Beatrice Schindler Rangvid

01:2018 WORKING PAPER

GENDER DISCRIMINATION IN EXAM GRADING?

DOUBLE EVIDENCE FROM A GRADING REFORM AND A FIELD EXPERIMENT

VIVE – DANISH CENTRE OF APPLIED SOCIAL SCIENCE

GENDER DISCRIMINATION IN EXAM GRADING?

DOUBLE EVIDENCE FROM A GRADING REFORM AND A FIELD EXPERIMENT

Beatrice Schindler Rangvid

DANISH CENTRE OF APPLIED SOCIAL SCIENCE

COPENHAGEN, DENMARK

Working Paper 01:2018

GENDER DISCRIMINATION IN EXAM GRADING? DOUBLE EVIDENCE FROM A GRADING REFORM AND A FIELD EXPERIMENT [*]

Beatrice Schindler RANGVID
The Danish Centre of Applied Social Science (VIVE)

December 2017

Abstract
Girls, on average, obtain higher test scores in school than boys, and recent research suggests that part of this difference may be due to discrimination against boys in grading. This bias is consequential if admission to subsequent education programs is based on exam scores. This study assesses the causal effect of blind grading, exploiting two separate identification strategies. The first derives from a unique full cohort natural experiment with a grading reform, providing exogenous variation in blind grading. The other strategy derives from a field experiment where the exact same exam papers are scored twice (blind and non-blind). Both strategies use difference-in-differences methods. Although imprecisely estimated, the point estimates indicate a blind grading advantage for boys in essay writing of approximately 5-8% SD, corresponding to 9-15% of the gender gap in essay exam grades. The effect appears to be more pronounced among low performers. Moreover, evaluators tend to give higher grades to boys' essays when they are led to believe these essays were written by girls. Additional analyses for math suggest a (poorly determined) blind grading effect in favor of girls of 1-3% SD. The overall tendencies are in accordance with statistical discrimination as a mechanism for grading bias in essay writing and with gender-stereotyped beliefs of math being a male domain.

Keywords: discrimination; gender bias; education economics; difference-in-differences
JEL Classification: I20

[*] Contact information: bsr@vive.dk; The Danish Centre of Applied Social Science (VIVE), Herluf Trolles Gade 11, 1052 Copenhagen K, Denmark. The usual disclaimer applies. Thanks to the Ministry of Education for giving me access to the experimental data. I thank Paul Bingley, Vibeke Myrup Jensen and C. Kirabo Jackson for valuable comments on earlier versions of this paper. The paper has also benefited from comments by seminar participants at the University of Copenhagen and The Danish National Centre for Social Research (SFI), as well as by participants at the ESPE 2017 conference.

1. Introduction

Student assessments such as tests and exams are used in countries all over the world to monitor student performance, inform students about their academic ability and guide enrolment in different tracks and educational programs. Research on the evaluation of students, however, suggests that exam grades are not invariably unbiased measures of student achievement. Investigating grading bias is important, because test scores are the main indicator of ability and performance that students receive in school. Discrimination in grading could have long-lasting effects by reinforcing erroneous beliefs of inferiority and by discouraging students from making human capital investments.[1] The perceived fairness of grading is likely to affect students' motivation, self-confidence, longer-term school outcomes and, thereby, future employment perspectives. A number of studies highlight that non-blind grading[2] of tests and exams induces grading bias against specific groups of students, such as boys and ethnic minority students (Lavy, 2008; Hinnerich et al., 2011, 2015).[3] Grading bias may be a factor contributing to these educational achievement gaps in gender and ethnicity.

Sources of grading bias may derive from personal ties or from statistical discrimination and stereotyping. When teachers grade their own students' exams, personal ties between the teacher and student may affect the evaluation of exams. Even when exams are graded by an external grader, however, expectations may affect the way in which examiners perceive and grade student performance. For example, if exam papers reveal students' names or in other ways reveal the gender of the student, this may directly influence the test scores given by graders via their priors about boys' performances and through gender stereotypes and attitudes. Given the existing performance gap between boys and girls in, e.g., the humanities, graders may expect individual male students to have lower skills on average. On the other hand, long-standing beliefs about math and science being male fields might produce bias against girls in these subjects. The underlying hypothesis is that the stereotypes and image of the student that this calls forth in the grader affect perceived student performance and thus test scores. It is unclear, however, how these expectations affect grading. Graders may give a higher score than they would have otherwise if the observed performance surpasses their expectations or if they want to provide encouragement. On the other hand, they may give boys lower scores in the humanities if low expectations prevent them from recognizing performance. Also, negative attitudes toward some groups (perhaps due to more-disruptive behavior in the case of boys) may lead graders to give biased scores.

A way to avoid grading bias due to personal ties, statistical discrimination and stereotyping is to blind the grading procedure. The general literature on blind assessment dates back to the 1990's in studies examining submission to scientific journals or conferences (Blank, 1991; Carlsson et al.

[1] Note that expectations and how they affect grading are probably not due to deliberate discrimination of students by the examiners, but beliefs and expectations may unintentionally affect behavior and lead to implicit bias.

[2] Under non-blind grading, the exam paper is marked with students' names or other information that allows the grader to deduce the gender or ethnic background of the student. Blind grading means that the grader does not have information from which he/she can deduce the student's gender or ethnic background.

[3] Hinnerich et al. (2011), however, do not find evidence of gender grading bias.

2012) and orchestra auditions (Goldin & Rouse, 2000).[4] Specifically in the field of education, there are several strands of literature on blind grading. One strand compares non-blind course grades with blind (or external) exam grades (e.g., Burgess & Greaves, 2013; Cornwell et al., 2013; Falch & Naper, 2013; Rangvid, 2015) and generally finds that educationally disadvantaged groups (boys and immigrants) are disadvantaged by teacher course grades.[5] Other studies have examined grading bias by comparing scores from school exams graded under different grading procedures. Lavy (2008) examined grading bias in school exams in a natural experiment setting comparing test scores for the same students from two school exams that use different grading procedures: a non-blind score of an internal school exam and a blind score from a similar external exam with blind assessment. He finds statistically significant discrimination against boys under non-blind grading in all examined subjects. In a field experiment for Sweden, Hinnerich et al. (2011) carried out a study that compares blind and non-blind scores by gender for the exact same exam papers by analyzing a random sample of essay exams that were graded twice using different procedures: non-blind (as part of the national exam) and blind (as part of the scientific study).[6] They find no evidence of gender bias.

To my knowledge, this study provides the first large-scale natural experiment evidence where blind and non-blind grading is assessed for the *same* test. The main identification approach used in this study assesses blind and non-blind grading at the same exam, thus holding constant across treated and untreated students all other dimensions that might bias the estimate when the blind and non-blind scores come from different exams, e.g., the examination questions, the stakes of the exam (high/low), or the type of grader (own teacher/external grader). Identification comes from a grading reform that quasi-randomly assigns students to treatment. While previous studies that use large-scale natural experiment methods produce *within*-student estimates, this approach exploits *between*-student variation in treatment assignment. Although randomization should take care of selection issues between students, I take additional steps to alleviate remaining concerns. First, I include controls for ability and socio-economic background to account for remaining selection. Second, I exploit a field experiment to provide additional evidence on the effect of blind grading. The field experiment is based on a random sample from the same examination and cohort as used for the reform approach and provides within-student estimates. This approach exploits a different (and independent) source of identification by comparing blind and non-blind scores for the exact same exam papers. The results from both approaches combined provide particularly strong causal evidence.

[4] Moreover, there is a more general literature examining discrimination in economics (Ayres & Siegelman, 1995; Neumark et al., 1996; Ladd,1998; Szymanski, 2000; Bertrand & Mullainathan, 2004; Riach & Rich, 2006; Petit, 2007; Lahey, 2008; Carlsson & Eriksson, 2017).

[5] Some studies investigate grading bias in fully experimental settings (Van Ewijk, 2011; Hanna and Linden, 2012; Sprietsma, 2013) to assess the effect of students' supposed origin (immigrant background or caste) on scores in essay writing. These studies randomly assign immigrant and native (or high and low caste) first names to a set of essays, thus making some teachers believe a given essay was written by a native (or high caste) student, while others believe it was written by an immigrant (or low caste) student. Results are mixed.

[6] Note that since the non-blind assessments in Lavy's and Hinnerich et al.'s studies are carried out by the student's own teacher, the effect of blind grading cannot be estimated net of the effect of teacher grading.

The natural experiment derives from a grading reform that provides exogenous variation in assignment to blind grading. The objective of the reform was to implement blind grading in core subjects of the school-leaving examination in Denmark by replacing the student's full name on the exam paper with a student identification number. However, the procedure was flawed. For convenience, preexisting student numbers were chosen. Since these are not random numbers but include the first four characters of the student's (first) name, truncation after four characters may or may not conceal the gender of the student – depending on the name (see section 3 for details)[7]. In this setting, only some students are affected by the reform, because students with student identification numbers (hereafter: SN) that reveal the gender are still graded non-blind – just as before the reform. Consequently, the reform does not blind the grading process for all students but de facto produces an arguably random assignment to blind grading. This flaw in the blinding procedure provides a unique opportunity to identify the effect of blind grading within the same examination in a full-cohort natural experiment.

The second identification strategy provides within-student estimates and comes from a field experiment where a random sample of the same exam papers are subject to blind and non-blind grading within the same empirical framework as the reform-based approach. The non-blind grades are the original grades from the school-leaving examination (for students with identification numbers that clearly reveal gender). The blind grades are obtained from regrading the very same exam papers with any identifying information removed.

The main contribution of this study is that it is the first to provide evidence on the effect of blind grading from a (full-cohort) natural experiment for the same exam. Another particular strength of this study is that it provides evidence from two independent sources of identification within the same empirical framework: one providing estimates from a large-scale natural experiment within the same school leaving examinations and the other providing within-student estimates from a field experiment. The availability of the combined evidence from two rigorous identification methods provides particularly strong causal evidence.

