IFAU – INSTITUTE FOR
LABOUR MARKET POLICY
EVALUATION

Are boys discriminated in Swedish high schools?

Björn Tyrefors Hinnerich
Erik Höglin
Magnus Johannesson

WORKING PAPER 2010:14

The Institute for Labour Market Policy Evaluation (IFAU) is a research institute under the Swedish Ministry of Employment, situated in Uppsala. IFAU's objective is to promote, support and carry out scientific evaluations. The assignment includes: the effects of labour market policies, studies of the functioning of the labour market, the labour market effects of educational policies and the labour market effects of social insurance policies. IFAU shall also disseminate its results so that they become accessible to different interested parties in Sweden and abroad.

IFAU also provides funding for research projects within its areas of interest. The deadline for applications is October 1 each year. Since the researchers at IFAU are mainly economists, researchers from other disciplines are encouraged to apply for funding.

IFAU is run by a Director-General. The institute has a scientific council, consisting of a chairman, the Director-General and five other members. Among other things, the scientific council proposes a decision for the allocation of research grants. A reference group including representatives for employer organizations and trade unions, as well as the ministries and authorities concerned is also connected to the institute.

Postal address: P.O. Box 513, 751 20 Uppsala
Visiting address: Kyrkogårdsgatan 6, Uppsala
Phone: +46 18 471 70 70
Fax: +46 18 471 70 71
ifau@ifau.uu.se
www.ifau.se

Papers published in the Working Paper Series should, according to the IFAU policy, have been discussed at seminars held at IFAU and at least one other academic forum, and have been read by one external and one internal referee. They need not, however, have undergone the standard scrutiny for publication in a scientific journal. The purpose of the Working Paper Series is to provide a factual basis for public policy and the public policy discussion.

ISSN 1651-1166

Are boys discriminated in Swedish high schools?[*]

by

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

November 22, 2010

Abstract

Girls typically have higher grades than boys in school and recent research suggests that part of this gender difference may be due to discrimination of boys. 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 boys than for girls. 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 boys. 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.

Keywords: discrimination, field experiments, grading, education, gender.
JEL-codes: C93, I20, J16.

[*] We have benefited from comments and suggestions from Bertil Holmlund, David Strömberg and two anonymous referees. Alexander Höglin, Christoffer Tyrefors and Karolina Wallin provided excellent research assistance. The views expressed in this paper are the authors' and do not represent those of the Swedish Fiscal Policy Council nor its members. Financial support from the Institute for Labour Market Policy Evaluation (IFAU), Jan Wallander and Tom Hedelius Foundation, the Swedish Research Council, and the Swedish Council for Working Life and Social Research is gratefully acknowledged.

[♣] 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.

Table of contents

1 Introduction

Gender differences are present both in school and in the labor market. A puzzling empirical regularity is that while girls outperform boys in school, they generally have lower wages when entering 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 boys 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 & 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 regraded 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.

[1] See for example the OECD PISA reports from 2002, 2003 and 2006 for gender differences in different subjects and the recent papers by Castagnetti & Rosti (2009), Hajj & Panizza (2009), Bedard & Cho (2010), Guo et al. (2010), and Lai (2010). Also the historical male advantage in mathematics and science has been reduced. Campbell et al., 1999 Jay R. Campbell, Catherine M. Hombo and John Mazzeo, Trends in Academic progress: Three decades of student performance, National Center for Education Statistics 2000-469 (19For an overview of gender differences in the labor market, see Altonji & Blank (1999).

Previous work by Lindahl (2007) suggests that boys might be discriminated in the Swedish 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 score is only one input for the final grades on the school leaving certificates, and girls may have outperformed boys 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.

Our study is important to a wider audience for several reasons. Firstly, it is important to test if we can confirm the Lavy (2008) result that boys are discriminated against using an even more rigorous methodology (i.e. using the exact same test for both the blind and the non-blind grading). Secondly, given the importance of gender equality it is fundamental to obtain more well-controlled empirical evidence on the occurrence of gender discrimination in different settings and countries. Thirdly, to compare blind and non-blind grading is important to decide whether it is motivated with policies to grade exams blindly. Currently these policies differ between countries.

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 boys than for girls. 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.

[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 & Sund (2008) find no such effects using data on Swedish school leaving certificates.

[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.

