Are men discriminated in Swedish high schools?[*]

Björn Tyrefors Hinnerich[♣], Erik Höglin[♠] and Magnus Johannesson[♦]

Latest version: April, 2010

Abstract

Women typically have higher grades than men in school and recent research suggests that part of this gender difference may be due to discrimination of men. We rigorously test this in a field experiment where a random sample of the same tests in the Swedish language is subject to blind and non-blind grading. The non-blind test score is on average 15% lower for men than for women. Blind grading lowers the average grades with 13%, indicating that personal ties and/or grade inflation are important in non-blind grading. But we find no evidence of discrimination against men. The point estimate of the discrimination effect is close to zero with a 95% confidence interval of ±4.5% of the average non-blind grade.

Key words: discrimination, field experiments, grading, education, gender.
JEL codes: C93, I20, J16.

[♣] Dep. of Economics, Stockholm University, Sweden; e-mail: bjorn.hinnerich@ne.su.se and School of Economics and Management, Aarhus University, Denmark.
[♠] Swedish Fiscal Policy Council, e-mail: erik.hoglin@finanspolitiskaradet.se.
[♦] Dep. of Economics, Stockholm School of Economics, Sweden; e-mail: magnus.johannesson@hhs.se

1. Introduction

Gender differences are present both in school and in the labor market. A puzzling empirical regularity is that while women outperform men in school, they generally have lower wages in the labor market. While a large body of literature has studied gender differences and discrimination in the labor market, much less is known about the causes of gender differences among individuals before entering the labor market.[1]

A recent study by Lavy (2008) indicates that part of the gender difference is due to discrimination of male students. He used a large data set from high school in Israel and compared two different test scores for the same individuals: one school score based on a non-blind grading of a school exam by the student's own teacher and one test score on a similar test graded blindly by an external examiner. He found a statistically significant discrimination of men in all the examined tests. A limitation of the Lavy study is that it does not involve a comparison of blind and non-blind grading of the exact same tests; the author for instance notes that "schools are allowed to deviate from the score on the school exam to reflect the student's performance on previous exams" (p. 2086). Moreover, the mere fact that both students and teachers know that one test is graded locally and the other is graded externally may affect performance on the tests. Lab experiments in economics suggest that subtle changes in context and framing can affect behavior (Levitt and List 2007).

Ideally we would like to compare blind and non-blind grading of the very same tests. In this study we carry out such a test by randomly drawing a sample of compulsory national tests in the Swedish high school. These tests are graded blindly by teachers with no information about the student's identity and the blind test scores are compared with the original non-blind test scores graded by the student's own teachers.

Previous work by Lindahl (2007) suggests that men might be discriminated in the Swedish High School. She compared the non-blind test scores on national tests with the grades on the school leaving certificates, and found that for a given test score on the national test, female students obtained higher grades than male students on the school leaving certificate.[2] However, the national test scores is only one input for the final grades on the school leaving certificates,

---

[1] See for example the OECD PISA reports from 2002, 2003 and 2006 for gender differences in different subjects. Also the historical male advantage in mathematics and science has been reduced. For an overview of gender differences in the labor market, see Altonji and Blank (1999).

[2] Moreover, a number of studies have investigated if the effect is related to the gender of the teacher and the gender/ethnic congruence between student and teacher, e.g. Dee (2005). However, Holmlund and Sund (2008) find no such effects using data on Swedish school leaving certificates.

and women may have outperformed men in other tasks.[3] To credibly attribute inequality to discrimination, it is imperative that the variation being examined is not due to differences in the skills being tested. Our strategy to study the same tests twice using the variation between blind and non-blind grading, fulfills this criterion.

In line with previous work we find a substantial gender gap in the non-blind test scores; the non-blind test scores are on average 15% lower for men than for women. We furthermore find that blind grading substantially lowers the grades; on average the blind grades are 13% lower than the non-blind grades. This is consistent with personal ties between teachers and students affecting the grading and/or grade inflation, i.e. a tendency to increase grades to attract students to the school. However, even though the blind grading substantially lowers the grades, it does not affect the gender difference in grades. The point estimate of the discrimination effect is close to zero with a 95% confidence interval of ±4.5% of the average grade.