The main focus in this study is on essay writing, because there is probably more room for discretion in grading essays than in grading math exams. I also provide results for math. To preview the results, I find that – although imprecisely estimated – the point estimates indicate a blind grading advantage for boys in essay writing of approximately 5-8% of a standard deviation, corresponding to 9-15% of the gender gap in essay exam grades. The effect is more pronounced among low performers. Moreover, evaluators give higher grades to boys' essays when they are led to believe these essays were written by girls. An additional analysis reveals a (poorly determined) blind grading effect in favor of girls of 1-3% SD in math. The results for essay writing are in accordance with statistical discrimination as a driver of grading bias, while the results for math are compatible with the notion of gender-stereotyped beliefs of math being a male domain.

The paper proceeds as follows. Sections 2-5 describe the institutional background, identification strategy, empirical model/data and results from the reform-based approach to identification.

[7] In theory, this strategy could also be used to identify ethnic bias. However, a preliminary analysis has shown that too few immigrant students have student numbers that effectively conceal their immigrant background (see footnote 18 for further explanation).

Section 6 presents results from the field experiment, and the last section concludes.

2. Institutional background

After ten years of compulsory education in a comprehensive school system, students in Denmark sit school-leaving exams. The school-leaving exams consist of a set of mandatory tests and a small number of voluntary tests. All students take exams in Danish, math, English and science. The mandatory parts of the exams in English and science are oral exams.[8]

Until 2014, the school-leaving examinations were low-stake for the students. Beginning in 2015, admission to vocational education and training has been contingent on passing the exams in the core subjects of Danish and math. Furthermore, from 2019, admission to high-school programs will depend on performance at the school-leaving examinations. Thus, exam scores from the school-leaving examinations have become consequential for admission to upper secondary programs.

Until 2015, exams in essay writing and math (problem solving) and the written exams in foreign languages (English, German, French) were graded jointly by the student's teacher and an external grader (a centrally appointed teacher from another school), and students put their full names on the exam papers. Thus, the exam was graded partly by a grader who personally knew the student from class (the teacher) and an external grader who could infer student gender via the name on the exam paper. Knowing the student personally from class teaching (personal ties) but also just being aware of students' gender (statistical discrimination, stereotyping) have been shown in the literature to be potential sources of grading bias.

To eliminate these sources of bias, a grading reform in 2016 introduced two changes to the procedure. First, beginning in 2016, grading is performed by an external grader alone, eliminating potential bias from personal ties between the teacher and the student. Second, to eliminate bias due to stereotyping, an attempt was made to blind the grading procedure by replacing student names on the exam papers with student identification numbers. To replace students' names on the exam papers, the Ministry of Education chose a preexisting set of identification numbers that were in use already for other school-related IT purposes.[9] These identification numbers were chosen partly for convenience but also because these numbers were only semi-anonymous, which reduced the perceived risk of mixing up student identities in the grading process.[10]

[8] The remaining two mandatory exams are decided by the Ministry of Education at the class level: one exam from the science group, the other from the humanities. For each class, the Ministry of Education draws one test subject from the science group (i.e., either biology or geography) and one from the humanities (English (written), Christian studies (oral), history (oral), social studies (oral) and German or French (written or oral). Tests in each of the five subjects within the humanities are drawn evenly across classes, such that each subject is covered by approximately 20% of a student cohort. A similar procedure applies to the two subjects within the science group, which each are taken by approximately 50% of the students.

[9] The student identity number is a username *('UNI-login')* that students use to access, e.g., the platform administering digital tests that are part of the school-leaving examinations (e.g., a multiple choice test in geography), the national tests and the school's intranet.

[10] This part of the reform is a temporary solution and will be replaced by a digital exam management solution that will implement a fully blinded grading procedure (planned to be launched in 2019).

Taken together, the objective of these changes in the grading scheme was to ensure full objectivity in the grading of the exams. Having the exams scored by an external grader effectively removes any local component from the grading procedure (i.e., removes the potential influence of personal ties between the grader and the student).[11] Using the student numbers, however, might be less successful at removing bias due to statistical discrimination or stereotyping, because, due to a flaw in the blinding procedure, the reform does not effectively conceal the gender of all students.

While the reform is targeted at specific exams[12], these changes are part of a wider agenda to make the grading procedure of all written exams fully external and blind. For example, exams in geography and biology are computer-based multiple choice exams that are automatically scored. While multiple choice exams lend themselves easily to computerized scoring, the reform is targeted at types of exams where automatic scoring is not feasible or desirable, such as essay writing or math problem solving.

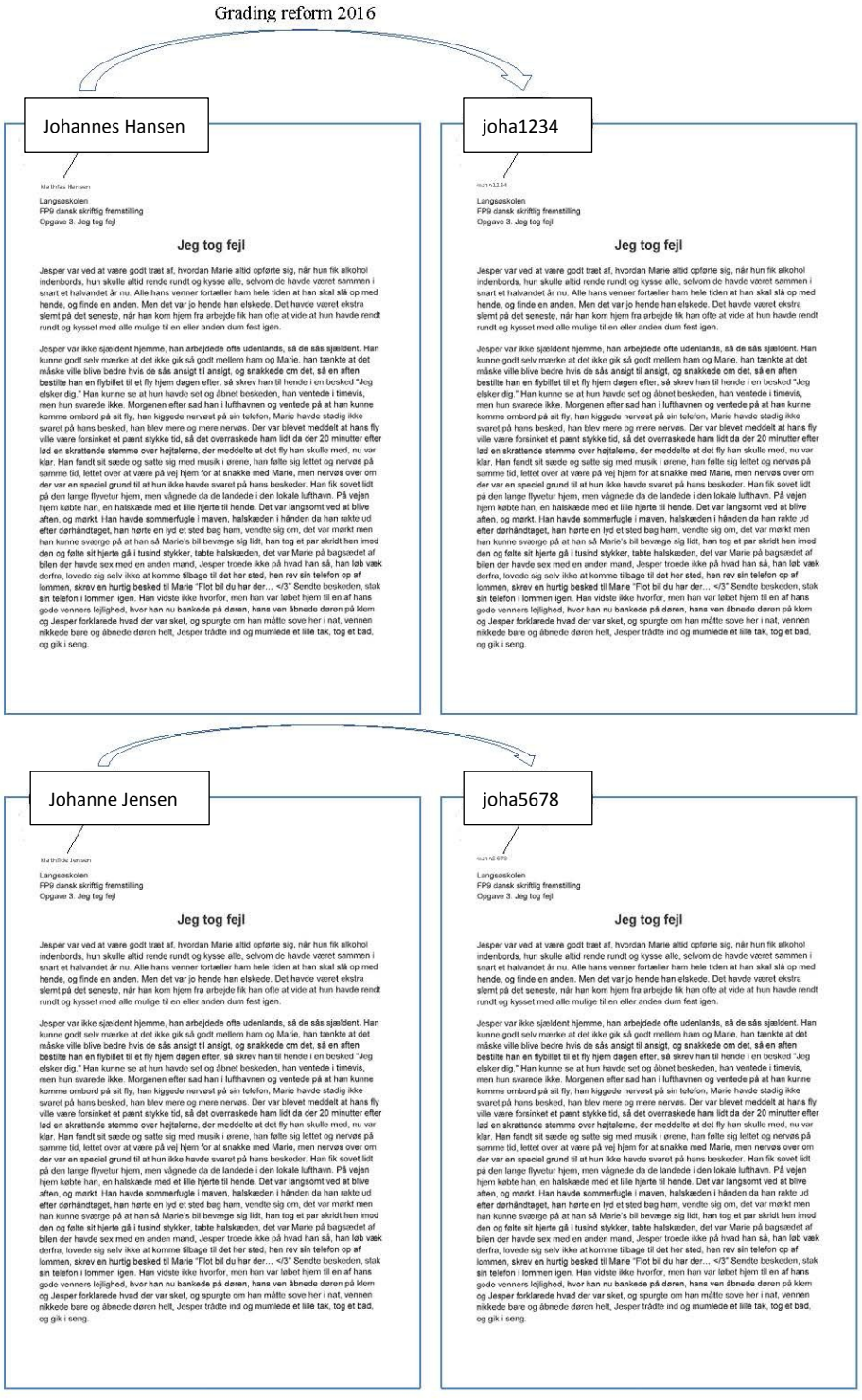
3. Identification strategy

This section describes how the student identification number provides the quasi-random assignment to treatment that is crucial for identifying the causal effects of blind grading. As mentioned in section 2, from 2016 onwards, students no longer write their names on the exam papers but instead mark exams with their student identification number. These numbers had already been in use for various digital school-related tasks. The SNs begin with four characters followed by four digits. The characters correspond to the beginning of the student's first name, e.g., *joha* for boys named Johannes. The four digits that follow do not hold any information related to the student's name or gender. They simply serve to create a unique student identifier. For example, a boy named Johannes Hansen would, until 2015, have submitted his exam paper marked with his full name, but from 2016, he would mark the exam with his SN, e.g., *joha1234* (see Figure 1).

[11] See section 3 for details. To investigate whether this part of the reform could be used provide causal evidence of the effect of external grading (vs. grading by students' own teachers), I considered using a difference-in-differences design comparing exam scores of boys and girls before and after the reform (of boys and girls whose gender is not blinded by the reform). However, I carried out preliminary analyses and found that the gender test score gap is not stable in the years before the reform, which would be crucial for the validity of the identification strategy.

[12] The changes in the grading scheme are targeted at specific exams: essay writing (Danish), problem solving (math), and the written exams in foreign languages (English, German, French). All students sit exams in the core subjects of Danish and math. Assessment in both subjects is divided into several parts. In Danish, essay writing, reading, spelling and oral skills are assessed and graded separately. In math, separate grades for problem solving and math proficiency are given. While all students sit exams in the core subjects essay writing and problem solving, the other exams targeted by the reform are administered by random draw to a subsample of classes.

