over the past few decades. Numerous authors have concluded that this gender gap originates from a gender-biased culture and environment rather than from innate biological differences in ability (Guiso et al., 2008; OECD, 2015). Relatively few authors, however, have explored gender gaps in grading as opposed to test-scores. In the Czech Republic, Straková, Potužníková, and Tomášek (2006) found that on average girls receive higher grades than boys in all measured subjects across the majority of types of education.

The gender grading gap then captures the extent to which a boy and a girl with the same level of skills (based on anonymous testing) obtain different grades via non-anonymous evaluation, typically by teachers. It constitutes a gender bias in grading when contrasted with other forms of evaluation. In most schooling systems, grades represent an established form of feedback to pupils and their parents about their performance and/or learning progress. Appropriate feedback can help pupils and parents to learn about the efficiency of their effort and foster their further progress. Grading is also an important factor in forming aspirations and making career-related decisions; Federičová (2016) shows that students rely heavily on grades when deciding whether to apply to selective high-schools. Symmetrically, incorrect or biased feedback can impede educational attainment and cause sub-optimal skills acquisition with life-long consequences. It can foster a mismatch between students' skills and the educational track they choose, and consequently lead to occupational mismatch on the labor market.

As Hinnerich, Höglin, and Johannesson (2011) put it, given the growing focus of contemporary societies on gender equality, it is

☆ The research was supported by the Czech Science Foundation (grant P402/12/G130).
* Corresponding author.
E-mail address: daniel.munich@cerge-ei.cz (D. Münich).

https://doi.org/10.1016/j.stueduc.2018.07.006
Received 13 October 2017; Received in revised form 13 July 2018; Accepted 16 July 2018
0191-491X/ © 2018 Elsevier Ltd. All rights reserved.

desirable to explore the nature and the causes of the gender grading gap in different settings and countries more intensively. So far, the gender grading gap has been investigated in India, Israel, Norway, Portugal, Sweden, the United Kingdom and the United States, although there have been substantial differences in the methodology used and to some extent also in the findings. The gender gap in grading 15-year-olds in the Czech Republic has been researched by Matějů and Smith (2014) and Matějů and Simonová (2013). We confirm their findings here using different, and to some extent better data, and further explore the possible role of differences in pupils' attitudes to subjects and in their perception of stress during exams.

In the following section we offer a thorough literature review; Section 3 then describes the key features of the Czech educational system related to the phenomena we are investigating and the data we use. Section 4 presents our quantitative findings about gender achievement and the grading gap, and in the final section we conclude, placing our findings within the relevant policy framework and discussing their implications for further research.

2. Literature review

A gender gap in grading occurs when a teacher gives students of different genders grades that systematically differ but not due to their performance; this can be caused by numerous factors including intentional or unintentional discrimination. The typical result is that boys on average receive worse grades than girls, despite their level of academic skills being the same. Lavy (2008) introduced a very explicit term for this phenomenon: "anti-male bias in grading". Other terms describing the same notion are gender grading gap, gender differences in teachers' grading or gender bias in teachers' grading and these are used interchangeably in the presented study.

The existing literature about the gender grading gap is neither vast, nor consistent. Research methodology in this area always relies on comparing two different evaluation schemes, i.e. anonymous and non-anonymous assessment, although there are differences in the particular types of assessment used. Falch and Naper (2013) propose a classification of assessments based on three dimensions: (i) anonymous vs. non-anonymous, (ii) high-stakes vs. low-stakes, (iii) one-day vs. assessment over time. For instance, teachers' grading represents non-anonymous, high-stakes assessment over time, whereas admission exams (e.g. to high schools or university studies) can be considered as anonymous, high-stakes, one-day assessment. Even though grading by teachers, whether one-off or over a period of time, is a prototypical example of non-anonymous assessment, some authors have also included other forms of evaluation. These differences in methodology can directly translate into differences in results (cf. Hanna & Linden, 2012).

Table 1 summarizes the existing findings about the gender grading gap. Most studies found a gender grading gap in favor of girls in some or all subjects, reporting that girls typically receive higher grades than boys with the same academic skills. Only two studies did not find any bias in grading – Hanna and Linden (2012) and Hinnerich et al. (2011). The latter investigated only one subject – Swedish as a native language, while the former used a very different methodology than all the other studies. As non-anonymous evaluation, Hanna and Linden (2012) asked teachers to evaluate an exam sheet with a particular gender, age and caste assigned. Falch and Naper (2013) suggest that the gender grading gap is caused by teacher-student interaction. This hypothesis is consistent with Hanna and Linden's finding that the gender grading gap does not occur at all when there is no teacher-student interaction. This disparity also suggests that personal ties between a teacher and a student might be important for the occurrence of the gender grading gap. The most distinct results are reported by Gibbons and Chevalier (2008) and Lavy and Sand (2015), who found bias against boys in some subjects and bias against girls in other subjects. These are the only cases when bias in grading against girls was identified in the empirical literature; the particular subjects concerned differ in these studies.

Two other differences between teachers' grading and anonymous evaluation that are documented in the empirical literature should be noted. The inflation of teachers' grading describes a common situation when pupils graded by their own teachers receive better grades compared to anonymous evaluation (Falch & Naper, 2013; Hinnerich et al., 2011). Lower variance of teachers' grading reflects the fact that teachers frequently under-assess high-achieving students and over-assess low-achieving students (Gibbons & Chevalier, 2008).

Several authors provide various possible explanations for why grades could be biased in favor of one gender (Angelo, 2014; Falch & Naper, 2013; Lavy, 2008): (i) different intensity of competition during evaluations, (ii) teacher-student interaction, (iii) non-cognitive skills reflected in teachers' grading. Lavy (2008) offers several other possible explanations, although his data does not confirm them – these are that anonymous and non-anonymous assessment might emphasize different parts of curriculum, hence measure different skills, interaction between teacher's grading bias and student's ability (in combination with lower variance of teachers' grading), statistical discrimination against boys or different timing of exams. Below, we briefly discuss the three aforementioned explanations and the empirical evidence in support of them.