In the next section we describe the Swedish high school system and our data collection in more detail. In section 3 we discuss our empirical strategy. The results are presented in section 4 and section 5 concludes the paper.

2. The Design of the Study

2.1 The Swedish high school system

After nine years of compulsory schooling, the vast majority of the Swedish youth enroll in high school education. High school lasts for three years and can be either vocational training or on an academic track. Both the academic track and the vocational programs offer the same set of core subjects, comprising Swedish, English, math, and social studies. Basic courses in the core subjects are compulsory and, upon completion, the student earns basic eligibility for college education.[4] In addition to the core subjects, students on the academic track complete advanced courses in either math/science or humanities/social studies. Students in vocational programs specialize in their field, e.g. cooking, construction and automobile mechanics.

Students' achievements in different subjects are graded on a four-tiered scale: Fail, Pass, Pass with Distinction and Excellent. To calculate a grade point average (GPA), the grades are translated into a cardinal scale with 0 for Fail, 10 for Pass, 15 for Pass with Distinction and 20 for

---

[3] There is no formal relation between the test score on a national test and the final grade in the subject, which makes a comparison between the two types of grades difficult to use for investigating discrimination.

[4] Some college educations, e.g. medical schools and college programs aiming at a degree in engineering, have additional requirements, such as completed high school courses in science and/or advanced math.

Excellent. Grades are absolute and the core subjects have nationally stipulated prerequisites for each grade. The prerequisites are exclusively based on knowledge criteria. Hence, conditional on the level of knowledge, grades must not reflect participation, diligence or ambition. In practice however, teachers enjoy great discretion when setting grades. Grades are not externally evaluated, so teachers could base their grades on anything they observe.

Compulsory national tests are given in the core parts of Swedish, English and math. We focus on the test in Swedish, since we posit that grading a Swedish test allows for more arbitrariness than, for example, math. Every academic year, two national tests in Swedish are constructed by the National Agency of Education in conjunction with the Department of Scandinavian Languages at Uppsala University. The tests have three parts, one oral and two written. We use data from the second, more extensive, written test for the academic year 2005/2006. In this test, students are asked to write an essay based on one out of nine topics within a common theme.[5] Students choose their topic with full discretion.

The written part of the national test is graded on the same scale as the subjects. Teachers are given written guidelines stating the prerequisites for each grade. Once more, they have great discretion in the actual grading. Moreover, the teachers grade their own students. No means are taken by the national authorities to ensure that the guidelines are followed, and no evaluations of the schools are conducted.

In terms of gender differences, the Ministry of Education in 2004 showed that women outperform men in most subjects at all education levels in the Swedish school system (Ubildningsdepartementet 2004). The overall GPA was 10 % lower for men and 7 % more men did not earn pass in the 9th grade. The gender difference was less distinct in mathematics and science than in languages and religion. These differences are also confirmed in the yearly national tests (Skolverket 2006; Lindahl 2007). Historically the gender gap has increased in subjects such as languages and religion, but the historical advantages for men in math and science has turned into a disadvantage.

2.2 Data collection and sampling procedure

The Swedish school system directly provides us with one of the components needed, the non-blind grade. To obtain blind test scores, we drew a random sample of 2880 students from 100

---

[5] We use the fall test of 2005 and the spring test of 2006. The themes were "Leva Livet" (Live Your Life) and "Hur mår du?" (How are you?), respectively.

schools eligible to take the test.[6] Out of the 2880 students in the sample, we received *complete* information, which is the actual test, the test score and the student's identity, for 1712 students.[7] Absenteeism is the main cause for not taking the test, but tests were also missing due to inferior administrative routines at the schools. Out of the 96 participating schools, not all schools had proper filing procedures in line with the guidelines of the National Agency of Education. In the end, 94 schools were able to deliver the required material.

We had all tests rewritten on a word processor and the student identities as well as their teachers' notes were deleted. We did this to ensure that the re-graders would not be able to identify the students' gender or be influenced by the non-blind grade. Naturally, nothing else was changed.

As a final step, we selected about 35-50 tests into groups and hired 42 teachers from a teachers' agency to regrade one group each. The re-grading teachers did not know which student's test they regraded and they had no information regarding the purpose of the study. We required re-graders to have been grading national tests in Swedish before. The teachers were provided the official written guidelines stating the prerequisites for each grade and topic.