Figure 1: Marking of exam papers before and after the reform



Obviously, using the student identification number conceals only part of the student's name. The crucial question is whether truncating the name after 4 characters is sufficient to conceal students' gender such that the exam grader cannot know whether the exam is submitted by a boy or a girl.[13] Thus, if a student's gender is not revealed by the four characters, the blinding procedure works as intended. For example, consider again the student Johannes Hansen with the student identification number *joha1234*. When Johannes' exam paper is only marked with his student identification number and not with his name, the exam grader cannot know whether the exam is written by a boy named Johannes or a girl named Johanne, because the four characters in their names are identical, and consequently, the student identification numbers of girls named Johanne look exactly like those of boys named Johannes (see Figure 1). In this case, exam graders are not able to infer the gender of the student from the SN, i.e., marking the exam with the SN effectively conceals the student's gender. The same is true for other names, e.g., Mathias/Mathilde and Nicole/Nicolaj. Thus, in such cases, the blinding procedure works as intended. Note that students type the exams on computers. Thus, unlike with handwritten exams, students' handwriting does not enable graders to predict the gender of the student.

However, this is not always the case. Consider, for instance, another common name: Caroline. Her SN could be *caro2947*. Since male names starting with *caro* do not exist (at least in Denmark[14]), exam graders will be able to infer that this specific exam is submitted by a girl. This is similarly the case for SNs beginning with *clar* (Clara) or *mari* (Marie or Maria). For boys, SNs beginning with *jaco* (Jacob) or *marc* (Marcus) clearly reveal that the student is a boy. SNs like these are exclusively, or overwhelmingly, used by a specific gender. These examples show that the SN blinds gender for some students but not for others. Thus, whether the new procedure actually blinds student identification with respect to gender depends on (the beginning of) students' names.

However, the treatment (blinding with the SN) can only have an impact on grading if the exam graders (i) realize that the SN may hold information on gender and if they are able to (ii) correctly infer the gender of the student. First, for the treatment to be able to affect grading, graders must be aware that the SN may hold information about gender, i.e., they must know that the first four characters of the SN are the beginning of the student's first name. We can safely assume that this is the case, because the exam graders are teachers (from other schools), and teachers happen to have a user-name/teacher-number for use with school-related IT that is constructed in exactly the same way as the student number.

Second, graders need to be able to correctly infer the gender of the students from the SN. A survey among all external graders provides evidence that this is indeed the case. In the survey,

[13] Since it is not the student's own teacher but an external grader who is grading the exam, the only information the grader can infer from the student's first name is the gender of the student. Because the grader does not personally know the student, grading bias can only arise from statistical discrimination or stereotyping (not personal ties). Thus, teachers' beliefs about girls performing, on average, better than boys may affect non-blind grading results, while teachers' knowledge about the individual student's skills or behavior cannot.

[14] A calculation shows that 100% of students with first names beginning with *caro* are girls (in this specific student cohort in Denmark).

graders were asked whether they could infer student gender from the SN when they graded the exam papers. The overwhelming majority answered that they could infer the gender of the student in more than 75% of the cases. Based on the questions from the survey, we cannot know whether graders always indeed *correctly* identified student gender. Yet, graders can use their experience from interacting with students on a daily basis to guess correctly in most cases. Thus, at least for a large number of students, it is probably safe to assume that they could.

Thus, I assume that exam graders are able to judge whether the SNs most likely refer to a male or female student or whether both are about equally likely. For use in the empirical analysis, I construct three categories of SNs: one including SNs that will lead the grader to believe the exam was written by a boy, one for SNs associated with girls and, finally, a category including SNs that do not reveal the gender of the student. I propose to use the empirical gender distribution for each four-letter combination of the SN to form these categories.[15] Thus, using student-level data from administrative registers, I calculate the share of boys among students with this particular SN for each four-letter combination. I use this variable as an approximation of the grader's gender perception of each four-letter SN.

As an example, Table 1 shows the empirical share of male students for each SN for ten common four-letter combinations. The three SNs in the upper panel (*anto*, *marc*, *jaco*) are SNs with a male share of (close to) 100%, and they consequently refer predominantly to male names (Anton, Marcus, Marc and Jacob). I term SNs with a high share of male students as *male-specific* SNs. When graders see exam papers with such SNs, it is highly likely that they will perceive the paper as being written by a male student. Common SNs in the corresponding category for overwhelmingly female SNs include *caro, clar* and *mari*, which in most cases refer to female names (Caroline, Clara, Marie/ Maria).[16] I term SNs with a high share of female students as *female-specific* SNs. Apart from the clearly gender-specific SNs, there is a range of SNs shared almost equally by boys and girls. These SNs have an empirical male share of approximately 50% (e.g., *nico*, *math*, *joha*). I term these SNs *gender-neutral*. When the graders see exam papers marked with such SNs, they will not know whether the paper is written by a boy or a girl. Thus, students with first names with gender-neutral SNs have their gender blinded by the use of the SN compared with when they put their full name on the exam paper.

[15] This might not correctly model each single grader's perception, but on average, this should provide a consistent estimate.

[16] In principle, this identification strategy could also be used to investigate bias by ethnicity. However, it turns out that immigrant students are not quasi-randomly allocated to the 'ethnicity-blinding treatment' with respect to their *country of origin*. (They are, however, randomly allocated with respect to the 'gender-blinding treatment', just as native students are.) Treated students, i.e., students with *ethnicity*-neutral SNs, were mainly students from other Western countries or other developed countries, which had first names similar, but not identical, to Danish students (e.g., Aleksandar, Aleksej or Aleksandr, which are similar to the Danish version of the name, *Aleksander*). Immigrants from non-Western countries only rarely share the beginning of their first name with Danes. Therefore, this strategy cannot be used to examine bias against immigrants from non-Western countries, who make up the large majority of immigrant students in Denmark.

Table 1: Examples of 4-letter combinations with varying percentages of male students

4-letter combination (beginning of student number)	Percentage male students	# in 2016 (class 9 cohort)	Example of first names covered by student number	Type of student number
marc	100%	616	Marcus	Male-specific
jaco	100%	507	Jacob	
anto	96%	244	Anton	
nico	66%	753	Nicole, Nicolaj	Gender-neutral
math	64%	1,653	Mathias, Mathilde	
joha	50%	479	Johannes, Johanne	
mari	8%	1,123	Marie, Maria	Female-specific
caro	0%	585	Caroline	
clar	0%	270	Clara	

Figure 2: Percentage of student numbers with different empirical male shares (by gender)

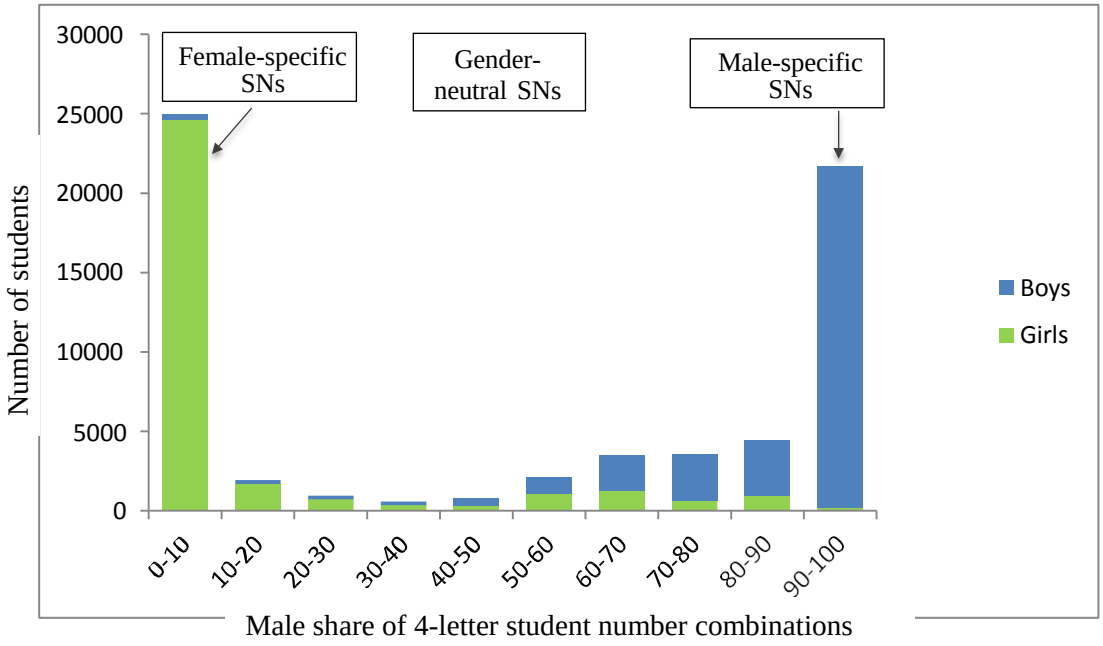


Figure 2 shows the distribution of the percentage of boys for all four-letter combinations (approximately 4,000) that were used by the 2016 year 9 cohort. Figure 2 shows that approximately 65% of boys have a male-specific SN (with an empirical male share of at least 90%). Thus, two out of three boys have an SN that does not conceal their gender. Therefore, with respect to the objective of the grading reform, replacing students' name with the SN on the exam papers does not make a

significant difference for the majority of boys. Figure 2 also shows the corresponding numbers for girls. Roughly 75% of girls have SNs with an empirical male share of less than 10%, corresponding to a female share of more than 90%. This means that 3 out of 4 girls have SNs that do not conceal their gender. Note that only a small share of boys and girls have their gender effectively blinded by the reform.

Students with gender-neutral SNs are those receiving 'treatment' in this natural experiment strategy (because their gender is blinded by the reform). Students with gender-specific SNs are untreated, since their gender is perceivable to the exam grader, just as if exams were marked with students' names. To mimic the perfect experiment, I choose a quite narrow definition of specific and neutral SNs in the main specification of the empirical analysis. The perfect experiment is to compare completely blind and completely non-blind grading, which in my identification strategy corresponds to comparing students with SNs that are 100% male or female (non-blind scores) with boys and girls who have clearly neutral SNs, i.e., SNs that are equally likely to belong to a boy or girl (blind scores). In practice, slightly broader categories must be chosen due to sample size concerns. Therefore, in the main estimation sample, I include students with SNs with a male share of at least 98% in the case of boys and less than 2% in the case of girls (corresponding to a female share of more than 98%) and students with SNs with a male share very close to 50% (47.5-52.5%). I conduct robustness checks using broader cut-offs, but this does not change the main conclusions (see section 5, Table 5).