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 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. Since, students should be evaluated according to absolute criteria in their final grades in each subject, the test aims at helping the teachers to measure some of the knowledge

[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.

criteria that should determine the final grade. The final grade will be important when applying to universities after completion of high school. However, there is no formal relation between the national test and the final grade in the subject and there is indeed substantial variation proving the fact that the test is only one of the determinants for the final grade in the subject.[5] Thus, if the knowledge level is observed independently of the national test, the national test score could be completely ignored by the teacher when setting the final grade. 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.[6] 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, but 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.[7]

In terms of gender differences, the Ministry of Education in 2004 showed that girls outperform boys in most subjects at all education levels in the Swedish school system (Ministry of Education 2004). The overall GPA was 10 % lower for boys and 7 % more boys 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 (Swedish National Agency for Education, 2006; Lindahl, 2007). Historically the gender gap has increased in subjects such as languages and religion, while advantages for boys in math and science have turned into a disadvantage.

[5] See, for example, Lindahl (2007).

[6] 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.

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 schools eligible to take the test.[8] 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 1713 students.[9] 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 re-grade one group each.[10] The re-grading teachers did not know which student's test they regraded and they had no information regarding the purpose of the study. The teachers were provided the official written guidelines stating the prerequisites for each grade and topic.

[7] In 2010, the Swedish government launched a first evaluation in order to ensure objectivity of grading.

[8] 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.

[9] 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

Since there were only a few characteristics that could be used to match the re-grader with the Swedish population of teachers, we required re-graders to have been grading national tests in Swedish before. With a slight majority of female teachers in Swedish high schools, we also 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 female, and 88 % were born in Sweden. Moreover, the re-grading teachers had 7.8 years of teaching experience, were born 1969 on average and were located all over Sweden.[11]

3 The empirical estimation approach

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

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

where $Male$ is an indicator taking the value of 1 if student *i* is a boy and 0 otherwise. We define gender discrimination as gender differences in the test results conditional on ability. To put it differently: If grades are not discriminatory, then two students of different gender producing the same quality of the test should get the same grade.[12] If not, one of them is discriminated. Thus, we could interpret β as a discrimination effect. If negative, then boys are discriminated and if positive, girls 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.

[10] The agency is represented all over Sweden and was established 1999.

[11] The oldest was born in 1953 and the youngest in 1983.

[12] We think it is appropriate to use the label "discrimination" here. According to the written guidelines the teacher should only grade the test according to the quality of the test, and nothing else. However, it is possible that a discrimination effect could be due to discrimination with respect to some unobserved characteristic that is correlated with gender. But even if this is the case, it would still result in discrimination. It is very difficult to separate such

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

$$Testscore_{iB} = \alpha_B + \delta ability_i + u_{iB}. \tag{2}$$

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 as:

$$\Delta Testscore_i = \alpha + \beta Male_i + u_i \tag{3}$$

where $\Delta Testscore_i = Testscore_{iNB} - Testscore_{iB}$, $\alpha = (\alpha_{NB} - \alpha_B)$ and $u_i = u_{iNB} - u_{iB}$.

It is worth noticing that an explicit assumption is that δ carries no subscript, i.e. ability is assumed to affect the non-blind and blind test score in the same way. We argue that there is no reason for ability to systematically affect the test score differently in the two equations, given that grading is based on absolute knowledge criteria and that both the teachers and the re-graders were given the very same detailed instructions for grading the test.

Our discrimination 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 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 as our baseline model:

$$Testscore_{ij} = \alpha + \gamma Male_i + \lambda NB_j + \beta\left(NB_j * Male_i\right) + \varepsilon_{ij} \tag{4}$$

where *j* denotes either blind or non blind grading. The coefficient γ measures the extent to which girls are outperforming boys. Note that γ, in contrast to β, could be biased

indirect discrimination from direct discrimination due to preferences. Since other studies use the label discrimination when facing the same methodological problem we stick to that convention here (see Altonji & Blank, 1999).

because of absenteeism. For example, assume boys are poorer than girls (γ<0). If the worst students (more boys) are absent, then we would underestimate γ in absolute terms. We will therefore also add control variables to equation (4) to test the robustness of our estimate of γ. Since *NB* is an indicator with values 1 if the test was graded non blind and 0 otherwise, λ is our measure of the inflation caused by non-blind grading. β has the same interpretation as before.

To test the robustness of the discrimination effect (β) by adding individual invariant covariates such as school fixed effects and year of birth, we will use equation (3) instead of equation (4), as equation (4) saturates all these effects.