Since there is a slight majority of female teachers in Swedish high schools, we required the share of female teachers to be 50-60%. Moreover, we required that 75% of the teachers were certified in order to match the corresponding national share. Out of the 42 regrading teachers, 81% were certified, 52% were women, and 88% were born in Sweden

3. The empirical estimation approach

---

[6] Being eligible means that a student attends a class that is participating in the course Swedish B. To perform the random sample, we obtained a complete list of all 467 Swedish public high schools for 2005/06 and the schools enrollment data from the National Agency of Education. Based on this data, we used a two-step procedure to ensure that each student is equally likely to end up in our sample. In the first step, we weighted all schools by the number of enrolled students in the final year 2005/06. We then chose 100 schools, where the probability of each school being chosen corresponds to its weight in the population. Since Swedish public high schools are subject to a law requiring that documents produced at the schools should be made available to anyone asking for them, we phoned these 100 schools and asked for the classes that took the test either in the fall of 2005 or the spring of 2006. Out of 100 schools we were able to establish contact with 96. After receiving the lists of students in each class, we randomly drew 30 students from each school. Using this procedure, we thus ended up with a sample of 2880 students where all students in the population had the same probability of being sampled.

[7] It is worth noticing that the National Agency of Education requires that all tests and test results should be properly filed and also handed out to any citizen according to the Swedish constitution. As compared to the statistics from the yearly collection of test scores, not tests, that Statistics Sweden does for 200 representative High Schools, we have approximately the same success. For Swedish B, their total response rate for 2006 was about 62 %, as compared to 59 % in this study. Moreover, we did receive about 100 more tests but either the grade was lost, or the wrong test was submitted. According to National Agency of Education, about 10 % of the missing values are due to administrative causes. The rest is due to the fact that eligible students are absent. See: www.skolverket.se/content/1/c4/20/08/kursprovrapport%20vt06.pdf

Let a non-blind (*NB*) test score be determined by student *i*:s ability in a broad sense, the examiner's potential prejudice of gender and an error term. Assume it to be linearly related as

$$Testscore_{iNB} = \alpha_{NB} + ability_i + \beta Male_i + u_{iNB},\qquad(1)$$

where *Male* is an indicator taking the value of 1 if student *i* is male and 0 otherwise. We define gender discrimination as gender differences in the test results conditional on ability. Thus, we could interpret β as a discrimination effect. If negative, then male students are discriminated and if positive, female students are discriminated. The classical problem with this formulation is that we do not observe ability. If ability is correlated with gender, e.g. if female students of school age are more mature or for some reason study harder, then estimating this equation without conditioning on ability would bias β downwards and we could falsely conclude that boys are discriminated, when in fact female students are more able.

Given our set up of the study, this endogeneity problem can be taken care of. Consider an examiner that has no information about gender (*B* for 'blind'). Then, we simply have $\beta = 0$ and

$$Testscore_{iB} = \alpha_B + ability_i + u_{iB}\qquad(2)$$

The difference between (1) and (2) yields the standard difference-in-difference formulation where ability is differenced away and β measures the pure discrimination effect.

Our estimate could still be biased through selection. However, only 6 out of 100 schools did not respond or submitted no information on tests which makes selection very unlikely to be problematic at the school level. For students being absent on the test date to create a problem, we need their potential difference in test scores to be related to gender. It is not a problem for our identification strategy that this group would perform differently from the students taking the test.

Apart from the discrimination effect we also want to estimate the effect of blind grading per se. Hence, we choose to use the interaction formulation of the difference-in-difference model in our analysis:

$$Testscore_{ij} = \alpha + \gamma Male_i + \delta NB_j + \beta Male_i * NB_j + \varepsilon_{ij},\qquad(3)$$

where *j* denotes either blind or non blind grading. The coefficient γ measures the extent to which women are outperforming men. Since NB is an indicator with values 1 if the test was graded non blind and 0 otherwise, δ is our measure of the inflation caused by non-blind grading. β has the same interpretation as before.

4. Results

4.1 Descriptive results

Out of the 2880 students, we are able to determine gender of 2861 by either second last digit in the social security number or first name. However, as noted before, we only have 1712 observations were *both* the blind and the non blind test score is recorded due to absenteeism or substandard administrative routines. Figure 1 depicts the distributions of the blind and non-blind test scores for these observations. In the Figure, we clearly see that female students have higher grades than male students in both the non-blind test score and the blind test score. There is also a clear tendency of an overall down-grading for both genders in the blind grading.