2.1. Competitive environment

The majority of research compares two different evaluations related to forms of assessment that possibly differ in terms of the intensity of competitiveness and stress to which they subject the students. Several studies about performance under stress suggest that men outperform women, even if they would perform similarly in a non-competitive environment (in an experimental setting: Gneezy, Niederle, & Rustichini, 2003; in a real-world setting: Ors, Palomino, & Peyrache,

Table 1
An overview of the existing research about the gender grading gap (studies are listed in alphabetical order, including the findings of this study).

Author	Country	Age	Finding
Angelo (2014)	Portugal	Grade 12	Bias against boys in math and Portuguese
Bonesrønning (2008)	Norway	Grade 9	Bias against boys in math
Cornwell et al. (2013)	US	Kindergarten - grade 5	Bias against boys in reading, math and science
Emanuelsson and Fischbein (1986)	Sweden	Grade 6, 8, 9	Bias against boys in Swedish, math and English
Falch and Naper (2013)	Norway	Grade 10	Bias against boys in math and English, small bias also in Norwegian
Gibbons and Chevalier (2008)	UK	Age 14	Bias against boys in mathematics and science and against girls in English
Hanna and Linden (2012)	India	Age 7–14	No evidence of bias in exam comprising math, language skills and art section
Hinnerich et al. (2011)	Sweden	Grade 10–12 (high school)	No evidence of bias in Swedish
Lavy (2008)	Israel	Grade 10–12	Bias against boys in chemistry, computer science, math, physics, bible studies, biology, English, history, literature
Lavy and Sand (2015)	Israel	Grade 5–6	Bias against boys in English and against girls in math
Lindahl (2007a)	Sweden	Grade 9	Bias against boys in math, English and Swedish
Matějů and Smith (2014)	Czech Republic	Grade 9	Bias against boys in math and Czech language
Münich & Protivínský (2018)	Czech Republic	Grade 9	Bias against boys in math and Czech language

2013). Jurajda and Münich (2011) employ data from a pilot run of a state-wide examination on an entire cohort of Czech students finishing secondary school and applying to universities to show that – compared to men of similar general skills and subject-of-study preferences – women perform similarly well when competition is less intense, but perform substantially worse (are less likely to be admitted) when applying to very selective universities. Falch and Naper (2013) exploit variation in the importance of a national exam across different counties in Norway to investigate whether the gender grading gap only occurs when comparing teachers' grading to high-stakes exams. Their conclusion is that the gender grading gap is present to the same extent regardless of whether the anonymous exam is low-stakes or high-stakes. Lavy (2008) and Matějů and Smith (2014) compare teachers' grading with low-stake anonymous exams and still find grading bias against boys in all subjects. Based on the existing literature, it seems implausible that the gender grading gap is caused by differences in the competitiveness of the assessment environment. In this paper, we further extend the empirical evidence comparing gender bias in grading directly with students' susceptibility to stress.

2.2. Teacher-student interaction

Lavy (2008) investigates whether the gender grading gap is caused by behavioral differences on the part of the teachers or the students. His analysis proves that the gender grading gap varies substantially with teachers' characteristics (gender, age, years of experience, even family size), suggesting that the gap is highly likely caused by the teachers' behavior rather than the students'. Lindahl (2007b) shows that a student performs better if the ratio teachers of their gender at the school increases. Mechtenberg (2009) proposes a game-theoretical model explaining the gender grading gap by principal-agent framework to describe the teacher-student relationship. Falch and Naper (2013) present data that convincingly contradicts Mechtenberg's prediction but agree that it is better to consider teacher-student interaction as the explaining factor, rather than only teachers' or only students' behavior.

2.3. Non-cognitive skills

It is conceivable that grading by teachers does not only reflect cognitive skills, but also takes into account other skills such as in-class behavior and attitudes. Such non-cognitive skills do not translate into anonymously assigned exam scores. Only a few studies have taken non-cognitive skills into account. Cornwell et al. (2013) use data about children's approaches to learning, including, for example, engagement in the classroom, interpersonal skills, ability to control their own behavior, and internalization of problems while learning. They show that when controlling for non-cognitive skills, the gender grading gap in favor of girls disappeared, or in some cases even turned into a small grading bias against girls. However, their evidence is based on younger children than other studies hence it does not shed light on the role of non-cognitive skills in grading at older ages. Moreover, Burgess and Greaves (2013) conclude that classroom behavior cannot explain observed differences between anonymous and non-anonymous assessment for pupils with different ethnic backgrounds. The effect of non-cognitive traits on teachers' judgment has long been investigated in psychological literature, although the focus is typically on younger children and on the assessment of basic cognitive skills (e.g. word and letter recognition, auditory memory) rather than school exams assessing study achievements. Such literature clearly shows that the subjective evaluation can be easily confounded by unrelated personality or behavioral traits (cf. Bennett, Gottesman, Rock, & Cerullo, 1993; Pedulla, Airasian, & Madaus, 1980; Shaywitz, Shaywitz, Fletcher, & Escobar, 1990).

Grades play an important role in students' attainment perception, therefore bias in teachers' grading can impact students' future lives in a variety of ways. Students and parents typically consider grades to represent correct feedback about the student's academic performance and progress and use this information when they make further decisions about the student's educational track and career. If such feedback is biased, those decisions may be suboptimal, which could ultimately lead to labor market inefficiencies. Lavy and Sand (2015) support this hypothesis, presenting evidence that bias in teachers' grading at primary schools has an impact on students' later enrolment in advanced level math courses at high schools, which are often a prerequisite for university studies in computer science or engineering. Another recent study using longitudinal data on 5th graders in the Czech Republic documents the extent to which students rely on grades as an unbiased signal about their academic abilities when considering whether to apply to selective schools (Federičová, 2016). Conditional on the predicted probability of admission, and thus effectively controlling for academic skills, students with the best grades in both Czech language and mathematics apply to selective schools significantly more often than students who do not obtain the best grade in either Czech language or mathematics. The application rate for the top-graded students is approximately 50% higher, and this is true even in regions where admission is primarily based on test scores rather than school grades. These results suggest that students and their parents, for lack of other reliable information, consider grades a valid indicator of academic skills. A study by Breda & Hillion (2016) investigates gender differences in assessment during teacher recruitment and confirm a strong bias in favor of women in male-dominated fields (e.g. math, physics, chemistry and philosophy) and less pronounced bias in favor of men in female-dominated fields (literature and languages). This identified bias in the evaluation of professional exams indicates that the gender differences in grading may extend far beyond the educational system.