4 Results

4.1 Descriptive results

Out of the 2880 students, we are able to determine gender of 2861 by either the second last digit in the social security number or first name. However, due to absenteeism or substandard administrative routines at the schools, we only have 1713 observations were *both* the blind and the non blind test score is recorded. 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.

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.

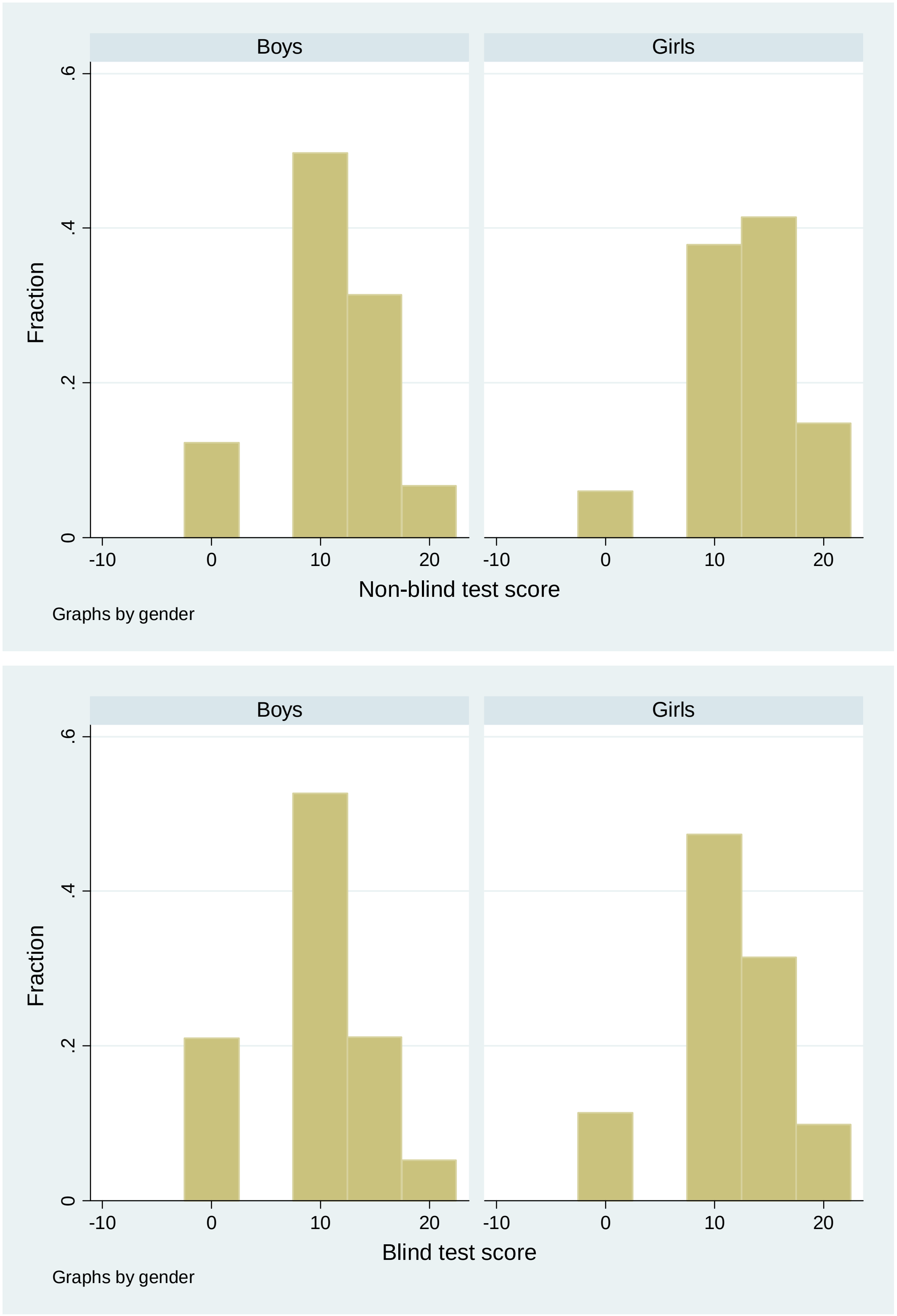

Figure 1 The distribution of test scores for the non-blind and the blind grading procedures