[Figure 1 about here]

Moreover, Figure 2 measures the difference between non-blind and blind test scores. The blind and non-blind test scores are identical for about 50% of the students, whereas the scores differ for the remaining students. The most noteworthy difference is that 5 female students received the highest grades in the non-blind procedure, while they received the lowest grade when graded blindly.

[Figure 2 about here]

Table 1 contains the summary statistics for the 1712 complete observations.[8] In the Table we also report the significance levels for the difference between non-blind and blind test scores and for the difference-in-difference measuring the discrimination effect.[9] In line with previous studies,

---

[8] In this table and in the rest of the paper, we use the cardinal scale used by the national authorities to calculate GPAs, i.e. 0, 10, 15, 20 for Fail, Pass, Pass with Distinction and Excellent.

[9] The p-values are reported both with a parametric test (an independent samples t-test for between subjects comparisons and a paired t-test for the within subjects comparisons) and a non-parametric test (the Mann-Whitney test for between subjects comparisons and the Wilcoxon test for within subjects comparisons).

female students on average get higher grades than male students. The average grade for men is 15% lower than the average grade for women in our data, and this difference is highly significant. Blind grading significantly decreases the average score by 13%, consistent with grade inflation. However, this decrease is of a similar magnitude for both men and women, and the difference between the blind and the non-blind test score is almost identical for men and women. We thus find no evidence of discrimination. The point estimate of the discrimination effect is -0.0013, consistent with discrimination of men. But the effect is close to zero and highly non-significant. To further test the significance of the results we turn to the regression analysis results.

[Table 1 about here]

4.2 Regression results

Table 2 presents the results from the estimation of the regression equation in (3).[10] The variable of interest, the interaction between the male and the non-blind indicator, measures the potential discrimination. The point estimate in the base-line estimation in the first column in the Table is close to zero; the interpretation of the point estimate of -.0013 is that male students get a .0013 lower non-blind grade due to their gender. The sign of this point estimate is consistent with our hypothesis of discrimination of men, but the estimate is very far from significant. Taken at face value, it suggests a discrimination effect of less than .01% of the average non-blind test score.[11] Making use of a standard 95% confidence interval the confidence interval for the discrimination effect is ±4.5% of the average non-blind grade. We conclude that there is no evidence in favor of discrimination of either men or women.

The other estimates in column 1 show that men perform worse and that blind grading is associated with lower grades for both genders. The highly significant point estimate of -1.93 on the *Male* indicator means that the non-blind test score is 15% lower for men than for women, controlling for discrimination. The estimate of 1.53 on the variable *Non-blind test* is also highly significant and means that the blind test score is on average 13% lower than the non-blind test score. As can be expected these results are very similar to the comparisons of mean differences in Table 1.

---

[11] To estimate the discrimination effect in relative terms we use the average non-blind test score for men and women of 11.971 in Table 1.

In the remaining regression equations in Table 1, we add fixed effect for the re-grading teacher, the schools, the rewriter (that rewrote the tests on a word processor) and controls for student's year of birth.[12] The estimates are very robust to the inclusion of these control variables and this also serves as verification that the sampling procedure was appropriate.

[Table 2 about here]

4.3 Extensions and robustness

As explained in section 2 students in the Swedish high school system can chose between two types of high school programs: academic track or vocation training. It is possible that the discrimination effect could differ between these two sub-groups. We therefore as a robustness check estimate our results separately for academic track and vocational training students.[13] These results are presented in Table 3. The point estimate goes in the direction of male discrimination in the academic track and female discrimination in the vocational track, but both effects are far from significant. We also test if the coefficient of the discrimination variable differs significantly between the two groups, but this difference is also far from significant.[14]

[Table 3 about here]

Our dependent variable is not continuous as we only observe four possible grades: 0, 10, 15 and 20. However, in the OLS regressions it is treated as a continuous variable. To test the importance of this assumption we also estimate an interval regression (also known as grouped data regression) with maximum likelihood (Long and Freese 2006). The drawback of implementing this model is that we do not know the exact bounds of the intervals, but in the estimation below we put the bounds at the midpoint between each of the grades.[15] The interval regression results are shown in the two last columns in Table 3. The results are similar to the OLS estimation. The point estimate of the discrimination effect is -0.112, which is somewhat higher in absolute terms compared to the OLS. But the point estimate only corresponds to a discrimination effect of 0.9%

---

[12] Most of the students were born in the year 1987 (84%). Another 14 % were born in either 1986 or 1988. We also have month of birth for a smaller sub-sample. However, nothing substantial changes when adding it as a control.