4. Empirical model & data

The main empirical strategy uses the quasi-random assignment of students to blind grading induced by a flaw in the blinding procedure in the 2016 grading reform. In terms of the difference-in-differences literature, one may think of the empirical setting as a comparison between a treatment (blind scores) and control group (non-blind scores). Students with a gender-specific SN (i.e., their exam papers bear clear information about the student's gender) are the control group. Students with a neutral SN that does not reveal the gender of the student are treated. I use this setting to conduct a difference-in-differences analysis of the effect of blind grading.

A key prerequisite for identification in a difference-in-differences setting is that treatment is (quasi-)randomly assigned. The identifying assumption in this study is that treated and untreated students have similar ability. Specifically in my study, this assumption means that boys with gender-specific SNs (e.g., Jacob, Marcus, Anton) have similar ability as boys with gender-neutral SNs (e.g., Johannes, Mathias or Nicolaj) – and similarly for girls. While there is no reason to believe that this assumption should be violated, it is important to recognize that – because I can directly control for ability in the regressions – I only need *conditional* random assignment to hold. The identifying assumption for conditional random assignment is that parents must not have chosen the name of their child with respect to whether the first four characters are gender-neutral or gender-specific (in ways that are systematically correlated with unobserved characteristics that affect academic achievement). It is difficult to find convincing arguments for why this assumption should not hold.

The assumption for identification by random assignment is usually investigated formally by means of a balancing test. Essentially, students' assignment to treatment should not be related to ability. In other words, there should be no substantial differences in the mean values of ability when we compare treated and non-treated individuals, i.e., students with gender-specific and neutral SNs. This can be tested by means of simple t-tests. If the tests confirm that gender-specific and neutral SNs are as good as randomly assigned, one can just compare average exam results in the two groups. If the tests show that random assignment does not hold, differences in ability can be directly controlled for by including the relevant variables as controls in the grading effect estimations.

Table 2 shows the means of unbiased measures of ability in a range of subjects (reading, math, biology, physics, geology and English).[17] For reading, I have test scores from years 4, 6 and 8 from the national tests[18]. For math, test scores are available for grades 3 and 6. English is tested in year 7 and biology, physics and geology in year 8. The means for students with gender-specific vs. gender-neutral SNs are displayed along with the difference in means between the two, separately for boys and girls. Stars indicate the significance levels of t-tests of the differences. The overall impression is that ability is similar for students with gender-specific vs. gender-neutral SNs, because only 3 out of 18 subject areas display significant differences – and these even go in different directions.[19] Yet, to alleviate concerns about remaining bias in the estimates, the regressions in section 5 include a large set of relevant conditioning variables to control for any observable differences across groups.

Table 2: Balancing test for random assignment on unbiased pre-determined test scores across students with specific or neutral student numbers

	Boys				*Girls*			
Pre-determined testscores (unbiased)	Gender-specific SNs (untreated)	Gender-neutral SNs (treated)	Differ-ence	Signifi-cance	Gender-specific SNs (untreated)	Gender-neutral SNs (treated)	Differ-ence	Signifi-cance
Reading, grade 8	54.7	56.0	1.3		59.9	59.2	-0.7	
Reading, grade 6	56.7	54.6	-2.1		62.2	59.7	-2.5	
Reading, grade 4	54.1	51.8	-2.3		58.5	55.1	-3.4	**
Math, grade 6	58.2	57.1	-1.1		58.7	56.5	-2.2	
Math, grade 3	54.2	52.4	-1.8		52.4	48.3	-4.1	**
Biology, grade 8	55.3	55.0	-0.3		56.6	55.9	-0.7	
Physics, grade 8	59.8	59.9	0.1		56.1	56.1	0.0	
Geology, grade 8	53.1	54.6	1.5		50.5	50.3	-0.2	
English, grade 7	57.1	60.8	3.7	**	56.9	58.6	1.7	

Note: Gender-specific SNs have a male share of 0-2% for girls and 98-100% for boys.
*Gender-neutral SNs have a male-share of 47.5-52.5%. ** indicates significance at the1%-level.*

[17] These test scores are automatically scored by computers and are therefore blind scores, i.e., unbiased with respect to gender.
[18] National tests were introduced in 2010 after the cohorts that are included in the regressions attended year 2. Thus, test results for year 2 are not available for this sample.
[19] The size of the significant differences is 15% of a standard deviation for boys in English, and 14% and 17% of a standard deviation for girls in reading (year 4) and math (year 3).

The identification strategy is a difference-in-differences approach. The first difference is between gender-specific (untreated) and gender-neutral (treated) SNs, and the second difference is between boys and girls. Formally, the regression equation for the effect of the grading reform on the gender gap is:

$$Testscore_i = \beta_1 + \beta_2 Male_i + \beta_3 Blind_i + \beta_4 (Male_i \times Blind_i) + \beta_5 X_i + \varepsilon_i \qquad (1)$$

where $Testscore_i$ is the exam grade, $Male_i$ is a gender indicator equal to 1 for boys and 0 for girls, $Blind_i$ is a dummy indicating whether the student number is neutral (=1; treated) or specific (=0; untreated), and $(\text{Male}_i \times Blind_i)$ is the interaction of the two, which allows the treatment effect to differ by gender. Control variables, X_i, include unbiased measures of student ability and parental background (e.g., education, income and labor market status), and ε_i is the error term. The parameter of main interest is β_4, representing the effect of treatment on boys relative to girls. If β_4 is positive, this would signify that boys receive better grades relative to girls when the grading process is blinded.

Data

The dataset used in the empirical analysis is put together from different administrative registers hosted by Statistics Denmark. For the analysis, I use data for the entire cohort of year 9 students who took the school-leaving exams in 2016. The dataset contains information on roughly 65,000 students in public and private schools. Data on pupil background are linked to test scores via a unique personal registration number.

The estimation sample includes students with specific or neutral SNs. I drop students with SNs that only appear infrequently (less than 100 students in the entire 2016 year 9 cohort of 65,000 students), because in these cases, graders may not have a reasonable opportunity to form accurate expectations about students' gender. In the section on robustness, I provide results without this constraint. The final estimation sample contains roughly 27,500 students.[20]

Exam scores from the school-leaving exam in essay writing in Danish and problem solving in math are the main outcomes in the grading bias estimations. I choose grades in essay exams in Danish as my main outcome variable, because there is more room for discretion in grading essays than in grading math. However, as an additional analysis, I also present results for math.

Exam scores are reported on a 7-tiered grading scale that directly translates to the international ECTS scale.[21] For comparison with other studies, I standardized exam scores to a distribution with zero mean and a unit standard deviation, meaning that the effect of blind grading should be interpreted as the share of a standard deviation of the exam score. As controls, I include a number of test scores from the national tests in various subjects. I also add controls for the socio-economic background of the student. Table A2 provides descriptives on all variables used in the estimations in section 5.

[20] Ability as measured by the (unbiased) national tests is slightly higher in the estimation sample than in the full cohort. However, the differences are not substantial; see Table A1.
[21] The scores in the grading scale are 12/A, 10/B, 7/C, 4/D, 02/E, 00/Fx, -03/F.

5. Results (reform-based approach)

This section presents results from the grading bias regressions using the reform-induced identification strategy. Table 3 presents the results. The results from a simple model without controls are shown in the first column. Column 2 adds ability measures at the student level. The model in column 3 adds controls for SES, and the last column adds school fixed effects. Only the estimate of main interest – the interaction effect between being treated and being male, corresponding to β_4 in equation 1 – is shown in the tables, while the full results for model (4) are available in the appendix, (Table A3). This estimate shows whether the treatment effect is different for boys, indicating that one gender profits (more) from blind grading than the other. A priori, we would expect the treatment effect to be larger for boys than for girls ($\beta_4 > 0$). This would indicate that boys gain more than girls[22] from having a neutral SN, i.e., from blind grading. In all models, cluster-robust standard errors are calculated at the school level.

Table 3: Effect of blind grading for boys (compared with girls) in essay exams

	(1)	(2)	(3)	(4)
Interaction: male x blind grading (β_4)	0.096 (0.070)	0.070 (0.060)	0.058 (0.060)	0.047 (0.060)
Ability		x	x	x
SES			x	x
School fixed effects				x
R^2	0.073	0.309	0.320	0.434

Notes: Dependent variables are standardized scores. A constant and the main effects are always included. Standard errors corrected for clustering at the school level are reported in parentheses. N = 27,452 in all specifications.

In the base-line estimation in column 1, the point estimate is 0.096, indicating a blind grading effect of nearly 10% of a standard deviation for boys relative to girls. However, the estimate is imprecisely estimated. When controls are included in models (2) to (4), the coefficient estimate drops to approximately half the size. However, the estimate in the fully specified model is of non-negligible size. The point estimate is 0.047 SD, corresponding to 9% of the gender exam grade gap in essay writing. The estimate is poorly determined, however.

The sign and size of the point estimate – a blind grading effect in favor of boys of approximately 0.05 SD – are virtually equal to the result in Lavy (2008)[23]. However, although equal in size, Lavy's estimate is highly significant, while the point estimate in my study is poorly determined. Thus, while my study also reveals a tendency toward a blind grading effect in favor of boys, the standard errors are too large to produce significant results.

Taken at face value, the point estimates tend to be supportive of the 'bias against boys' hypothesis, but they are imprecisely estimated, and thus, the null hypothesis of no discrimination could not be rejected. I now turn to consider potential caveats in the empirical design, which all tend to bias the

[22] Actually, we would expect boys to gain and girls to lose from blind grading.
[23] Lavy (2008) finds a non-blind disadvantage against boys of -0.053 SD in literature (Tb. 3), corresponding to a blind grading effect in favor of boys of the same size.

estimate towards zero. A major concern with any blind/non-blind setting is that the blind grader also (at least partly) observes the relevant characteristics of the student or, conversely, that the non-blind grader fails to observe or recognize the relevant student characteristics. Both mechanisms would introduce a downward bias in the result. I discuss these concerns below.

First, in the present setting, it is unlikely that non-blind graders do not notice the SN on the exam papers or do not recognize the information on gender embedded in the SN. It is difficult to avoid *noticing* the SN that is written on each single page of students' exam papers. Furthermore, as argued in section 3, through their work as teachers, they have an identically structured number themselves. Thus, they should know that the SN holds the beginning of the first name. Moreover, in a survey, they answered that they *recognize* students' gender. This is evidence that they both notice the SN when they grade the exam papers and that they are aware that the SN contains information on gender.