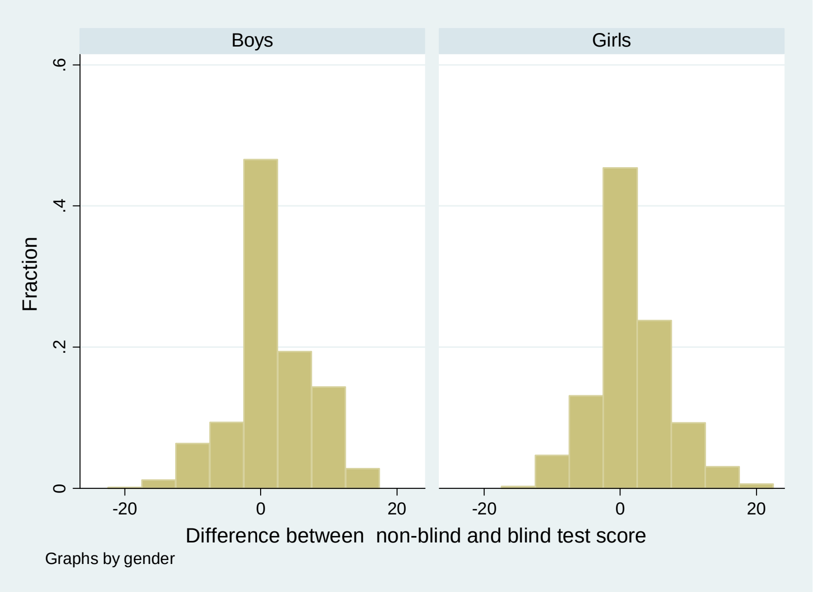

Figure 2 The distribution of the difference in test scores for the non-blind and blind grading procedures

Table 1 contains the summary statistics for the 1713 complete observations.[13] 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.[14] In line with previous studies, female students on average get higher grades than male students. The average non-blind test score is 15 % lower for boys than for girls 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 boys and girls, and the difference between the blind and the non-blind test score is almost identical for boys and girls. We thus find no evidence of discrimination. To further test the significance and robustness of the results we turn to the regression analysis results.

[13] 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.

[14] 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).

Table 1 Test scores and differences in test scores

Sample statistics	N	Mean	Std. Dev
Non-blind test score	1713	11.97607	4.999183
Blind test score	1713	10.4495	5.484892
Difference	1713	1.526562	
p-value of diff. (paired t-test)		<0.0001	
p-value of diff. (Wilcoxon test)		<0.0001	
Non-blind test score, boys	858	11.00816	5.072743
Non-blind test score, girls	855	12.94737	4.732002
Difference	1713	-1.930952	
p-value of diff. (t-test)		<0.0001	
p-value of diff. (Mann-Whitney test)		<0.0001	
Blind test score, boys	858	9.481352	5.591522
Blind test score, girls	855	11.42105	5.200705
Difference	1713	-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, boys	858	1.526807	5.906692
Non-blind test - Blind test score, girls	855	1.526316	5.526512
Difference	1713	.0004907	
p-value of diff. (t-test)		0.9986	
p-value of diff. (Mann-Whitney test)		0.6157	

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

4.2 Regression results

Table 2 presents the results from the estimation of the regression equation (4). The main 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 0.0004907 is that girls get about a .0005 lower non-blind test score on average due to their gender. The sign of this point estimate is not consistent with our hypothesis of discrimination of boys, but the estimate is very far from significant. Taken at face value, it suggests a discrimination effect of less than 0.005 % of the average non-blind test score. 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 boys or girls.

Table 2 Regression results on the effect of gender discrimination on the non-blind test score and robustness of the male indicator variable

Variables	1	2	3	4	5
Discrimination	.0004907	.0004907	.0004907	.0004907	.0234859
	(.2733793)	(.2863008)	(.3078964)	(.3240635)	(.3165188)
Male	-1.939701	-1.974668	-2.017402	-2.028517	-1.985367
	(.3031551)	(.2820646)	(.2895858)	(.3045571)	(.2990608)
Non-blind test	1.526316	1.526316	1.526316	1.526316	1.48503
	(.3739996)	(.3773052)	(.3963086)	(.3998068)	(.3753924)
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	3426	3426	3426	3426	3314
R^2	.0542	.1103	.2005	.2005	.2131

Note: 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)).

The other estimates in column 1 show that boys 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 boys than for girls, 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.[15]

In order to check for robustness of the estimate of γ we add fixed effects 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.[16] The estimate is very robust to the inclusion of these control variables. Note that the coefficients of discrimination and the non-blind test will not change in these additional estimations, by definition, since equation (4) saturates all these effects. The change of estimates in column 5 is only due to 56

[15] A difference in difference estimator as in equation (3) or equation (4) is mathematically equivalent to the difference of the difference of group means as reported in Table 1.