[13] We have not been able to get full information on vocational and academic tracks for the full sample. Thus, we miss some observations.

[14] The p-value of a z-test of the difference in the discrimination coefficient between equation 1 and 3 in Table 3 is 0.281 and the p-value of a z-test of the difference between equation 2 and 4 in Table 3 is 0.483.

[15] The four grades are thus divided into the following four intervals: <5, 5-12.5, 12.5-17.5, >17.5.

of the average non-blind test score (with an upper 95% confidence interval bound of 5.7%), and it is still highly non-significant. The male indicator variable in the interval regression is -1.99 compared to -1.93 in the OLS regression, and the coefficient of the *Non-blind test* variable is 1.65 in the interval regression compared to 1.53 in the OLS.

5. Concluding remarks

Our study contributes to the increasing literature testing for discrimination in economics (Ayres and Siegelman 1995; Ladd 1998; Szymanski 2000; Bertrand and Mullainathan 2004). We failed to find any evidence of discrimination of men in the Swedish high school. Our point estimate is very close to zero with a relatively narrow confidence interval. So we cannot confirm the results of Lavy (2008) for high school students in Israel. This could either be because there is discrimination in Israel but not in Sweden or because the difference between the school scores and the national scores studied by Lavy is due to other factors than discrimination. Further work is needed to differentiate between these two explanations. It should also be emphasized that we only test for discrimination in one subject/test (Swedish) and it cannot be ruled out that there is discrimination in other subjects in the Swedish high school. We also cannot rule out small effects of discrimination that are within our estimated confidence interval.

Our results suggest that comparing the grades between national tests and the school leaving certificates as done by for instance Lindahl (2007) is not a valid method to detect discrimination. Instead it is necessary to compare blind and non-blind grading on the exact same test as done in the present study. It would be relatively simple for the responsible national authorities to generate such data on a large scale by routinely using blind grading on a sample of the national tests in addition to the standard grading by the student's own teachers. Such data would be a valuable source for continuously testing and monitoring for discrimination in grading. Implementing a system of blind grading on the national tests would also be one way of ensuring against discrimination as well as grade inflation on the tests. But even if the national tests are graded blindly, there is still scope for grade inflation and discrimination in the final subject grades as these are not only based on the national tests.

According to our results blind grading leads to substantially lower grades than non-blind grading, i.e. there is a tendency for teachers to give their own student's a too high grade. It is likely that this tendency can depend on the incentives for teachers and the competition between schools (Jacob and Levitt 2003). In Sweden a system of competition between high schools for students was relatively recently implemented, and concerns have been raised about grade inflation

due to this system (Wikström and Wikström 2005). By giving higher grades, which are important for university admission, high schools can attract better and more students. The personal ties between students and their teachers may also in itself put an upward pressure on grades.

It has been seen in many studies that women outperform men at school (e.g..PISA 2000, Ministry of Education 2004) and our data confirms this. To continue studying the sources of this gender gap is important. As this difference does not appear to be due to discrimination and is unlikely to depend on innate differences in ability (Feingold 1988; Hyde et al. 1990, 2008; Guiso et al. 2008), the most plausible explanation is that women provide more effort in school. To investigate why this is the case and to what extent it varies with different learning environments are crucial for designing policies to decrease the gender gap in school.

References

Altonji, J.G., Blank, R.M., 1999. Race and gender in the labor market. In: Ashenfelter, O., Card, D. (Eds.), Handbook of Labor Economics vol. 3. Amsterdam: Elsevier Science, pp. 3143–3259

Ayres, I., Siegelman, P., 1995. Race and gender discrimination in bargaining for a new car. American Economic Review 85, 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, 991-1013.