The second concern is that student gender in some cases can be inferred even with blind grading (i.e., blind scores are not fully blind). A typical example from previous studies is that if exams are handwritten rather than typed, the handwriting itself may reveal student gender, because girls' handwriting often is distinguishable from boys' (Lavy, 2008; Hinnerich, 2011). Yet, at the school leaving exams in Denmark, all students use computers to type their essays. Therefore, the potential threat to the identification strategy of revealing gender through handwritten exams is not present in this study.

I find some evidence, however, that gender may be revealed in other ways. First, there is evidence from the survey among graders that students erroneously wrote their full names on the exam paper instead of their student number in some cases. Second, students have also sometimes used their full name as part of the exam text. For example, when asked to write a newspaper article, some students use their full name in the byline. I cannot check this in the full population register data, but the original exam papers are available for the random sample in section 6 (regrader study). I manually checked the exam papers from the regrader study and found that students revealed their names in approximately 10% of the essay exams. There is no way of knowing, however, whether this number is representative of the overall situation. Third, blind graders can detect a clue about the student's gender from the student's choice of essay topic. Students choose from a number of essay topics, and this choice may have revealed information on their gender, as indicated by Van Ewijk (2011) and Hinnerich et al. (2011).[24] Moreover, some topics and literary genres were more prone to reveal students' gender through the contents of the text than others. For example, one topic was a first-person narrative requiring the student to choose the gender of the first-person storyteller. An examination of the essays from the regrader study (section 6) shows that girls tend to choose female storytellers, while boys choose males. Another topic was on educational choice, which tends to reveal students' gender through gendered choices (e.g., police officer or nurse).

[24] Schools choose between two forms of exams: with access to the internet or without access. Schools have to inform the Ministry of Education about their choice some months before the exam is held. At the beginning of the exam, students choose between different topics for their essays. The exam with internet access came with a choice between four topics, while the exam without internet access came with six choices.

Unfortunately, I cannot directly account for this in the regressions in this section that use full population data, because information on students' choice of topic, etc., is not available in the register data. This information, however, is available for the regrader study in the next section. Thus, in the analysis in section 6, I can directly control for these potential sources of bias. Indeed, the point estimate from a regression using data from the regrader study in a specification with topic fixed effects and including only properly blinded exams is substantially larger (0.079 vs. 0.047 SD; Table 8, col. 3), lending support to these concerns. This suggests that the results in Table 3 are probably lower bound estimates of the true effect of grading bias.

Robustness and extensions

Table 4 provides results from robustness checks where the cut-offs for the definitions of specific and neutral student numbers are different from the main specification. To ease comparison, the main results from Table 3, col. 4, are repeated in the first column. In the main specification, gender-specific SNs are defined as SNs that are at least 98% male or female. Columns 2 and 3 present results with broader definitions. In column 2, SNs that are at least 90% male or female are defined as gender-specific, while column 3 uses 80% as the cut-off. These variations produce very similar point estimates.

Columns 4 and 5 report results from models where the definition of the gender-neutral category is extended to 45-55% male/female in model 4 and to 40-60% in model 5 (in the main specification, the cut-offs are 47.5-52.5%). Using these broader categories increases the share of treated students in the estimation sample from 3% in the main specification to nearly 10% using the broadest definition. With broader definitions for the gender-neutral category, the point estimates decrease. This does not change the main conclusion, however, that the sign of the point estimate is positive but imprecisely estimated.

Table 4: Sensitivity checks

	(1) Main result, repeated	(2) Broader untreated: 0-10% & 90-100%	(3) Broader untreated: 0-20% & 80-100%	(4) Broader treated: 45-55%	(5) Broader treated: 40-60%
Interaction: male x blind grading	0.047	0.046	0.046	0.012	0.014
	(0.060)	(0.059)	(0.058)	(0.036)	(0.034)
N	27,452	32,378	36,517	29,001	29,338
R^2	0.434	0.428	0.426	0.432	0.432

Notes: Dependent variables are standardized scores. A constant and the main effects are always included. Standard errors corrected for clustering at the school level are reported in parentheses.

In Table 5, I present additional results. To check whether grading bias varies over the ability distribution, I estimate separate regressions for low, average and high achievers measured by terciles of scores from the national test in reading in year 8. The national tests are blind scores (automatically scored by the computer) and can therefore be used as a proxy for ability. Columns 2 to 4 in Table 5

present the results. The point estimates indicate that grading bias tends to be larger at the lower end of the ability distribution, suggesting that low-performing boys in particular are hurt by non-blind grading. However, the null hypothesis that the estimates are equal cannot be rejected.

Table 5: Results by ability and for opposite-gender student identification numbers

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Main result, repeated	Low achievers	Average achievers	High achievers	Including students with opposite-gender SNs (80%)	Including students with opposite-gender SNs (85%)	Including students with opposite-gender SNs (90%)
Interaction: male x blind grading (neutral SN)	0.047	0.084	0.061	0.040	0.040	0.045	0.045
	(0.060)	(0.116)	(0.143)	(0.125)	-0.059	(0.060)	(0.060)
Interaction: male x opposite sex SN					0.071	0.115*	0.150*
					(0.045)	(0.051)	(0.074)
N	27,452	6,426	7,436	7,737	33,723	33,395	32,862
R^2	0.434	0.449	0.389	0.381	0.428	0.427	0.428

Notes: Dependent variables are standardized scores. A constant and the full set of controls are always included. Standard errors corrected for clustering at the school level are reported in parentheses.

Last, as explained above, in this experimental set-up, students with gender-specific SNs are untreated and students with neutral SNs are treated. We actually observe yet another treatment, however: boys who are treated with student identification numbers that appear to be female-specific and vice versa for girls (*double dose treatment*). For example, boys named Marius or Laurits are treated with SNs that are predominantly *female* (*mari*: mainly Maria, Marie, Marianne; *laur*: Laura[25]). In such cases, external graders are led to believe that they are scoring exam papers written by girls. Similarly, girls named Patricia or Augusta have SNs that are predominantly male, leading external graders to believe they are scoring exam papers written by boys named Patrick or August.[26] Table 6 lists these three different types of SNs: (i) gender-specific SNs that reveal students' gender and thus provide the same information to external graders as before the reform when students marked the exam papers with their names (= no treatment), (ii) gender-neutral SNs that are uninformative about gender (= treatment corresponding to blind grading) and (iii) opposite-gender specific SNs that suggest that students are of the other gender (= double dose treatment: boys appear to be girls and girls appear to be boys).

[25] 92% and 89% of students whose names begin with *mari*/*laur* are girls.
[26] 91% and 86% of students whose names begin with *patr*/*augu* are boys.

Table 6: Opposite-gender SNs and their relation to no treatment and blind treatment

Types of SNs	*Treatment status*	*Variables in regression*
(1) Gender-specific SN: *Same information about students' gender as before reform, i.e., with students' names on exam papers.*	Untreated	(Omitted category)
(2) Gender-neutral SN: *Reveals no information about student gender, equivalent to blind grading*	Main treatment	***Blind*** Treatment with neutral SN *(interacted with male indicator)*
(3) Opposite-gender SN: *Information suggesting that student has the opposite gender.*	Secondary treatment	***Opposite*** Treatment with opposite-sex SN *(interacted with male indicator)*

If gender grading bias is present, one would expect the effect of being treated with an opposite-gender SN to be even larger than when the student has a gender-neutral SN. To investigate this, I add the sample of students with opposite-gender-specific SNs to the analysis. I define opposite-gender SNs as student numbers that are mostly used by the opposite gender. I present results using three different cut-offs: 80, 85 and 90%, for example, boys who have an SN that is at least 80% female-specific and vice versa for girls. In this specification, two treatment indicators are interacted with the gender dummy: treatment with a gender-neutral SN and treatment with an opposite-gender SN. Formally, I write

$$Testscore_i = \beta_1 + \beta_2 Male_i + \beta_3 Blind_i + \beta_4 (Male_i \times Blind_i) + \beta_5 Opposite_i + \beta_6 (Male_i \times Opposite_i) + \beta_7 X_i + \varepsilon_i \qquad (2)$$

where $Opposite_i$ is an indicator for having an opposite-gender specific SN and ($\text{Male}_i \times Opposite_i$) is the interaction with the male indicator (the other terms are the same as in equation 1). As before, if β_4 is positive, this would signify that boys receive better grades relative to girls when the grading process is blinded, i.e., when they have gender-neutral SNs. If β_6 is positive, this would signify that boys receive better grades relative to girls when students have opposite-gender SNs – i.e., boys look like girls and girls look like boys to the examiner – than when their true gender is revealed (no treatment is the reference category).

The results are shown in Table 5, columns 5-7. As one would expect, the point estimate of the simple treatment (having a gender-neutral SN) is virtually identical to the result in the main specification. The point estimate for having an opposite-gender SN, however, is much larger than for neutral SNs. This suggests that having an SN that leads graders to believe the student is of the opposite gender is more advantageous for boys than for girls. When the definition of opposite-gender-specific SNs is narrow, i.e., those with an SN that is shared by at least 85% of the opposite gender (col. 6 and 7) are defined as treated, the point estimate for boys' advantage under blind grading is significant and exceeds 10% of a SD. This result provides further evidence of the existence of gender bias in the grading of essays.

This setup mimics the experimental studies of Van Ewijk (2011), Hanna & Linden (2012) and

Sprietsma (2013). In these studies, essays are not blinded but are randomly assigned immigrant and native first names (high and low caste in Hanna & Linden). This makes some graders believe a given essay was written by a native student, while others believe it was written by an immigrant student. Thus, the estimates in these studies are not blind grading effects but opposite-group effects. However, while I examine differences by gender, the existing experimental studies all focus on ethnicity.

Exam scores in essay writing are chosen as the main outcomes in this paper, because we expect bias to be larger in essay writing both due to the higher degree of discretion in the grading of essays compared with math and due to the larger gender gap in essay scores (providing scope for statistical discrimination). In contrast, math achievement does not vary much by gender, and therefore, statistical discrimination due to differences in achievement is unlikely to be an important mechanism.[27] Yet, even if achievement in math is similar for boys and girls, long-lived gender stereotypes of mathematics being a male domain might still produce a bias against girls in math.