[16] Most of the students were born in the year 1987 (84%). Another 14 % were born in either 1986 or 1988. We lack data for 56 students. We also have month of birth for a smaller sub-sample. However, nothing substantial changes when adding it as a control.

missing observations on student year of birth. Reassuringly, the discrimination estimate is not substantially different in this sub-sample.

If the randomization was improper, then we could simply capture compositional effects. E.g. some schools have a conservative grading policy and if randomization failed then we might have disproportionably many boys or girls in these schools. The same argument holds for the reassessing teacher being conservative, and for the rewriting procedure and the age of student. This can be tested for by adding fixed effects for schools, re-grading teacher and rewriter and the year of birth of the student. However, as pointed out before, it is easier to use equation (3) for this purpose, since we are mainly interested in sensitiveness of the coefficient of discrimination. Note also that by adding controls to equation (3) we also allow for different schools or older students to have greater or smaller impact on the difference in grades, in addition to potential efficiency enhancements. Column 1-4 in Table 3 presents the results. The coefficient is robust and randomization seems to have worked properly and the main conclusion from Table 2 holds.

In general, a major concern with any non-blind/blind set up is that the blind assessor also can observe the variable that is supposed to be non-observable (gender in this study). It is reasonable that *some* students reveal their gender in their texts. This means a bias towards zero of the discrimination effect. With a larger number of observations, we could thus find a lower bound of the discrimination effect. It is reasonable that choosing some topics to write about could be correlated with the easiness of identify the gender of a student in the blind setting. For example, if the student write about alcohol (one of the 18 topics), then possibly gender could more easily be deduced since alcohol consumption differs across gender.

That some topics might carry a gender signal to the re-grader does not necessarily create a bias. For example, assume boys are discriminated in the non-blind setting. Then if a boy chooses a boyish subject then he will get the same grade from the re-grader, holding other determinants constant. But at the same time the girl that choose the boyish topic will be treated as a boy and get a decreased grade, leaving the coefficient on discrimination unchanged in a difference in difference set up *if* the proportion of girls and boys is representative.

Table 3 Robustness of the discrimination effect

Variables	1	2	3	4	5
Discrimination	.1701372	.0424138	.0417103	.1091431	.0098424
	(.236257)	(.2624792)	(.2639508)	(.2699595)	(.2938945)
Regrader fixed effect	Yes	Yes	Yes	Yes	Yes
School fixed effect	No	Yes	Yes	Yes	Yes
Re-writer fixed effect	No	No	Yes	Yes	Yes
Student year of birth	No	No	No	Yes	Yes
Topic fixed effect	No	No	No	No	Yes
N	1713	1713	1713	1657	1657
R^2	0.0958	0.2099	0.2188	0.2195	0.2335

Note: 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)).

To put it differently, controlling for the proportion of boys or girls in each topic would take care of this problem. Figure 3 shows that there are significant differences in the proportion of boys across topics. The most extreme topics a' priory also seem to attract girls or boys disproportional. Except for the topic on alcohol there is one topic on beauty, one on cellular phones, indicated in Figure 3 that shows clear gender marks.

However, if the choice of topic will affect the probability to discover the gender of students with certainty, then we need each topic to have its own intercept. Thus, including a fixed effect for the choice of topic, should serve as a reasonable robustness check for both the two problems. As discussed before, the students choose 1 out of 9 topics each time a test take place.[17] In column 5 a topic fixed effect for topics is added, and even though the coefficient changes somewhat, the previous conclusion still remain.[18]

[17] Since we have two rounds of test we observe 18 topics. Moreover, some students have failed to indicate topic chosen, which means we have another category of unknown topics.

[18] Note that given missing observation on some students year of birth, the estimate in column 5 in Table 3 should have the discrimination effect in column 5 in Table 2 as a benchmark.