The gender bias in teachers' grading can also turn out to be a self-fulfilling prophecy, in the form of the Pygmalion effect (Hanna & Linden, 2012). Rosenthal and Jacobson (1968) found that a teacher's expectation regarding a student's intellectual growth can affect the student's intellectual development in the direction of the expectation. In other words, if a teacher expects a student to perform badly, the student does so as a direct consequence of the teacher's expectations; and vice versa. The gender grading gap thus not only provides incorrect feedback about students' academic performance at one moment but can lead to a long-term gender achievement gap.

3. Institutional settings and data

The main analysis of this study is based on a comparison of anonymous and non-anonymous evaluation scheme in a similar manner as in other empirical studies of gender bias in grading. The non-anonymous evaluation is represented by grades on a school report in the middle of 9th grade. The grades are awarded autonomously by teachers and summarise academic performance of a student in each subject in the previous academic term (half of a school year). Standardized test scores from an entrance exam to high schools are used as anonymous assessment. The entrance exam occurs towards the end of 9th grade, approximately three months after students received the school reports. To provide a better context, the following paragraphs describe briefly the educational system of the Czech Republic and continue in detail with procedural aspects of admissions to high schools, the standardized tests used for the exams and some limitations of the comparison with teachers' grades.

The Czech Republic operates a free compulsory education system for children aged 6 to 15. The first stage of compulsory education comprises five grades at elementary schools. The second stage, which corresponds to lower secondary education, lasts four years and the majority of pupils continue to attend the same school establishment for this second stage. About 10% of each age cohort, however, transfers into selective lower secondary grammar schools that offer more intensive and elaborate preparation for academic track upper-secondary level education. At the age of 15, effectively all pupils transfer from

lower- to upper-secondary education, which is provided by school establishments with three different tracks:

- The academic track focuses on preparation for tertiary education. It lasts four years and concludes with an official abitur-type exam called Maturita, which is a prerequisite for studies at tertiary level. Schools in this track are called Gymnasia.
- The vocational track offers curricula focusing more on particular skills applicable on the labor market upon graduation. It also lasts four years and concludes with a Maturita exam, therefore allowing for studies at tertiary level.
- Apprenticeship and vocational centers offer shorter courses focused on manual skills; these typically last 2–3 years, and result in an apprentice's certificate, which does not allow for further tertiary studies.

In the academic year 2015/16, 23% of students enrolled in upper secondary education entered the academic track, 47% entered the vocational track and the remainder chose various apprenticeship tracks (Vojtěch & Paterová, 2016). Most schools in the academic track experienced high excess demand for places, while the situation for vocational track schools was very varied. Admission to upper secondary schools is guided by the pupils' performance in admission exams, which are organized autonomously by each upper secondary school. Only the dates of these exams are set centrally. Each applicant can apply to a maximum of two schools and sits the entrance exams to those schools on two different days.

In 2014, many upper secondary schools outsourced a standardized test from SCIO, a company offering educational outcome assessment tools. The majority of those schools belonged to the selective academic track, offering a general curriculum. The regional authorities made such standardized SCIO tests mandatory in five out of 14 regions of the Czech Republic; as one of these five regions did not gather any information about the pupils' school of origin, we use data from the remaining four regions.[1] The exams took place on two consecutive days, hence the dataset contains two outcomes of the same type (not content) of test for many individual applicants. Out of the sample restricted to the selected four regions, 83.7% of pupils sat an exam on the first day and 67.8% pupils on the second, with about half (51.5%) sitting tests on both days. Our core analysis uses data from the first day, and we use the data from the second day for robustness checks.

The standardized SCIO exam comprises tests of Czech language, mathematics and general reasoning ability. In addition, students are asked to complete a complementary questionnaire about their attitudes toward education, their grades on their last school transcript (summarizing the students' academic performance during the first half of the current school year) and their level of self-reported stress in various situations. Most students completed the questionnaire: the average response rate was 71.6% for girls and 72.3% for boys. The scores from standardized tests constitute the anonymous assessment and are contrasted with the non-anonymous evaluation represented by reported grades from corresponding subjects. All applicants are labeled by meaningless codes for the purposes of the admission, and the test results are anonymously assessed by external evaluators who know neither the identity nor the gender of the student whose test they are marking. Moreover, the tests contain only a closed answer set of questions enabling computer processing of answers.

Contrary to admission test scores, school grades are fully in the discretion of teachers and are assigned on a discrete scale ranging from 1 (the best performance) to 5 (the worse performance). Although students are graded continuously throughout the school year, the grades on the official school transcripts issued in the middle and at the end of each school year are considered by far the most important. However, there are no explicit guidelines for assigning a particular grade. Moreover, ministerial guidelines stipulate that a grade for performance in a subject should also reflect the pupil's diligence and attitude to studying if affecting her performance (Vyhláška č. 48/2005 Sb.). This implies that when grading, teachers may also consider, intentionally or not, the pupil's behavioral and personal attitudes and traits. Nevertheless, it is widely shared wisdom that the grade reflects the pupil's academic performance during, and academic achievement at the end of the evaluated period. Since the curriculum of the last year at lower secondary school typically focuses on preparation for upper-secondary school admissions, both the transcript grades and the admission tests relate to a similar curriculum and skill set. Therefore, comparing the test scores in mathematics and Czech language with the students' latest school report grades for the same subjects enables us to identify the gap between skills (measured by the tests) and the grades awarded by teachers for the individuals concerned. The advantage of our data compared to Matějů and Smith (2014) is that ours contains scores from tests that are more closely related to the national school curricula than the international PISA tests. Moreover, students take school admission tests much more seriously than PISA tests, and we also have a larger sample of schools. On the other hand, our data do not include pupils who decided not to apply to selective schools.