Table 7: Results for Math

Math	(1)	(2)	(3)	(4)
Interaction: male x	0.026	0.008	-0.006	-0.034
blind grading	(0.069)	(0.049)	(0.049)	(0.048)
Ability		x	x	x
SES			x	x
School fixed effects				x
R^2	0.001	0.473	0.494	0.578

Notes: Dependent variables are standardised scores. A constant and the main effects are always included. Standard errors corrected for clustering at the school level are reported in parentheses. N = 27,388 in all specifications.

The results for math are shown in Table 7. As in Table 3 for essay writing, the first results are from a simple model without controls, and controls are then added progressively in columns 2 to 4. In the simple model, the positive point estimate suggests a small positive effect of blind grading for boys relative to girls (0.026 SD). The point estimate drops, however, when controls are added and the point estimate in the fully specified model in column 4 is negative. Contrary to essay writing, boys are disadvantaged by blind grading in math (while girls profit), but the effect is small (0.034 SD) and imprecisely estimated.[28] Robustness checks corroborate these results (Table A4). Overall, it is interesting that the sign of the coefficient is different for math compared with essay writing. The tendency that girls seem to profit from blind grading in math is compatible with the notion of

[27] The average exam grade in essay writing for boys in the 2016 year 9 cohort is 5.70 compared to 7.28 for girls, corresponding to 0.55 SD of the raw exam scores. By comparison, in math (problem solving), the average for boys is 6.76 compared to 6.70 for girls. This difference amounts to only 0.02 SD of the raw math scores.

[28] While it is much less likely that students' gender is revealed by the contents of a math exam (compared with an essay exam), a disadvantage for identification of the math model is the fact that math exams are often at least partly written by hand rather than typed (essay exams are always typed).

mathematics as a male domain.

To conclude, the results in this section – based on reform-induced variation in assignment to blind grading – suggest that there might be small (imprecisely estimated) effects of blind grading in favor of boys in essay writing and in favor of girls in math. As discussed in section 1, these results are based on between-student variation. In the next section, I provide additional evidence from within-student estimates that are retrieved from blind and non-blind grading of a random sample of the same tests.

6. Evidence from a field experiment

A considerable strength of this study is the availability of two independent identification strategies for the same cohort and exam. In section 5, the results of an analysis where identification was provided by a grading reform were presented. In this section, I present results from an additional analysis where identification derives from a field experiment comparing blind and non-blind scores for the exact same exam papers. My setup is similar to that used by Hinnerich et al. (2011). They carry out a study that compares blind and non-blind scores for the same exam papers by analyzing a random sample of essay exams that were graded twice: non-blind (as part of the national exam) and blind (as part of the scientific study).

Likewise, in my analysis, a sample of 250 essay exams has been graded blindly by graders with no information about students' identities. These 250 exams from the 2016 school-leaving examinations in essay writing were drawn among students with gender-specific SNs (see section 3). Thus, the gender of these students was clearly visible in the original grading, and the original exam grading was therefore non-blind. These non-blind test scores from the original school-leaving examinations are then compared to the blind test scores obtained by blind regrading of these exam papers. Before regrading, the student identification numbers have been manually deleted from the exam papers, such that gender could no longer be inferred from information on the exam paper, and the regrading was thus blind.

Since there is both a blind score and a non-blind score for each exam, the effect of blind grading on test scores is identified by using a difference-in-differences strategy comparing the difference in the blind and non-blind scores for boys and girls. The interaction formulation of the difference-in-differences model may be written as:[29]

$$\text{Testscore}_{ij} = \delta_1 + \delta_2 \text{Male}_i + \delta_3 \text{Blind}_{ij} + \delta_4 \left(\text{Male}_i \times \text{Blind}_{ij}\right) + \mu_i + \varepsilon_{ij} \qquad (3)$$

j denotes the evaluation procedure: blind (j=B) or non-blind (j=NB). δ_2 and δ_3 identify the effects of gender and of the evaluation procedure on the test scores. The coefficient of interest is δ_4, i.e., the additional effect of blind grading. If δ_4 is positive and significant, this indicates that boys have an advantage when graded blindly relative to girls.

[29] Note that while this equation is similar to equation 1, in equation 2, each student contributes two observations (one blind score and one non-blind score), while equation 1 has only one observation per student (either blind or non-blind).

The difference-in-differences nature of eq. (3) implies that differences between students, e.g., ability, cancel out in this model with regard to the estimated coefficient of interest, δ_4, as long as they have the same effect on the blind and non-blind scores. There is no immediate reason to believe that there are variables that affect the two scores differently. Graders were only informed that the regrader study is part of the follow-up research related to the grading reform but not on the specific purpose of the study. Moreover, the regrading was performed by graders who were also grading the original examinations. They received the same material as for the original exam grading, and they received compensation equivalent to what they would receive for grading regular exams. However, a drawback in this type of study is that while the non-blind scores are high-stakes, the blind scores are low-stakes, since they are only used for scientific purposes.[30] In the robustness section, I examine this assumption further.

An algebraically identical estimate of δ_4 can also be retrieved from the following difference formulation of the difference-in-differences model:

$$\text{Testscore}_{iB} - \text{Testscore}_{iNB} = \alpha + \gamma \text{Male}_i + \varepsilon_i \qquad (4)$$

I use this difference formulation as the baseline model. To test the robustness of γ, I also include an immigrant indicator and regrader- and topic-fixed effects. The regrader data are not linked to the administrative registers, and therefore, I do not have further SES controls available. However, since gender and immigrant background are the only characteristics that are discernible from the student numbers (non-blind scores), they are the only student characteristics that can influence the difference in blind and non-blind exam scores.

Sample and descriptives

Initially, a sample of 360 exams was drawn: 180 exams from boys with male-specific SNs and 180 exams from girls with female-specific SNs. However, the final sample decreases to 251 exams with valid information on the key variables. While the original sample had equal shares of boys and girls, the final sample has a slightly smaller percentage of boys (47%). The most-frequent reasons for the reduction in the sample size were that regraders did not score the exams[31] (in 26% of the missing cases), the school did not send the original exam papers for the students requested (25%), the student was exempted from sitting the exam (23%), and inferior administrative routines[32] (17%). The reasons for attrition were generally equally distributed across gender. Being exempted from the exam, however, was considerably more frequent among boys in this (rather small) sample: 12% of boys but only 2% of girls were exempted from sitting the essay exam.

Table 8 shows averages for both the original exam grades and the grades from the regrading. Overall, the exam grades are higher than the grades obtained by regrading (for both genders). Blind grades are 13% lower than non-blind grades. This corroborates the findings of previous studies. In the study most closely related to this, Hinnerich et al. (2011) find exactly the same results, i.e., that

[30] This drawback is common to my study and to Hinnerich et al. (2011).
[31] These graders dropped out, although they originally had agreed to participate in the regrading.
[32] E.g., students were not enrolled in the school that they attended according to the administrative registers.

the blind grades, on average, are 13% lower than the non-blind grades.[33] Hinnerich et al. explain this difference partly by bias due to personal ties between the teacher and the student in non-blind grading and partly by teachers' incentive to inflate their students' grades due to competition for students between high schools. These are not potential mechanisms driving my results, however, because in my study the non-blind grades are given by external graders as well. However, a potential mechanism both in my study and in the Hinnerich study is that differences between high- and low-stakes exam grades induce the external graders give better grades for high-stakes exams (=the original grades at the school leaving examination). First, graders may grade more leniently for high-stakes examinations because they know that these grades actually matter for the future educational career of students, contrary to the (low-stakes) scientific regrading study. They may therefore simply be more compassionate with students for high-stakes examinations. Second, students may complain about a grade they receive for the school leaving examination. Graders may thus be incentivized to give higher grades in order to avoid complaints. These mechanisms are potentially important only in the original grading procedure; they are not relevant for the blind grades obtained by regrading.

In comparing grades by gender (Tb. 8), note that the relevant comparison is not whether boys' grades differ from girls' but whether the ratio of boys' to girls' grades under non-blind grading are different from that ratio under non-blind grading. Boys receive lower grades than girls under both non-blind and blind grading, but the differential is smaller under blind grading. While boys under blind grading have average grades of only 1.43 below that of girls, their mean is 1.61 lower under non-blind grading.

Table 8: Descriptives for the regrader estimation sample (essay exams)

	All		Boys		Girls		Difference
	Mean	#obs	Mean	#obs	Mean	#obs	(boys-girls)
Exam grade (non-blind)	5.94	251	5.09	119	6.71	132	-1.61
Regrade grade (blind)	5.20	251	4.44	119	5.88	132	-1.43

Main results

Table 9 shows the main results for the exam in essay writing. I run regressions with the difference between the blind and non-blind grades as the dependent variable (eq. 4). The parameter of interest in this specification is the estimate of the male dummy indicating whether boys receive better grades (relative to girls) when assessed blindly. All test scores were standardized to a distribution with zero mean and a unit standard deviation, meaning that the effect of blind grading should be interpreted as the share of a standard deviation of the test score. To account for a possible correlation in observations scored by the same grader, standard errors accounting for clustering at the grader level are used.[34]

[33] Moreover, in a study of double-blind versus single-blind peer reviewing, Blank (1991) finds that acceptance rates are lower and referees are more critical when the reviewer is unaware of the author's identity.

[34] Note that since the sample of exams was drawn at random across the entire student cohort, almost all students attend different schools. Therefore, neither cluster correction of standard errors at the school

I present results from four different specifications, beginning with a simple baseline model and then gradually adding controls (Table 9). The first model only includes the male indicator and an indicator for being native or immigrant. The point estimate is positive (3.5% of a SD), indicating that boys receive better grades relative to girls when assessed blindly, which is consistent with the hypothesis of male discrimination in non-blind grading and, thus, with the results from the reform-based approach in section 5. The point estimate, however, is only imprecisely estimated.