Cameron, A.C., Gelbach, J.B. and Miller, D.L. 2006. Robust Inference with Multi-Way Clustering. NBER Technical Working paper 327.

Dee, T.S., 2005. A teacher like me: does race, ethnicity or gender matter? American Economic Review 95, 158-165.

Feingold, A., 1988. Cognitive gender differences are disappearing. American Psychologist 43, 95-103.

Guiso, L., Monte, F., Sapienza, P., Zingales, L., 2008. Culture, gender and math. Science 320, 1164-1165.

Holmlund, H., Sund, K., 2005. Is the gender gap in school performance affected by sex of the Teacher. Labour Economics 15, 37-53.

Hyde, J.S., Fennema, E., Lamon, S.J., 1990. Gender differences in mathematics performance: a meta-analysis. Psychological Bulletin 107, 139-155.

Hyde, J.S., Lindberg, S.M., Linn, M.C., Ellis, A.B., Williams, C.C., 2008. Gender similarities characterize math performance. Science 321, 494-495.

Jacob, B.A., and Levitt, S.D., 2003. Rotten apples: an investigation of the prevalence and predictors of teacher cheating. Quarterly Journal of Economics 118, 843-877.

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

Lavy, V., 2008. Do gender stereotypes reduce girls' human capital outcomes? Evidence from a natural experiment. Journal of Public Economics 92, 2083-2105.

Levitt, S.D., and List, J.A., 2007. What do laboratory experiments measuring social preferences reveal about the real world? Journal of Economic Perspectives 21 (spring), 153-174.

Lindahl, E., 2007. Comparing teachers assessments and national test results: evidence from Sweden. IFAU Working Paper 2007:24.

Long, J.S., and Freese, J., 2006. Regression models for categorical and limited dependent variables using Stata, 2nd ed. College Station, TX: Stata Press.

Skolverket 2002, Likvärdighet i en skola för alla: historisk bakgrund och kritisk granskning, Tallberg Broman, Ingegerd, Rubenstein Reich, Lena och Hägerström, Jeanette. Rapport 02:123, Forskning i fokus. Stockholm: Skolverket.

Skolverket, 2006. Könskillnader i måluppfyllelse och utbildningsval. Technical Report 286.

Szymanski, S., 2000. A market test for discrimination in the English professional soccer leagues. Journal of Political Economy 108, 590-603.

Thompson, S.B. 2009. Simple Formulas for Standard Errors that Cluster by Both Firm and Time. Working paper.

Utbildningsdepartementet, 2004. Könsskillnader i Utbildningsresultat. Utbildningsdepartementet skriftserie 2004, rapport 7.

Wikström, C., Wikström, M., 2005. Grade inflation and school competition. An empirical analysis based on the Swedish upper secondary schools. Economics of Education Review 24, 309-322.

Tables

Table 1. Test scores and differences in test scores.

Sample statistics	N	Mean	Std. Dev
Non-blind test score	1712	11.97138	4.996878
Blind test score	1712	10.44393	5.481632
Difference	1712	1.527453	
p-value of diff. (paired t-test)		<0.0001	
p-value of diff. (Wilcoxon test)		<0.0001	
Non-blind test score, men	858	11.00816	5.072743
Non-blind test score, women	854	12.93911	4.728606
Difference	1712	-1.930952	
p-value of diff. (t-test)		<0.0001	
p-value of diff. (Mann-Whitney test)		<0.0001	
Blind test score, men	858	9.481352	5.591522
Blind test score, women	854	11.41101	5.195446
Difference	1712	-1.929655	
p-value of diff. (t-test)		<0.0001	
p-value of diff. (Mann-Whitney test)		<0.0001	
Non-blind test - Blind test score, men	858	1.526807	5.906692
Non-blind test - Blind test score, women	854	1.528103	5.529504
Difference	1712	-.0012965	
p-value of diff. (t-test)		0.9963	
p-value of diff. (Mann-Whitney test)		0.6216	

Notes: We report data on the test scores, where we have observations on both the blind and the non blind test score.

Table 2. Regression results on the effect of gender discrimination on the non-blind test score.