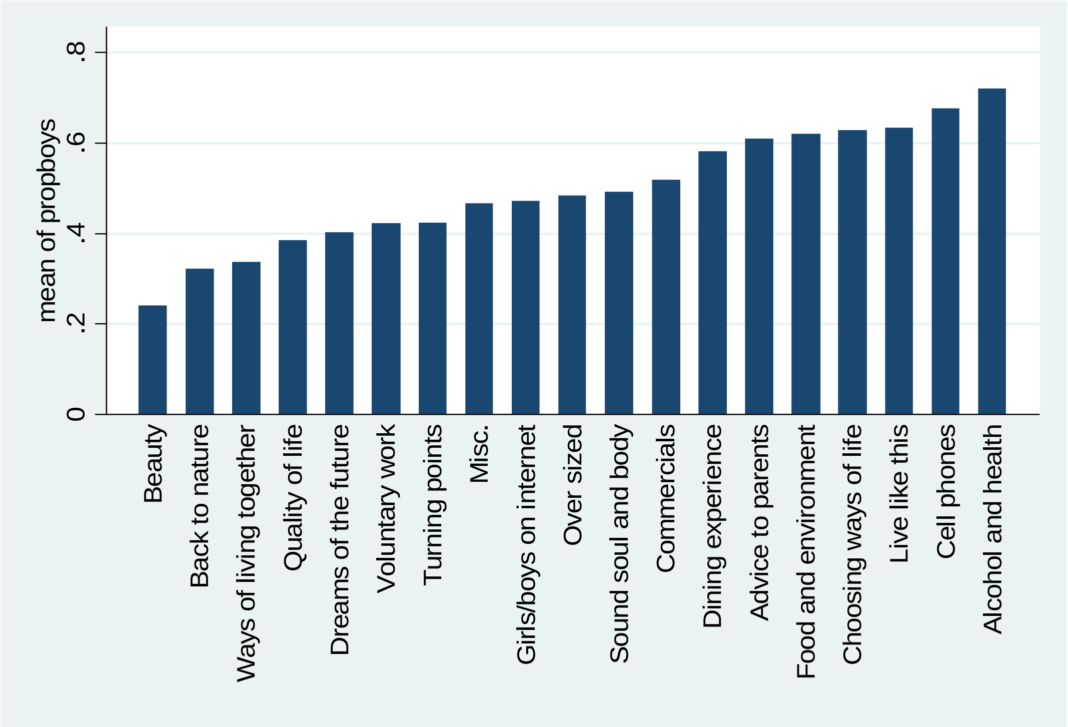

Figure 3 Proportion of boys for every topic

In general, a major concern with any non-blind/blind set up is that the blind assessor also can observe the variable that is supposed to be non-observable (gender in this study). It is possible that the re-grading teachers may be able to guess the gender of the student based on the text of the test. This could lead to a downward bias in our estimated discrimination effect. As the students choose among different topics, the choice of topic may reveal some information about gender. Figure 3 shows the fraction of boys in each topic and as can be seen in the graph this fraction varies between about 25 % and 70 % in the different topics. The topic “beauty” is least popular among boys and the topic “alcohol and health” is most popular. To control for the topic we add fixed effects for the topics in the final column in Table 3.[19] This has little effect on the results and the point estimate of the discrimination effect is still close to zero.[20]

[19] Since we have data from two rounds of the test we observe 18 topics. Moreover, as some students failed to indicate the chosen topic, we added a category for unknown topics (the Misc. category in Figure 3).

[20] Note that given the missing observation on some student’s year of birth, the estimate in column 5 in Table 3 should have the discrimination effect in column 5 in Table 2 as a benchmark.

4.3 Extensions

As explained in section 2 students in the Swedish high school system can chose between two types of high school programs: academic track or vocational training. It is possible that the discrimination effect could differ between these two sub-groups. We therefore, as a further robustness check, estimate our results separately for academic track and vocational training students using equation (3).[21] The results are presented in Table 4. 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. Moreover, the point estimate of discrimination in the academic track is decreased by more than 50 % when adding the full set of controls. We also test if the coefficient of the discrimination variable differs significantly between the two groups, but this difference is also far from significant.[22]

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) using equation (4) with maximum likelihood (Long & 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.[23] The interval regression results for the estimate of discrimination effect are shown in the last column in Table 4. Although the sign of the discrimination coefficient shifts from positive to negative, the estimated effect is still close to zero and far from significant.[24]

[21] We have not been able to get information on vocational and academic tracks for the full sample. Thus, we miss some observations. The share of girls on academic track is 50.5% and the share in vocational training is 49.5 %.

[22] The p-value of a z-test of the difference in the discrimination coefficient between equation 1 and 3 in Table 4 is 0.272 and the p-value of a z-test of the difference between equation 2 and 4 in Table 4 is 0.407.

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

[24] The male indicator variable in the interval regression is - 2.00 compared to -1.93 in the OLS regression, and the coefficient of the Non-blind test variable is 1.64 in the interval regression compared to 1.53 in the OLS.