In our regression analysis we use individual school fixed effects to control for school-related unobservable factors including stringency of grading. To secure a sufficient number of boys and girls from each school, only schools with at least three boys and three girls were included in the analysis.[2] We also excluded observations for 144 students that lacked information about their transcript grades.[3] Our final working sample includes 4 880 pupils from 293 schools; descriptive statistics are provided in Table 2.

4. Results

Although we focus primarily on identifying the gender grading gap, we present several descriptive results related to school grades and raw test scores first. Afterwards, we focus on the gender achievement gap and follow this with an analysis of the gender grading gap. Finally, we perform robustness checks of the results.

4.1. Test scores and grades

In primary and secondary schools, transcript grading (twice yearly) has to adhere to a discrete scale from 1 to 5, where 1 stands for the best performance. In most schools, the same grading scale is also used for regular grading throughout the year. Stringency of grading is to some extent school specific. Table 3 summarizes the relative frequency of girls and boys achieving each grade in Czech language and in mathematics. The statistics in Table 3 indicate that a notably larger proportion of girls than boys achieved the best grade (1). This difference is offset mainly by a larger proportion of boys achieving grade 3, but this is not already the case at grade 2. Grade 4 is rare, though more common in math, and grade 5 almost never appears in our data sample.

The maximum test score theoretically attainable in the SCIO school admission test is 30 points in mathematics and 40 points in Czech language. Applicants' mistakes are penalized by subtracting points,

[1] Standardized SCIO tests were made mandatory in the following regions: Liberec, Karlovy Vary, Pardubice, Ústí nad Labem and Zlín. In the Liberec region, no information was gathered about where school pupils were applying from.

[2] Small numbers of boys and girls from a given school in the dataset typically represent pupils who have applied to a school in a region different to that where their school of origin is located; we removed 1442 pupils from 493 schools. Furthermore, some students continued to vocational courses that do not require entrance exams and there are no available data about these students.

[3] Our inspections do not question the random nature of these omissions.

Table 2
Frequencies of applicants, relative frequencies of girls, and proportions of pupils applying to Gymnasia (selective upper-secondary schools, academic track).

Region	Frequency (%)	Girls %	Applicants to Gymnasia %
Karlovy Vary	730 (14.96)	50.3	18.5
Pardubice	577 (11.82)	57.2	41.4
Ústí nad Labem	1 782 (36.52)	51.2	27.3
Zlín	1 791 (36.70)	53.3	31.9
Total	**4 880**	**52.5**	**29.4**

Table 3
Relative frequency (in %) of girls and boys by transcript grades in Czech language and mathematics (grade 1 is the best).

		Relative frequency of grades				
	N	1	2	3	4	5
Czech language						
Girls	2,563	27.9	45.0	24.4	2.7	0.04
Boys	2,317	11.7	43.1	37.9	7.3	0.04
Total	4,880	20.2	44.1	30.8	4.9	0.04
Mathematics						
Girls	2,563	25.2	38.2	26.9	9.5	0.2
Boys	2,317	18.2	38.7	32.5	10.7	0.00
Total	4,880	21.9	38.4	29.5	10.0	0.1

Table 4
Summary statistics for the test scores by subject and gender.

	N	Mean	S.d.	Min	Q1	Median	Q3	Max
Czech language								
Girls	2,563	14.8	7.7	−6.7	9.3	14.7	20.0	38.7
Boys	2,317	12.8	7.3	−6.0	7.3	12.3	17.3	36.3
Total	4,880	13.9	7.6	−6.7	8.3	13.3	18.7	38.7
Mathematics								
Girls	2,563	6.9	6.2	−9.0	2.4	5.9	10.7	30
Boys	2,317	8.6	6.6	−5.0	3.4	7.6	12.8	30
Total	4,880	7.7	6.4	−9.0	2.8	6.6	12.1	30

therefore it is possible for a student to be awarded a negative overall score. Table 4 presents basic summary statistics for these test scores in our dataset, by subject and gender. It is clearly visible that boys perform on average better than girls in math, and girls perform better than their male peers in Czech language; this pattern is clear not only in the means but also in various percentiles. The gender differences in standard deviations (s.d.) are minor.

To allow for easy comparison, the analysis that now follows is based on standardized test scores (with mean 0 and standard deviation 1), so that the estimated regression coefficients express the effects in units of standard deviation.

4.2. Achievements by gender

Table 5 compares boys' and girls' achievement using a simple regression of standardized test scores on an intercept and a gender dummy only. In terms of test scores, boys are on average statistically significantly better in mathematics by 0.29 standard deviation and girls are better in Czech language by 0.25 standard deviation. These results are consistent with the findings of Matějů and Smith (2014) for the Czech Republic and show patterns similar to other countries, such as those presented by Machin and Pekkarinen (2008) who analyzed global data PISA 2003[4] and reported 15-year-old girls being better at reading by 0.30 standard deviation and boys outperforming girls in mathematics by 0.10 standard deviation.

[4] OECD Programme for International Student Assessment (PISA).

Table 5
Differences in standardized test scores in both subjects by gender. Coefficients express average differences in terms of standard deviations.