Variables	1	2	3	4	5
Discrimination	-.0012965	-.0012965	-.0012965	-.0012965	.0234859
	(.2724484)	(.2858932)	(.3075593)	(.3185108)	(.3154342)
Male	-1.929655	-1.967872	-2.015119	-2.025955	-1.981344
	(.3041689)	(.2829438)	(.2907206)	(.3023679)	(.2989651)
Non-blind test	1.528103	1.528103	1.528103	1.528103	1.48503
	(.3758494)	(.3792591)	(.3983403)	(.4004121)	(.3766113)
Regrader fixed effect	No	Yes	Yes	Yes	Yes
School fixed effect	No	No	Yes	Yes	Yes
Re-writer fixed effect	No	No	No	Yes	Yes
Student year of birth	No	No	No	No	Yes
N	3424	3424	3424	3424	3314
R^2	.054	0.1099	0.1969	0.2001	0.2127

Notes: A constant is always included. Two-way clustered standard errors reported in parentheses at the school and re-grader level (Cameron, Gelbach and Miller (2006) and Thompson (2009)).

Table 3. Robustness tests.

Variables	1	2	3	4	5	6
	Academic track		Vocational track		Interval regression	
Discrimination	-.2330128	-.1849376	. 42799	.338633	-.1124055	-.1044033
	(.3825167)	(.4554474)	(.4567069)	(.555356)	(.2885884)	(.2954716)
Male	-1.35641	-1.457964	-2.323837	-2.144782	-1.992324	-2.052662
	(.3788045)	(.3857754)	(.4451255)	(.458403)	(.2862677)	(.496655)
Non-blind test	1.4125	1.401515	1.700581	1.721068	1.645327	1.622824
	(.4342653)	(.4547545)	(.5225005)	(.553811)	(.262307)	(.4781412)
Full sets of controls	No	Yes	No	Yes	No	Yes
N	1580	1540	1388	1344	3424	3314
R^2	0.0413	0.2275	0.0708	0.2469		

Notes: A constant is always included. Two-way clustered standard errors reported in parentheses at the school and re-grader level (Cameron, Gelbach and Miller (2006) and Thompson (2009)) in column 1-4. STATA does not support two-way clustered standard errors for interval regressions and we present standard errors clustered at the school level in column 5-6. Clustering at the re-grader level gives somewhat lower standard errors.
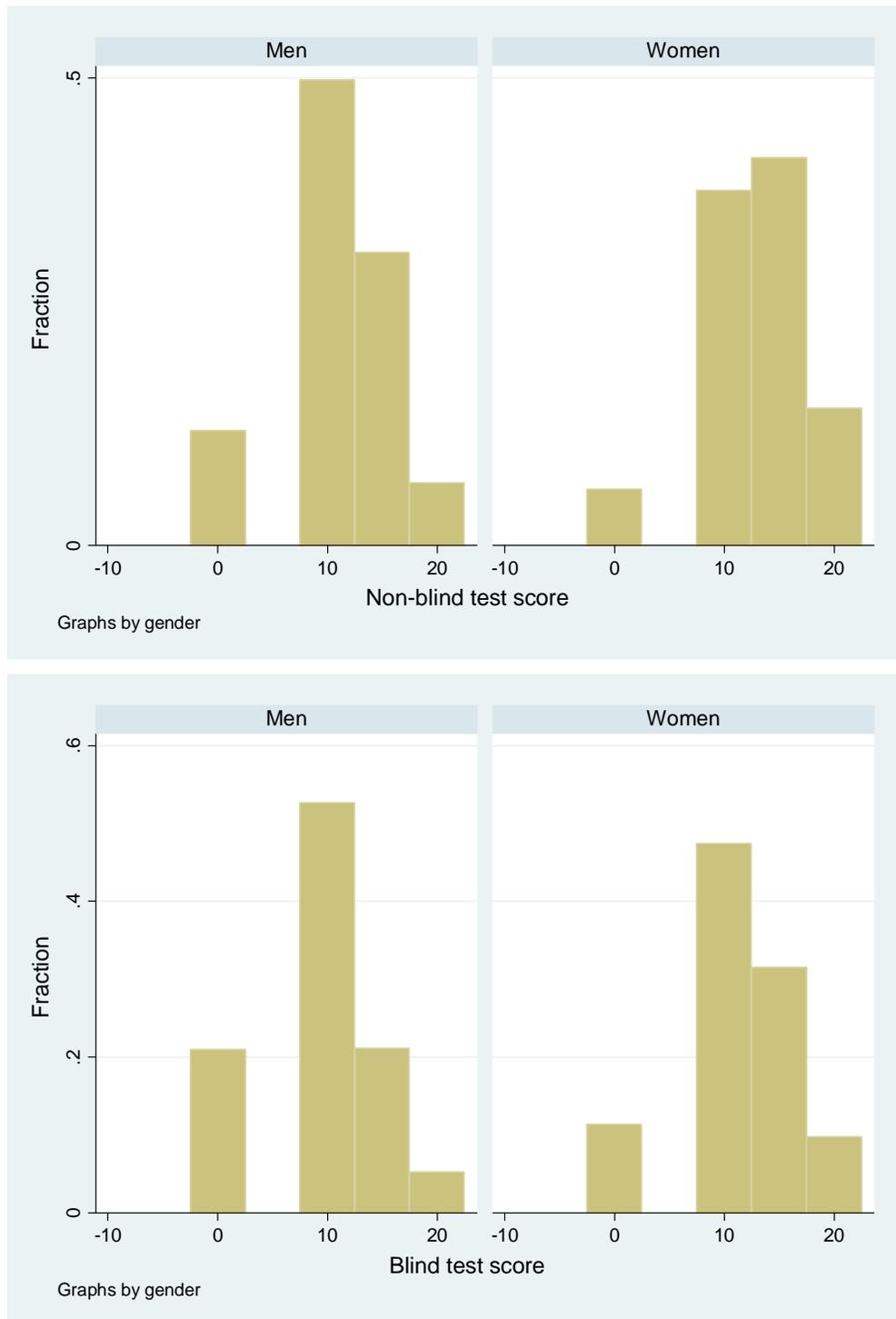
Figures