Table 4 Extensions

Variables	1	2	3	4	5
	Academic track		Vocational track		Interval regression
Discrimination	-.2294904	-.0948175	.42799	.5058355	-.1105371
	(.3863124)	(.4949792)	(.4567069)	(.529908)	(.2886335)
Full sets of controls	No	Yes	No	Yes	No
N	791	770	694	672	3426
R^2	0.0004	0.3759	0.0013	0.3664	

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. Clustering at the re-grader level gives somewhat lower standard errors.

5 Concluding remarks

Our study contributes to the increasing literature testing for discrimination in economics (Ayres & Siegelman, 1995; Ladd, 1998; Szymanski, 2000; Bertrand & Mullainathan, 2004). We failed to find any evidence of discrimination of boys 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 & 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 & 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 girls outperform boys at school 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 girls provide more effort in school. To investigate why this is the case and to what extent it varies with different learning environments is crucial for the design of policies aimed at decreasing 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(3), 304-321.

Bedard, K., & Cho, I. (2010). Early gender test score gaps across OECD countries. *Economics of Education Review*, 29(3), 348-363.

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.

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

Castagnetti, C., & Roste, L. (2009). Effort allocation in tournaments: The effect of gender on academic performance in Italian universities. *Economics of Education Review*, 28(3), 357-369.

Dee, T.S. (2005). A teacher like me: does race, ethnicity or gender matter? *American Economic Review Papers and Proceedings*, 95(2), 158-165.

Feingold, A. (1988). Cognitive gender differences are disappearing. *American Psychologist*, 43(2), 95-103.

Guiso, L., Monte, F., Sapienza, P., & Zingales, L. (2008). Culture, gender and math. *Science*, 320(5880), 1164-1165.

Guo, C., Tsang, M.C., & Ding, X. (2010). Gender disparities in science and engineering in Chinese universities. *Economics of Education Review*, 29(2), 225-235.

Hajj, M., & Panizza, U. (2009). Religion and education gender gap: Are muslims different? *Economics of Education Review*, 28(3), 337-344.

Holmlund, H., & Sund, K. (2008). Is the gender gap in school performance affected by sex of the Teacher. *Labour Economics*, 15(1), 37-53.

Hyde, J.S., Fennema, E., & Lamon, S.J. (1990). Gender differences in mathematics performance: a meta-analysis. *Psychological Bulletin*, 107(2), 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(5888), 494-495.

Jacob, B.A., & Levitt, S.D. (2003). Rotten apples: an investigation of the prevalence and predictors of teacher cheating. *Quarterly Journal of Economics*, 118(3), 843-877.

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

Lai, F. (2010). Are boys left behind? The evolution of the gender achievement gap in Beijing's middle schools. *Economics of Education Review*, 29(3), 383-399.

Lavy, V. (2008). Do gender stereotypes reduce girls' human capital outcomes? Evidence from a natural experiment. *Journal of Public Economics*, 92(10-11), 2083-2105.

Levitt, S.D., & List, J.A. (2007). What do laboratory experiments measuring social preferences reveal about the real world? *Journal of Economic Perspectives*, 21(2), 153-174.

Lindahl, E. (2007). *Comparing teachers assessments and national test results: Evidence from Sweden*. IFAU Working Paper 2007:24.

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

Ministry of Education. (2004). *Könsskillnader i utbildningsresultat* (in Swedish). Ministry of Education report series 2004, report 7. Stockholm: Ministry of Education.

Swedish National Agency for Education. (2006). *Könskillnader i måluppfyllelse och utbildningsval* (in Swedish). Technical Report 286. Stockholm: Swedish National Agency for Education.

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

Thompson, S.B. (2009). *Simple formulas for standard errors that cluster by both firm and time*. Working paper.

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(3), 309-322.

Publication series published by the Institute for Labour Market Policy Evaluation (IFAU) – latest issues

Rapporter/Reports

2010:1 Hägglund Pathric ”Rehabiliteringskedjans effekter på sjukskrivningstiderna”

2010:2 Liljeberg Linus and Martin Lundin ”Jobbnätet ger jobb: effekter av intensifierade arbetsförmedlingsinsatser för att bryta långtidsarbetslöshet”

2010:3 Martinson Sara ”Vad var det som gick snett? En analys av lärlingsplatser för ungdomar”

2010:4 Nordström Skans Oskar and Olof Åslund ”Etnisk segregation i storstäderna – bostadsområden, arbetsplatser, skolor och familjebildning 1985–2006”

2010:5 Johansson Elly-Ann ”Effekten av delad föräldraledighet på kvinnors löner”

2010:6 Vikman Ulrika ”Hur påverkar tillgång till barnomsorg arbetslösa föräldrars sannolikhet att få arbete?”