	(1)		(2)	
VARIABLES	Test score in mathematics		Test score in Czech language	
Intercept (base: boys)	0.154***	(0.016)	−0.130***	(0.015)
Girls	−0.292***	(0.030)	0.248***	(0.029)
Observations	4,880		4,880	
R-squared	0.024		0.018	
Number of schools	293		293	

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

4.3. Gender grading gap

The most prevalent method used in the empirical literature to identify gender bias in grading is a difference-in-differences estimation. In our dataset, the scales used for grading assessments by anonymous and non-anonymous evaluation are very different. Test scores represent continuous variable, while grades are of ordinal nature ranging from 1 to 5, where the weakest grades 4 and 5 occur only rarely. We estimate the gender grading gap by comparing subjective grades awarded by teachers with test outcomes. Both variables are standardized to have mean 0 and standard deviation 1, so that estimated coefficients represent effects expressed in terms of standard deviations. To foster intuitive understanding of estimated coefficients, we switch the order of grades to secure positive relationship with test scores. The form of our school-specific fixed-effects model is analogical to that used by Bonesrønning (2008) and Cornwell et al. (2013). This type of specification is more suitable than a difference-in-differences estimation given the nature of our data, and still enables us to identify inconsistency between the anonymous and non-anonymous systems of evaluation:

$$G_i = \beta_0 + \beta_1 S_i + \beta_2 W_i + \beta_3 W_i S_i + \theta_k + \varepsilon_i \quad (1)$$

where

- G_i is the standardized grade obtained by student i (either in mathematics or in Czech language),
- S_i is the standardized test score attained by student i in the corresponding subject,
- W_i is a binary variable = 1 if the student is a girl, and zero otherwise,
- θ_k are school-specific fixed-effects, for schools k = 1, …, N.

The intercept β_0 corresponds to the grade of an average boy from an average school. The coefficient β_1 is the effect of academic skills expressed in terms of grade change related with the increase of SCIO test score by one standard deviation. The coefficient of our main interest, coefficient β_2, captures the average gender difference in grades for students with average scores. It corresponds to the difference in the grades between a girl and a boy with the same skills, hence it can be interpreted as the gender bias in teacher's grading. It should be noted school-specific fixed-effects also control for different stringency in grading across schools. The coefficient on the interaction term β_3 captures possible gender difference in grading across the distribution of skills. The estimates of model (1) are summarized in Table 6.

The negative intercept β_0 indicates that boys receive in average lower grades than girls. The coefficient on test score β_1 is in line with the notion that better students are awarded higher grades. In particular,

Table 6
Estimates of the school fixed effects model (1) with grades as dependent variable and standardized test scores, gender and their interaction controlled by independent variables.

		(1)		(2)	
	VARIABLES	Grades in mathematics		Grades in Czech language	
β_0	Intercept	−0.171***	(0.0143)	−0.178***	(0.0117)
β_1	Score	0.553***	(0.0162)	0.576***	(0.0185)
β_2	Girl	0.341***	(0.0266)	0.337***	(0.0222)
β_3	Girl x score	0.115***	(0.0212)	0.0157	(0.0222)
	Observations	4,880		4,880	
	R-squared	0.351		0.388	
	Number of schools	293		293	

Robust standard errors in parentheses.
Statistical significance is based on errors clustered at school level.
All coefficients measure the effects in terms of standard deviation of grades.
**p < 0.05.
*p < 0.1.
*** p < 0.01.

one standard deviation difference in test scores yields on average 55.3% and 57.6% of standard deviation increase in grades in mathematics and in Czech language respectively. The statistically significant coefficient of our interest β_2 indicates the presence of the gender difference in teachers' grading in both subjects with similar magnitude (34.1% and 33.7% of standard deviation of grades in mathematics and Czech language respectively). It shows that teachers favor girls when awarding grades in both subjects, relative to their performance in tests. The effect of gender on teacher's grading roughly corresponds to skill difference of 60% in terms of its standard deviation in each subject. The coefficient on the interaction term β_3 represents the change in the gender difference in teachers' grading across the skill distribution. The insignificant coefficient in Czech language suggests uniform effect over the whole performance distribution, whereas positive and significant β_3 indicates substantially greater gender difference in grading among high-skilled pupils in mathematics.

4.4. Robustness of results

To investigate the robustness of our estimates, we compare them with estimates based on the data from the second admission exam day. The second exam day was mainly attended by students who had applied for two different schools. Because most students did so, there is a substantial overlap between both samples. While the batteries of test questions on the second day were different from the first day, both tests were designed in the same way, so that the total scores measure the same type of skills on both days in the same way.

The estimated coefficients in Table 7 are very similar to those presented in Table 6. The coefficient of interest β_2 capturing the gender differences in grading is again highly significant in favor of girls in both subjects. The effect in mathematics is slightly greater than that found on the first exam day (Table 6). The other coefficients are very similar to the coefficients estimated for the first exam day confirming robustness of the results.

4.5. Origins of gender differences in grading

We now turn to a summary of our further investigations into the possible origins of the estimated gender bias in teachers' grading. In particular, we explore the possible role of exam-related stress being perceived differently by boys and girls, and their different attitudes toward the subjects in question.

Stress is symptomatic of the high-stake character of school admission exams. To explore the possibly differentiated impact of stress on boys' and girls' test outcomes, we employ information provided by the

Table 7
Re-estimates of model (1) using data from the second exam day[a] (to be compared with estimates based on the first exam day in Table 6).

		(1)		(2)	
	VARIABLES	Grades in mathematics		Grades in Czech language	
β_0	Intercept	−0.204***	(0.0178)	−0.182***	(0.0155)
β_1	Score	0.548***	(0.0207)	0.591***	(0.0221)
β_2	Girl	0.393***	(0.0308)	0.332***	(0.0282)
β_3	Girl x score	0.129***	(0.0291)	−0.0238	(0.0281)
	Observations	3,017		3,017	
	R-squared	0.345		0.379	
	Number of schools	205		205	

Robust standard errors in parentheses.
Statistical significance is based on errors clustered at school level.
All coefficients measure the effects in terms of standard deviation of grades.
**p < 0.05.
*p < 0.1.
[a] The response rate on the complementary questionnaire was noticeably lower on the second day. Hence the size of the second day dataset is relatively smaller when compared with the proportions of students sitting the exam on each day.
*** p < 0.01.

tested students via a questionnaire. In the questionnaire, we asked a battery of questions about the intensity of stress the student experienced in various situations. As a proxy of perceived stress during the admission exam we use the students' self-reported intensity of stress perceived during important exams at school.