As noted before, a major concern with any non-blind/blind set-up is that the blind grader can also either observe or infer the variable that should be non-observable (here: gender). This would bias the estimate of the effect of blinding towards zero. I discuss two potential mechanisms. First, this could happen if regraders have been able to correctly guess the gender of the student based on the choice of topic for the essay (Hinnerich, 2011; van Ewijk, 2011). Because students choose among different topics for their essays, this choice may reveal information about student gender if some topics are more popular with boys or girls.[35] The 2016 exams had ten different topics to choose from, and even though there were minor differences in the popularity, overall, they were similarly popular with both genders. For eight out of ten topics, the fraction of boys was within the 40-60% interval. Yet, to be safe, Table 9, col. 2 presents results including topic fixed effects. As expected, this only marginally affects the estimate and leaves the overall conclusion unchanged, suggesting that the choice of topic was largely uninformative about student gender. Adding regrader fixed effects also does not affect the estimate very much (Table 9, col. 3).

Table 9: Estimation results (essay) for regrader study

	(1)	(2)	(3)	(4)
Coef (male indicator)	0.035	0.029	0.038	0.079
se	(0.100)	(0.092)	(0.102)	(0.106)
Topic fixed effects		x	x	x
Regrader fixed effects			x	x
Only properly blinded exams [a]				x
n	*251*	*249*	*249*	*233*
AdjR2	-0.005	*0.011*	*0.105*	*0.080*

[a] *This includes only exam papers that have been properly blinded before being sent to the regraders.*

Notes: Dependent variables are differences (blind-nonblind) of standardised scores. A constant and an immigrant indicator are always included. Clustered standard errors are reported in parentheses at the regrader level.

Second, blinding could be flawed if there were mistakes in the manual deletion of information on

level nor controlling for school fixed effects is relevant.

[35] Schools choose whether they offer the essay exam with or without access to the internet. Note that this is a choice at the school level. Since schools in Denmark are co-educational, this choice is unlikely to be related to student gender. The set of topics the students get to choose from differs according to whether the exam is taken with or without internet. Overall, there are ten different topics: four topics if the exam is taken with access to the internet and six without access.

student identification in the exam papers that were submitted to blind regrading. Since I have access to the blinded exam papers, I examined these and found that 18 exam papers (or 7%) had not been properly blinded.[36] As expected, the estimate for the effect of male advantage with blind grading increases when I exclude these from the regression (Table 9, col. 4). The effect size is 8% of a standard deviation, corresponding to 15% of the gender exam grade gap in essay writing. The estimate is poorly determined, however.

Compared to the main results from the natural experiment study (Table 3), which could not account for these three factors, the point estimate is larger (7.9% SD vs. 4.7% SD). Thus, the results from the second identification approach corroborate and even strengthen the result from the main approach in section 5.

Even this estimate of nearly 8% of a SD, however, is probably still a lower bound estimate, because the regressions cannot account for the fact that the content of the essays can give the graders a clue about the gender of the student (e.g., through the choice of the first person storyteller). The only results that are directly comparable are from Hinnerich et al. (2011). In contrast to this study, they find a point estimate close to zero.

Robustness and extensions

The difference-in-differences strategy in the main specification takes account of any covariates that have an equal effect on the blind and non-blind test scores. However, omitted covariates that affect the blind and non-blind test scores differently would be a threat to identification. For example, because girls demonstrate a higher ability at essay writing than boys,[37] identification would be threatened if the difference of the blind and non-blind grades varies by ability. This situation would arise if graders mark essays of low-achievers more leniently in the exam situation (when the grade actually matters for the student) than when assessing the essay for the regrader study (when the grade does not matter). If graders dislike giving students a low grade for their school leaving exams but have no such concerns in the regrader study, this may bias the estimate of the male indicator, because ability is correlated with gender. If low-performers are relatively favored by the non-blind (exam) grade,[38] not controlling for student ability would risk biasing the male indicator downwards.

[36] I.e., not all relevant information on the students' identities was deleted, such that student gender could be inferred by the regrader.

[37] On average, both blind and non-blind scores are markedly higher for girls (see Table 8).

[38] Note that this would violate the parallel trends assumption, since the outcome in the absence of treatment (=blinding) would not have been the same, because other circumstances affecting the evaluation also changed (real life exam situation vs. experimental situation). Any differences induced by the exam vs. experiment situation should not enter the effect estimate.

Table 10: Sensitivity: covariates with different effects on blind and non-blind test scores

	(1) Boys	(2) Girls
Interaction: Effect of blind grading	-0.087 (0.186)	-0.121 (0.194)
n	*389*	*436*
AdjR2	*0.032*	*0.040*

Notes: Dependent variables are standardized scores. A constant and an immigrant indicator are always included. Clustered standard errors are reported in parentheses at the regrader level. Exam papers that are not properly blinded are given untreated status.

One way of testing for this possibility is to include a measure of ability in essay writing as a control in the estimation, but a good proxy is not readily available.[39] Another way of testing is to estimate the effect of blind vs. non-blind grading separately for boys and girls, thereby avoiding the need to directly control for differences in ability. This is not possible, however, with the data from the regrader study alone, because all students are treated, and therefore, only the relative effect on boys vs. girls can be identified. Luckily, useful data from a related regrading study are available. The focus of the other study is *reliability* in grading rather than grading *bias*, and therefore, in that study, the exam papers from the school leaving exams were sent to regraders without any changes (in particular, without blinding them). Thus, in the related study, all students are untreated, and they can therefore act as a control group for the boys and girls in my regrader study. Thus, I pool the data from both studies and conduct separate difference-in-differences regressions for boys and girls.[40] The results are shown in Table 10. Note that the regressions estimate the absolute effect of blind grading on boys' and girls' test scores. Thus, while the blind test scores are 8.7% SD lower than non-blind scores for boys, for girls, they are even lower (12.1% SD). This means that boys are advantaged by blind grading compared to girls (or less disadvantaged). This corroborates the results from Table 9.

[39] A candidate for a proxy for ability in essay writing may be the blind grade itself, since this should be an unbiased estimate of essay writing. However, related analyses suggests that essay grades are measured with a substantial amount of error (the reliability ratio is 0.55, suggesting that nearly half of the variation is due to noise/error). Including an ability proxy with large measurement error as a control might severely bias the male estimate (in an unknown direction), since gender is strongly correlated with ability (e.g., Greene (1993), p. 284). This bias may be even acerbated by the fact that the outcome is measured not in level but in differences, which can increase the variance of the measurement error and reduce the variance of the signal (http://econ.lse.ac.uk/staff/spischke/ec524/Merr_new.pdf). Other candidates for ability proxies do not specifically address ability in *essay writing* (e.g., test scores from the national test in Danish measure reading skills). The correlation between the essay exam score and the national test scores for reading is on the low side (0.5) and is therefore unhelpful as a proxy.

[40] Contrary to the specification in Table 8, this estimates the interaction formulation of the DD model (eq. 2).

Table 11: Effect of blind grading for math exams

	(1)	(2)	(3)
Coef (male)	-0.002	-0.015	-0.010
se	(0.070)	(0.072)	(0.075)
Regrader fixed effects		x	x
Only properly blinded exams [a]			x
n	*264*	*264*	*250*
AdjR2	*-0.005*	*0.020*	*0.013*

[a] *This includes only exam papers that have been properly blinded before being sent to the regraders.*
Notes: Dependent variables are differences (blind-nonblind) of standardised scores. A constant and an immigrant indicator are always included. Clustered standard errors are reported in parentheses at the regrader level.

Finally, the regrading was not only carried out for essay exams but also for math exams. Table 11 provides estimates for bias in the assessment of math exams. All estimates are insignificant, though, providing no evidence of the presence of blind grading for math. While the sign of the point estimate switches for math – compared with essay – just as it did with the reform-based identification approach (Tables 3 & 7), the estimates are insignificant throughout, and the evidence is therefore inconclusive.

7. Conclusions

This paper has investigated the effects of blind versus non-blind grading using both a unique natural experiment and a field experiment. The primary conclusion is that while the data are consistent with the notion that boys perform better under a blind grading procedure in essay writing, the estimated effects show no statistical significance. The effect, however, is more pronounced among low performers. Moreover, evaluators give higher grades to boys' essays when they are led to believe that these essays were written by girls, thus strengthening the conclusion that boys are disadvantaged by non-blind grading. Conversely in math, I find a little (admittedly insignificant) advantage for girls under blind grading. The results for essay writing are in accordance with statistical discrimination, and the math results are consistent with gender-stereotyped beliefs of math being a male domain. However, while this paper provides little evidence – due to poorly determined estimates – that moving to a fully blind grading procedure will substantially decrease the gender gap in essay scores, the size of the point estimates calls for further research on this topic. In any case, the planned implementation of a digital exam management solution in Denmark in 2019 will ensure a fully blind grading procedure, removing concerns about biased grading for the school-leaving exams in the near future.

Furthermore, while a finding of the regrading approach in this paper is that grades are lower under blind grading, this result is almost certainly not generalizable to a situation in which the actual school leaving examinations are fully blind. Results from the reform-based approach yield a comparison of

blind and non-blind grading *within* the school leaving examinations. These results show that within the same examination situation, differences between blind and non-blind grades are very small, indicating that the differences found in the regrader study are most likely due to factors other than blind versus non-blind grading and are probably linked to the low-stakes nature of the regrading procedure.[41] This conclusion is corroborated by the fact that average exam grades did not drop when the (partly) blind grading procedure was implemented in 2016.

Finally, I discussed concerns inherent in studies on blind grading that a student's gender may be inferred even when the exam papers do not hold direct information on the student's gender. I argued that gender may be inferable for a non-negligible portion of the exams because the *contents* of the essays hold the potential to reveal students' gender. Thus, as an extension of the discussion on blind vs. non-blind grading, it might be useful to reflect on whether exam questions could be posed in ways that are less prone to unintentionally reveal students' gender or ethnicity to the grader. For example, instead of leaving it to the student to choose a protagonists' or first person storyteller's gender or ethnicity, determining the gender/ethnicity of the storyteller could be part of the exam question (e.g., "Tell the story from Peter's point of view.").

[41] As mentioned in section 6, graders may grade more leniently for high-stakes exams both out of sympathy for the students and to avoid complaints.