2010:7 Persson Anna and Ulrika Vikman ”In- och utträdeseffekter av aktiveringskrav på socialbidragstagare”

2010:8 Sjögren Anna ”Betygsatta barn – spelar det någon roll i längden?”

2010:9 Lagerström Jonas ”Påverkas sjukfrånvaron av ekonomiska drivkrafter och arbetsmiljö?”

2010:10 Kennerberg Louise and Olof Åslund ”Sfi och arbetsmarknaden”

2010:11 Engström Per, Hans Goine, Per Johansson, Edward Palmer and Pernilla Tollin ”Underlättar tidiga insatser i sjukskrivningsprocessen återgången i arbete?”

2010:12 Hensvik Lena ”Leder skolkonkurrens till högre lärarlöner? – En studie av den svenska friskolereformen”

2010:13 Björklund Anders, Peter Fredriksson, Jan-Eric Gustafsson and Björn Öckert ”Den svenska utbildningspolitikens arbetsmarknadseffekter: vad säger forskningen?”

2010:14 Hensvik Lena and Peter Nilsson ”Smittar benägenheten att skaffa barn mellan kollegor?”

2010:15 Martinson Sara and Kristina Sibbmark ”Vad gör de i jobb- och utvecklingsgarantin?”

2010:16 Junestav Malin ”Sjukskrivning som politiskt problem i välfärdsdebatten – det politiska språket och institutionell förändring”

2010:17 Hägglund Pathric and Peter Skogman Thoursie ”Reformerna inom sjukförsäkringen under perioden 2006–2010: Vilka effekter kan vi förvänta oss?”

2010:18 Sibbmark Kristina ”Arbetsmarknadspolitisk översikt 2009”

2010:19 Ulander-Wänman Carin ”Flexicurity och utvecklingsavtalet”

2010:20 Johansson Per and Erica Lindahl ”Informationsmöte – en väg till minskad sjukskrivning?”

2010:21 Grönqvist Erik, Jonas Vlachos and Björn Öckert ”Hur överförs förmågor mellan generationer?”

2010:22 Martinson Sara och Kristina Sibbmark ”Vad gör de i jobbgarantin för ungdomar?”

2010:23 Hinnerich Tyrefors Björn, Erik Höglin och Magnus Johannesson ”Diskrimineras pojkar i skolan?”

Working papers

2010:1 Ferraci Marc, Grégory Jolivet and Gerard J. van den Berg “Treatment evaluation in the case of interactions within markets”

2010:2 de Luna Xavier, Anders Stenberg and Olle Westerlund “Can adult education delay retirement from the labour market?”

2010:3 Olsson Martin and Peter Skogman Thoursie “Insured by the partner?”

2010:4 Johansson Elly-Ann “The effect of own and spousal parental leave on earnings”

2010:5 Vikman Ulrika “Does providing childcare to unemployed affect unemployment duration?”

2010:6 Persson Anna and Ulrika Vikman “Dynamic effects of mandatory activation of welfare participants”

2010:7 Sjögren Anna “Graded children – evidence of longrun consequences of school grades from a nationwide reform”

2010:8 Hensvik Lena “Competition, wages and teacher sorting: four lessons learned from a voucher reform”

2010:9 Hensvik Lena and Peter Nilsson “Businesses, buddies and babies: social ties and fertility at work”

2010:10 van den Berg Gerard J., Dorly J.H. Deeg, Maarten Lindeboom and France Portrait “The role of early-life conditions in the cognitive decline due to adverse events later in life”

2010:11 Johansson Per and Erica Lindahl “Can sickness absence be affected by information meetings? Evidence from a social experiment”

2010:12 Grönqvist Erik, Björn Öckert and Jonas Vlachos “The intergenerational transmission of cognitive and non-cognitive abilities”

2010:13 de Luna Xavier, Per Johansson and Sara Sjöstedt-de Luna “Bootstrap inference for *K*-nearest neighbour matching estimators”

2010:14 Hinnerich Tyrefors Björn, Erik Höglin och Magnus Johannesson “Are boys discriminated in Swedish high schools?”

Dissertation series

2010:1 Johansson Elly-Ann “Essays on schooling, gender, and parental leave”

2010:2 Hall Caroline “Empirical essays on education and social insurance policies”