Another factor that could lead to a gender grading gap is boys' and girls' different attitudes toward particular subjects, or toward education more generally. To capture this, we utilize information from the same questionnaire, where students expressed their attitude toward a range of different subjects on a scale from 1 (the least favorite) to 10 (the most favorite).

We re-estimated model (1) on a subsample of students less susceptible to stress (experiencing stress at level 4 or lower on a scale from 1 = "no stress at all" to 10 = "very stressed") and on a subsample of students who dislike the subject in question (giving the subject score 4 or lower on a scale from 1 = "the least favorite" to 10 = "the most favorite"). Table 8 summarizes the frequencies of girls and boys in these subsamples. Girls express a higher perception of stress, dislike math more, and dislike Czech language less than boys.

A re-estimation of model (1) on the three selected subsamples investigates whether the differences in susceptibility to stress and in attitudes toward the subject concerned translate into test score differences and into the gender grading gap identified as an observed difference between anonymous and non-anonymous evaluation. These estimates are presented in Table 9 for mathematics and Table 10 for the Czech language.

The major findings remain unchanged as girls are graded more favorably in both subjects despite their lower stress and subject dislike. The gender grading differences are still present, and the corresponding coefficients are highly significant in both subsamples and in both subjects. The magnitudes of gender effects in teachers' grading are very similar to those estimated on the full sample. This reveals that the gender grading gap can hardly be attributed to a gender specific impact of test stress or to boys' and girls' differing attitudes toward a particular subject.[5]

[5] This analysis can be alternatively performed by including the dummies for susceptibility to stress and attitudes toward subjects into the main equation of model (1), together with their interactions with gender. Estimation of such model only confirms results described above, the significance and the magnitude of gender bias in grading is not affected and the interaction terms are not significant.

Table 8
Absolute and relative frequencies of girls and boys in the full sample and in three subsamples – a subsample of students less susceptible to stress, a subsample of students who dislike mathematics, a subsample of students who dislike Czech language.

	Full sample		Less stressed		Dislike mathematics		Dislike Czech language	
	Count	Percent	Count	Percent	Count	Percent	Count	Percent
Girls	2 563	100.0	897	35.0	856	33.4	445	17.4
Boys	2 317	100.0	1 057	45.6	535	23.1	772	33.3
Total	**4 880**	**100.0**	**1 954**	**40.0**	**1 391**	**28.5**	**1 217**	**24.9**

Table 9
Estimates of model (1) for mathematics on three subsamples. Column 1 – full sample, column 2 – students less susceptible to stress, column 3 – students who dislike mathematics.

		(1)		(2)		(3)	
	VARIABLES	Full sample		Less stressed		Dislike mathematics	
β_0	Intercept	−0.171***	(0.0143)	−0.073***	(0.0198)	−0.520***	(0.0331)
β_1	Score	0.553***	(0.0162)	0.499***	(0.0216)	0.503***	(0.0423)
β_2	Girl	0.341***	(0.0266)	0.327***	(0.0399)	0.452***	(0.0526)
β_3	Girl x score	0.115***	(0.0212)	0.112***	(0.0312)	0.227***	(0.0513)
	Observations	4,880		1,954		1,391	
	R-squared	0.351		0.354		0.292	
	Number of schools	293		292		282	

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

Pupils' attitudes however seem to be related with received grades in the given subject, as the substantially lower intercepts in columns (3) suggest. An exploration of the more complex relationships between test scores, grading, susceptibility to stress and attitudes towards particular subjects is beyond the capacity of the data in our possession and we therefore have to leave this for further research.

Instead of model (1) we also considered its reversed alternative having continuous test scores as dependent variable and grade dummies as explanatory variables. While the findings for mathematics were essentially the same as in case of model (1), the estimated gender grading gap in case of the Czech language was much smaller. This discrepancy between estimates turned out to be due to different distribution (relative frequencies) of grades among girls and boys, and the present model (1) properly captures that.[6]

5. Discussion

Czech 15-year-old students exhibit patterns in gender achievement gaps similar to those reported by other authors (e.g. Buchmann, DiPrete, & McDaniel, 2008; Machin & Pekkarinen, 2008; Matějů & Smith, 2014). We find that boys outperform girls in mathematics by 0.29 standard deviation on average, whereas they lag behind girls in Czech language by 0.25 standard deviation. The Czech gender gap in test scores in mathematics is slightly larger than the global average 0.10 standard deviation reported by Machin and Pekkarinen (2008).

Furthermore, our analysis identified the presence of sizeable gender gap in teachers' grading in favor of girls both in mathematics and in Czech language. The gender effect amounts to 34% of standard deviation of grades and it corresponds to a difference in skills at 60% of standard deviation. Our further investigations revealed that the gender grading gap in Czech language is uniform over the whole performance distribution, while it is larger for high performing students in mathematics. Moreover, it can be explained neither by differences in the pupils' perception of stress nor in their attitudes towards math.

The gap we have identified could have important implications for the choice of higher school. Although we cannot investigate it with our data, we can refer to Federičová (2016), who used rich longitudinal data on a cohort of Czech pupils participating in the international survey TIMMS & PIRLS 2011 showing that a significant part of the gender difference in school application rates is due to unexplained gender differences in school grading, even in districts where grades have only low weight in the admission mechanism.

The findings presented in this paper are in line with the majority of studies investigating gender grading gaps. Out of the studies presented in Table 1, nine studies found a significant grading bias in favor of girls in mathematics and the same number also confirmed grading bias in languages. Of those studies that did not find any gender grading bias, Hinnerich et al. (2011) did not include an assessment of mathematical skills at all, while Hanna and Linden (2012) used only combined exams consisting of math, language skills and art sections. In the latter case, the research is based on slightly different methodology where there were no personal ties between the evaluators and the students, which might explain the absence of gender grading gap. The most surprising findings are that of Lavy and Sand (2015) who present evidence of gender grading gap in favor of boys in mathematics and of Gibbons & Chevalier who document gender bias against girls in English; however, the former study is based on a younger cohort of students, aged 10–11 years.