References

Ayres, I., & Siegelman, P. (1995): Race and gender discrimination in bargaining for a new car. *American Economic Review*, 85(3): 304–321.

Bertrand, M., & Mullainathan, S. (2004): Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review*, 94(4): 991–1013.

Blank, R. M. (1991): The Effects of Double-Blind versus Single-Blind Reviewing: Experimental Evidence from The American Economic Review. *The American Economic Review*, 81(5): 1041-1067.

Burgess, S. M., & Greaves, E. (2013): Test scores, subjective assessment and stereotyping of ethnic minorities. *Journal of Labor Economics*, 31(3): 535– 576.

Carlsson, M. & S. Eriksson (2017): The effect of age and gender on labor demand – evidence from a field experiment. IFAU Working Paper 2017:8.

Carlsson, F., Løfgren, Å. & T. Sterner (2012): Discrimination in Scientific Review: A Natural Field Experiment on Blind versus Non-Blind Reviews. *The Scandinavian Journal of Economics*, 114(2): 500-519.

Cornwell, C., Mustard, D., & Van Parys, J. (2013): Non-cognitive skills and gender disparities in test scores and teacher assessments: evidence from primary school. *Journal of Human Resources*, 48(1), 236–264.

Falch, T., & Naper, L. R. (2013): Educational evaluation schemes and gender gaps in student achievement. Econo*mics of Education Review*, 36, 12–25.

Goldin, C. & C. Rouse (2000): Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians. *American Economic Review*, 90(4): 715-741.

Greene, W. H. (1993): *Econometric Analysis (Second edition)*. Prentice-Hall.

Hanna, R. N. & L. Linden (2012): Discrimination in Grading. *American Economic Journal: Economic Policy*, 4(4): 146–168.

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

Hinnerich, B. T., Höglin, E. & M. Johannesson (2015): Discrimination against students with foreign backgrounds: evidence from grading in Swedish public high schools. *Education Economics*, 23:6, 660-676.

Ladd, H. F. (1998): Evidence on discrimination in mortgage lending. *Journal of Economic Perspectives*, 12(2): 41–62.

Lahey, J. N. (2008): Age, women, and hiring an experimental study, Journal of Human Resources, 43: 30-56.

Lavy, V. (2008): Do gender stereotypes reduce girls' or boys' human capital outcomes?

Evidence from a natural experiment. *Journal of Public Economics,* 92: 2083–2105.

Neumark, D., Bank, R., & K. Van Nort (1996): Sex discrimination in restaurant hiring: An audit study. *Quarterly Journal of Economics,* 111: 915-941.

Petit, P. (2007): The effects of age and family constraints on gender hiring discrimination: A field experiment in the French financial sector, *Labour Economics,* 14: 371-391.

Rangvid, B. S. (2015): Systematic differences across evaluation schemes and educational choice *Economics of Education Review,* 48: 41-55.

Riach, P. & J. Rich (2006): An experimental investigation of sexual discrimination in hiring in the English labor market, *The B.E. Journal of Economic Analysis and Policy*, 6(2): Article 1.

Sprietsma, M. (2013): Discrimination in grading: experimental evidence from primary school teachers. *Empirical Economics,* 45:523–538.

Szymanski, S. (2000): A market test for discrimination in the English professional soccer leagues. *Journal of Political Economy*, 108(3): 590–603.

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

Table A1: Ability in the full cohort and in the estimation sample

Pre-determined test scores (unbiased)	Full cohort	Estimation sample	Difference (Estimation sample - full cohort)
Reading, grade 8	56.1	58.1	2.0
Reading, grade 6	57.7	60.2	2.5
Reading, grade 4	54.7	56.9	2.2
Math, grade 6	56.8	59.4	2.6
Math, grade 3	51.9	53.7	1.8
Biology, grade 8	54.6	56.5	1.9
Physics, grade 8	56.7	58.6	1.9
Geology, grade 8	50.4	52.5	2.1
English, grade 7	56.5	57.7	1.2

Table A2: Descriptives of the estimation sample (reform-based approach)

Variable	Obs	Mean	SD	Min	Max
Outcomes (exam grades, year 9)					
Essay writing[a]	27,452	0.069	0.995	-3.167	1.843
Math (problem solving)	27,388	0.091	0.978	-2.952	1.598
Foreign languages	5,572	0.051	0.987	-3.061	1.441
Ability controls (test scores, national tests)					
Reading scores, year 8	21,599	0.088	0.958	-2.363	1.884
Reading scores, year 6	22,937	0.099	0.954	-2.257	1.683
Reading scores, year 4	23,389	0.086	0.967	-2.156	1.818
Math scores, year 6	22,926	0.101	0.966	-2.185	1.691
Math scores, year 3	21,077	0.074	0.979	-2.035	1.924
Biology scores, year 8	21,203	0.088	0.965	-2.435	2.066
Physics scores, year 8	21,303	0.081	0.969	-2.400	1.868
Geography scores, year 8	21,198	0.091	0.969	-2.088	2.098
English scores, year 7	21,815	0.047	0.975	-2.157	1.679
Socio-economic controls					
Male	27452	0.497	0.500	0	1
Lives with both parents	27452	0.669	0.471	0	1
Immigrant	27446	0.022	0.146	0	1
Mother's highest education (reference: lower secondary school)					
Vocational education and training	26974	0.388	0.487	0	1
High-school diploma	26974	0.057	0.232	0	1
Short tertiary education	26974	0.050	0.219	0	1
Bachelor	26974	0.288	0.453	0	1
University	26974	0.102	0.302	0	1
Father's highest education (reference: lower secondary school)					
Vocational education and training	26399	0.446	0.497	0	1
High-school diploma	26399	0.048	0.215	0	1
Short tertiary education	26399	0.073	0.260	0	1
Bachelor	26399	0.157	0.363	0	1
University	26399	0.123	0.329	0	1
Disposable income, mother (mio. DKK)	27132	0.160	0.082	0	>1
Disposable income, father (mio. DKK)	26619	0.199	0.288	0	>1
Mother's labour market status (reference: low-wage job)					
Self-employed	27132	0.043	0.202	0	1
High-wage job	27132	0.186	0.389	0	1
Medium-wage job	27132	0.280	0.449	0	1
Other wage levels	27132	0.095	0.294	0	1
Permanent income transfers	27132	0.084	0.277	0	1
Other employment categories	27132	0.044	0.205	0	1
Father's labour market status (reference: low-wage job)					
Self-employed	26619	0.086	0.281	0	1
High-wage job	26619	0.256	0.436	0	1
Medium-wage job	26619	0.140	0.347	0	1
Other wage levels	26619	0.153	0.360	0	1
Permanent income transfers	26619	0.061	0.239	0	1
Other employment categories	26619	0.026	0.158	0	1

[a] *Average year 9 exam grades are somewhat higher than in the full sample (full sample mean is zero).*

Table A3: Full results for main specification (reform-based approach)

Variable	Coef	se
Interaction: male x blind grading (variable of interest)	0.047	(0.060)
Male	-0.498***	(0.011)
Blind grading (treatment)	-0.021	(0.044)
Ability controls		
Reading scores, year 8	0.152***	(0.010)
Reading scores, year 6	0.104***	(0.010)
Reading scores, year 4	0.127***	(0.009)
Math scores, year 6	0.061***	(0.008)
Math scores, year 3	0.029***	(0.007)
Biology scores, year 8	0.011	(0.009)
Physics scores, year 8	0.001	(0.009)
Geography scores, year 8	0.077***	(0.009)
English scores, year 7	0.071***	(0.009)
SES controls		
Lives with both parents	0.026*	(0.011)
Immigrant	-0.012	(0.041)
Mother's highest education (reference: lower secondary school)		
Vocational education and training	0.075***	(0.018)
High-school diploma	0.088***	(0.026)
Short tertiary education	0.084**	(0.028)
Bachelor	0.096***	(0.021)
University	0.123***	(0.028)
Father's highest education (reference: lower secondary school)		
Vocational education and training	0.063***	(0.015)
High-school diploma	0.141***	(0.029)
Short tertiary education	0.078**	(0.025)
Bachelor	0.123***	(0.021)
University	0.135***	(0.024)
Disposable income, mother (mio. DKK)	0.127	(0.086)
Disposable income, father (mio. DKK)	0.059*	(0.025)
Mother's labour market status (reference: low-wage job)		
Self-employed	0.048(*)	(0.027)
High-wage job	0.045*	(0.019)
Medium-wage job	0.024	(0.016)
Other wage levels	-0.013	(0.019)
Permanent income transfers	-0.013	(0.022)
Other employment categories	-0.03	(0.028)
Father's labour market status (reference: low-wage job)		
Self-employed	-0.021	(0.02)
High-wage job	0.045**	(0.017)
Medium-wage job	0.019	(0.017)
Other wage levels	-0.005	(0.017)
Permanent income transfers	-0.002	(0.024)
Other employment categories	-0.027	(0.033)
Adj. R-sq.	0.434	
N	27,446	

[a] Interaction of gender and blind-grading indicators.

Table A4: Math results. Robustness checks and further results (reform-based approach)

	(1)	(2)	(3)	(4)	(4)	(7)	(8)	(9)	(10)	(11)	(12)
	Main result, repeated	Broader untreated: 0-10% & 90-100%	Broader untreated: 0-20% & 80-100%	Broader treated: 45-55%	Broader treated: 40-60%	Low achievers	Average achievers	High achievers	Including students with opposite-gender SNs (80%)	Including students with opposite-gender SNs (85%)	Including students with opposite-gender SNs (90%)
Interaction: male x blind grading	-0.034	-0.029	-0.029	-0.051	-0.049	0.091	-0.163	-0.095	-0.036	-0.033	-0.031
	(0.048)	(0.047)	(0.047)	(0.030)	(0.028)	(0.093)	(0.107)	(0.091)	(0.048)	(0.048)	(0.048)
Interaction: male x opposite sex SI									-0.024	-0.008	0.056
									(0.040)	(0.044)	(0.069)
N	27,382	32,288	36,433	28,939	29,272	6,779	7,791	8,308	28,730	28,402	27,873
R^2	0.578	0.571	0.570	0.577	0.577	0.512	0.460	0.467	0.576	0.577	0.577