Figure 1. The distribution of the test scores for the non-blind and the blind grading procedures.
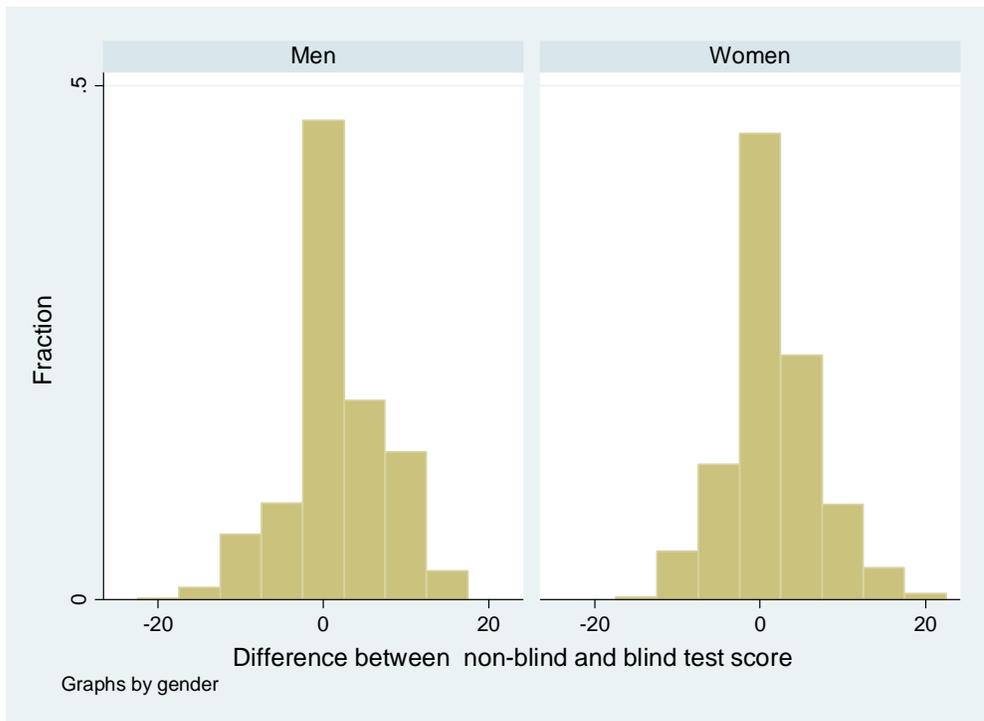
Figure 2. The distribution of the difference in test scores for the non-blind and blind grading procedures