Our approach has several limitations. The two pupils' assessments, teachers' grading and SCIO testing, were performed at different times and so the gender grading gap we have estimated may be contaminated by different attitudes of boys and girls to studying during the period between their grades being awarded and them sitting the tests (about 3 months). In this respect, however, Lavy (2008) documents that the gender grading gap cannot be explained by different assessment timing. Another possible limitation is that the curriculum tested in the admissions tests we used might differ to some extent from the curricula taught (and graded) in school. While this was surely the case for the international PISA test data employed by Matějů and Smith (2014), the

[6] We thank to one of the anonymous referees for very valuable insight helping us to capture this.

Table 10
Estimates of model (1) for Czech language on three subsamples. Column 1 – full sample, column 2 – students susceptible to stress, column 3 – students who dislike Czech language.

		(1)		(2)		(3)	
	VARIABLES	Full Sample		Less stressed		Dislike Czech language	
β_0	Intercept	−0.178***	(0.0117)	−0.165***	(0.0161)	−0.384***	(0.0218)
β_1	Score	0.576***	(0.0185)	0.598***	(0.0254)	0.510***	(0.0400)
β_2	Girl	0.337***	(0.0222)	0.332***	(0.0374)	0.334***	(0.0540)
β_3	Girl x score	0.0157	(0.0222)	−0.0101	(0.0349)	0.0805	(0.0559)
	Observations	4,880		1,954		1,217	
	R-squared	0.388		0.423		0.326	
	Number of schools	293		292		277	

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

domestic SCIO tests we used are notably closer to what is being taught in schools in the Czech Republic. Moreover, the teaching in the last year or two at lower-secondary schools is heavily focused on preparing students for high school admission exams. Needless to say, comparison of two different assessments is a weakness common to most existing studies estimating gender grading gaps. The only exception is a study by Hinnerich et al. (2011), which is based on repeated evaluation of the same exam sheets.

We have not found any evidence that students' attitudes toward subjects play a role in the gender grading gap. However, contrary to Cornwell et al. (2013) who could control for non-cognitive skills,[7] we can only rely on the students' self-reported attitudes toward subjects. Due to data limits, we also cannot unambiguously identify whether the gender grading gap is caused by teachers' characteristics, interaction between teachers and students or students' non-cognitive skills being reflected in teachers' grading. However, it should be kept in mind that the role of non-cognitive skills and their interaction with the gender grading gap is a complex issue. As Bennett et al. (1993) pointed out, whether non-cognitive skills are considered a threat to the validity of teachers' academic judgments depends primarily on the definition of academic judgment. Some teachers might, either intentionally or unintentionally, include these skills in their assessment because they consider students' behavior in class to be part of their academic performance. This conjecture would also explain why Hanna and Linden (2012) did not find a gender grading gap when they compared blind and non-blind evaluation of the same exam sheets without any personal ties existing between the students and teachers involved. Many other studies investigating gender grading gaps compare a one-day anonymous exam with teachers' overall assessment of the student over a longer period (e.g. Angelo, 2014; Bonesrønning, 2008; Falch & Naper, 2013; Lavy, 2008), where it is more likely that teachers would take into account the students' general approach toward learning and their in-class behavior, although this might be in contradiction to educational policy in some countries. For example, in Norway there are national guidelines that explicitly state that grades should only reflect achievement during the academic year and not students' effort. DiPrete and Jennings (2012) show that students' social and behavioral skills have a twofold impact on the learning process and educational outcomes. Part of the effect is reflected through educational assessment that typically rewards students who conform to social norms, while part is due to the positive impact of social and behavioral skills on the learning process itself.

[7] Cornwell et al. (2013) were able to control for non-cognitive skills such as students' general approach toward learning, their engagement in the classroom, their interpersonal skills, their ability to control their own behaviour and their internalization of problems during learning.

The presence of a sizeable gender effect on teachers' grading in both mathematics and native language in the Czech Republic might be explained by ministerial guidelines stipulating that transcript grades for performance in a particular subject can also reflect a pupil's diligence and attitude to studying if they affect her performance. Another possible explanation for our findings is that teachers grade girls more favorably in mathematics to improve their future performance, in accordance with the self-fulfilling prophecy and Pygmalion effect (Rosenthal & Jacobson, 1968). However, in that case there should be a gender grading gap favoring boys in Czech language, driven by the gender achievement gap in Czech language. But we do not find this to be present.

Future research should focus in greater detail not only on the relationship between the gender grading gap and non-cognitive skills, but also on the role of non-cognitive skills and behavior in the educational process in general, as these represent a crucial precondition for school success. If gender bias is indeed present in school grading systems, this may have important implications in selective schooling systems, such as the Czech one, where school grades play important role in admission to selective schools. It also may affect pupils' expectations and aspirations when making decisions about whether to apply to selective schools and compete with others, as documented by Federičová (2016). In such systems, gender differences in teachers' grading could have substantial impact on pupils' careers and eventually petrify gender differentials at adult ages.

Acknowledgement

We would like to thank Ondřej Šteffl and Jan Hučín from the Scio agency for kind data provision and assistance. We feel indebted to two anonymous referees who provided us with very helpful advice improving the quality, clarity and persuasiveness of the final manuscript. Last but not least we appreciate very much research inspiring personality of Petr Matějů.

References

Angelo, C. S. (2014). *Is there a bias towards girls in non anonymous evaluation?* Unpublished work.
Bennett, R. E., Gottesman, R. L., Rock, D. A., & Cerullo, F. (1993). Influence of behavior perceptions and gender on teachers' judgments of students' academic skill. *Journal of Educational Psychology, 85*(2), 347.
Bonesrønning, H. (2008). The effect of grading practices on gender differences in academic performance. *Bulletin of Economic Research, 60*(3), 245–264.
Breda, T., & Hillion, M. (2016). Teaching accreditation exams reveal grading biases favor women in male-dominated disciplines in France. *Science, 353*(6298), 474–478. http://science.sciencemag.org/content/353/6298/474.
Buchmann, C., DiPrete, T. A., & McDaniel, A. (2008). Gender inequalities in education. *Annual Review of Sociology, 34*, 319–337.
Burgess, S., & Greaves, E. (2013). Test scores, subjective assessment, and stereotyping of

ethnic minorities. *Journal of Labor Economics, 31*(3), 535–576.
Cornwell, C., Mustard, D. B., & Van Parys, J. (2013). Noncognitive skills and the gender disparities in test scores and teacher assessments: Evidence from primary school. *The Journal of Human Resources, 48*(1), 236–264.
DiPrete, T. A., & Jennings, J. L. (2012). Social and behavioral skills and the gender gap in early educational achievement. *Social Science Research, 41*(1), 1–15.
Emanuelsson, I., & Fischbein, S. (1986). A study on sex and schooling. *Scandinavian Journal of Educational Research, 30*(2), 71–84.
Falch, T., & Naper, L. R. (2013). Educational evaluation schemes and gender gaps in student achievement. *Economics of Education Review, 36*, 12–25.
Federičová, M. (2016). *Gender gap in application to selective schools: Are grades a good signal?* CERGE-EI Working Paper, No. 550, September 2015.
Gibbons, S., & Chevalier, A. (2008). Assessment and age 16+ education participation. *Research Papers in Education, 23*(2), 113–123.
Gneezy, U., Niederle, M., & Rustichini, A. (2003). Performance in competitive environments: Gender differences. *The Quarterly Journal of Economics, 118*(3), 1049–1074.
Guiso, L., Monte, F., Sapienza, P., & Zingales, L. (2008). Culture, gender, and math. *Science, 320*(5880), 1164–1165.
Hanna, R. N., & Linden, L. L. (2012). Discrimination in grading. *American Economic Journal Economic Policy, 4*(4), 146–168.
Hinnerich, B. T., Höglin, E., & Johannesson, M. (2011). Are boys discriminated in Swedish high schools? *Economics of Education Review, 30*(4), 682–690.
Jurajda, Š., & Münich, D. (2011). Gender gap in performance under competitive pressure: Admissions to Czech universities. *The American Economic Review, 101*(3), 514–518.
Lavy, V. (2008). Do gender stereotypes reduce girls' or boys' human capital outcomes? Evidence from a natural experiment. *Journal of Public Economics, 92*(10), 2083–2105.
Lavy, V., & Sand, E. (2015). *On the origins of gender human capital gaps: Short and long term consequences of teachers' stereotypical biases.* NBER Working Paper 20909, Cambridge, MA.
Lindahl, E. (2007a). *Comparing teachers' assessments and national test results: Evidence from Sweden.* Working PaperIFAU-Institute for Labour Market Policy Evaluation.
Lindahl, E. (2007b). *Gender and ethnic interactions among teachers and students: Evidence from Sweden.* Working PaperIFAU-Institute for Labour Market Policy Evaluation.
Machin, S., & McNally, S. (2005). Gender and student achievement in English schools. *Oxford Review of Economic Policy, 21*(3), 357–372.
Machin, S., & Pekkarinen, T. (2008). Global sex differences in test score variability. *Science, 332*(5906), 1331–1332.
Matějů, P., & Simonová, N. (2013). Koho znevýhodňuje škola: chlapce, nebo dívky? Rozdíly v dovednostech, školních výsledcích a vzdělanostních aspiracích dívek a chlapců devátých tříd základních škol [who is shortchanged in school: boys or girls? Gender gaps in skills, grading and educational aspirations among pupils of 9th grades of elementary schools]. *Orbis Scholae, 7*(3), 107–138.
Matějů, P., & Smith, M. (2014). Are boys that bad? Gender gaps in measured skills, grades and aspirations in Czech elementary schools. *British Journal of Sociology of Education, 36*(6), 871–895.
Mechtenberg, L. (2009). Cheap talk in the classroom: How biased grading at school explains gender differences in achievements, career choices and wages. *The Review of Economic Studies, 76*(4), 1431–1459.
OECD (2015). *The ABC of gender equality in education.* OECD Publishing.
Ors, E., Palomino, F., & Peyrache, E. (2013). Performance gender gap: does competition matter? *Journal of Labor Economics, 31*(3), 443–499.
Pedulla, J. J., Airasian, P. W., & Madaus, G. F. (1980). Do teacher ratings and standardized test results of students yield the same information? *American Educational Research Journal, 17*(3), 303–307.
Rosenthal, R., & Jacobson, L. (1968). Pygmalion in the classroom. *The Urban Review, 3*(1), 16–20.
Shaywitz, S. E., Shaywitz, B. A., Fletcher, J. M., & Escobar, M. D. (1990). Prevalence of reading disability in boys and girls. *JAMA the Journal of the American Medical Association, 264*(8), 998–1002.
Straková, J., Potužníková, E., & Tomášek, V. (2006). Vědomosti, dovednosti a postoje českých žáků v mezinárodním srovnání [knowledge, skills and attitudes of Czech pupils in international comparison]. In P. Matějů, & J. Straková (Eds.). *Ne)rovné šance na vzdělání, vzdělanostní nerovnosti v České republice [(Un)equal opportunities in education, educational disparities in the Czech Republic]*. Prague: Academia.
Vojtěch, J., & Paterová, P. (2016). *Vývoj vzdělanostní a oborové struktury žáků a studentů ve středním a vyšším odborném vzdělávání v ČR a v krajích ČR a postavení mladých lidí na trhu práce ve srovnání se stavem v Evropské unii 2015/16 [Trends in educational and professional structure of pupils and students in secondary and higher professional education in the Czech Republic and its regions and status of young people in the labor market in comparison with the European Union]*. Prague: Národní ústav pro vzdělávání [National Institute for Education].
Vyhláška č. 48/2005 Sb (2005). *o základním vzdělávání a některých náležitostech plnění povinné školní docházky [Notification no. 48/2005 Coll. concerning elementary education and selected terms of fulfilment of compulsory education]. Sbírka zákonů České republiky [Collection of Law of the Czech Republic]*. Prague: Ministry of Education, Youth and Sports of the Czech Republic ch. 48, pp. 319–327.