Biased grades? Changes in grading after a blinding of examinations reform

Magnus Bygren

Stockholm University, Sweden; Institute for Future Studies, Stockholm, Sweden

ABSTRACT
Group differences in average grades prior to and after a step-wise introduction of blinded examinations at Stockholm University are examined. Relative to students with ‘native’ names, students with ‘foreign’ names appear to experience weak positive bias in the grading of their examinations, but the estimated effect is sensitive to model specification. No substantial effects of blinding examinations with respect to male-female gaps are found. The results suggest that examiners – when the names of students are disclosed to them – if anything have a weak tendency to positively discriminate for students perceived to have an immigrant background, but they do not appear to discriminate on the basis of gender.

KEYWORDS
bias; discrimination; gender; migrant background; grading

Introduction
Grades given by teachers are crucially important for students, and directly affect their academic self-concepts, educational attainment and job chances. If there is bias in grading, this may affect individual life chances in a rather direct way, and stratification patterns more broadly. Humans are subject to some well-known biases when making decisions, and in this study I ask whether there is bias in the grading of examinations in higher education. Grading bias may be negative or positive, and implies that the grading of students in addition to performance is affected by the students’ membership in certain social categories.

Observing discrimination and disentangling it empirically from correlated non-discriminatory behaviours is a major challenge for research. I adopt a natural experiment design, exploiting that written examinations have been blinded step-wise at Stockholm University. Written examination is one of the more common ways of assessing students in higher education, and I compare the development of such examination results before and after the blinding of examinations for groups defined by their gender and ‘foreignness’. Although gender, ‘foreignness’ and other kinds of background characteristics to some extent can be inferred from handwriting and revealing idioms, blinded written examinations make it considerably more difficult for teachers to precisely target any conscious, or subconscious, discrimination or biases. The empirical strategy in this article therefore consists in comparing average differences in grades between groups prior to and after the introduction of blinded examinations, differencing out the grade difference development on courses in which blinded examinations were used throughout the period. The study
contributes more broadly to the literature on bias and discrimination in grading by providing a stringent empirical test of the effects of blind assessment compared to non-blind assessment.

**Mechanisms**

Different explanations have been proposed to describe what may produce discriminatory behaviours. Discriminatory bias implies that one or several irrelevant criteria have been used by an agent in an action or decision affecting a target, and the target of the biased decision in some way benefits or is harmed by the outcome of the action or decision. This type of bias may be more or less organizationally or institutionally embedded (Feagin & Eckberg, 1980), but most explanations of such discrimination focus on individual-level reasons, based on cognitive bias, favouritism/in-group bias or information bias.

Status-based discrimination theory takes as its point of departure that expectations ascribe levels of skill based on an individual’s personal characteristics (Berger, et al., 1977). Individuals with highly ranked status characteristics are expected to perform better, and the evaluation of their performance will therefore be more lenient, and vice versa. That is, expectations are biased in favour of a particular group. Being assessed to be skilled is to a certain extent dependent on culturally and traditionally defined ‘suitable properties’ in relation to the context at hand (Correll & Benard, 2006). Stereotypes of individuals as representatives of groups come into play in such evaluation contexts, and equivalent achievements may be judged differently because evaluation standards vary between groups with different status characteristics (Foschi, 1996). If the context of evaluation is traditionally masculine, men tend to be evaluated favourably, and vice versa (Ridgeway & Smith-Lovin, 1999; Wagner & Berger, 1997). These norms may also influence the ego’s own performance expectations, and thereby indirectly impact performance itself, and lead to self-fulfilling feedback loops (Correll, 2004; Steele, 2003)

**Biases in grading in higher education**

In contrast to the domain in which discrimination theories are usually applied – the labour market in general and recruitment situations in particular – the evaluation of written examinations in higher education is a situation less social and more formal. Discrimination in this setting implies a bias in the assessment of individual students because of their social category (Pager & Shepherd, 2008). In written examinations in higher education, the examiner has the task of evaluating and grading written answers to questions and for the examiner the stakes are relatively low as the outcome of such an evaluation is neither an employment relation, nor another form of enduring relation, at least not as a rule. The institutional setting of the evaluation is one in which a great deal of stress is placed on impartiality and the appreciation of individual achievement with ‘universalism’ as one of the institutional imperatives of modern science Merton (1973[1942]).

A consistent finding in studies of attitudes to immigrants and ethnic minorities is that the higher the level of education, the lower the level of expressed intolerance (Ceobanu & Escandell, 2010), and exposure to university education seems critical for instilling tolerant attitudes (Hainmueller & Hiscox, 2007). If anything, we would not expect intolerance towards women, or minorities, to be prevalent among university teachers. Further, university teachers appear to be more politically liberal in their attitudes by comparison with the population in general, and are supportive of anti-discrimination legislation (Klein & Stern, 2005).

Relatively, Behaghel, Crépon and le Barbanchon (2015) found evidence that employers who volunteered to subject job applicant resumes to blinding became less likely to interview and hire foreign candidates, the explanation being that employers who volunteered to participate in the experiment – which was designed to minimize discrimination – were a select group in the sense that they used a more lenient standard for ‘foreigners’, and had been more likely to hire foreign
candidates prior to participation. Once résumés were blinded, they became unable to attenuate negative résumé signals from foreigners and the hiring rates for these declined as a consequence. Reverse discrimination may in other words be practiced in certain tolerant contexts, which higher education can be seen as exemplifying.

**Empirical studies of biases in grading**

The weight of evidence in the small literature comparing blind to non-blind assessment suggests that grading bias in this context cannot be denied (Brennan, 2008). Although a nontrivial portion of studies report null findings, a tendency is that when there is evidence of bias, it disadvantages lower performing groups, such as boys, immigrants and children from less advantaged families.

Regarding **gender bias**, Lavy (2008) compared blind test results to non-blind among Israeli high school students, and found teachers to discriminate against boys. Comparing remotely marked written examinations to teacher assessments of British 11-year old students, Burgess & Greaves (2013) found boys to be penalized relative to girls. Cornwell, Mustard & Van Parys (2013) found that for primary schools in the U.S., boys who perform equally well as girls on reading, mathematics and science tests are graded less favourably by their teachers. This less favourable treatment vanishes once non-cognitive skills are taken into account, suggesting that orderly behaviours – more prevalent among girls – are rewarded when teachers grade. For U.S. pupils in elementary and middle school, girls tended to be rated by their teachers as more knowledgeable than boys in mathematics, in excess of direct cognitive assessments (Robinson & Lubinski, 2011).

Identifying bias by systematic differences in students’ scores between oral tests (non-blind toward gender) and anonymous written tests in the entrance examinations to the elite school Ecole Normale Superieure in France, Breda & Ly (2012) found raters to favour females in more male-connoted subjects, and males in more female-connoted subjects. Relative to remotely graded bachelor’s theses in Poland, graders who knew the students systematically over-rated female students (Krawczyk, 2018). Hinnerich et al. (2011) investigated the prevalence of this in the assessment of Swedish high school students by comparing two assessments of a compulsory national test in Swedish language proficiency. The students’ average test score dropped when the examinations were blinded, but there was no evidence of gender bias. Hinton & Higson (2017) similarly report a null effect of anonymization on gender differences in grades for UK undergraduates, as do Pitt & Winstone (2018).

Regarding **discrimination against minority groups**, Van Ewijk (2011) did not find any grading bias against essays of Dutch pupils assigned ethnic minority names compared to essays assigned ethnic majority names. Similarly, in a large-scale examination of assessments made at a UK university, there was no effect of blinding the examinations on ethnic differences in grades (Hinton & Higson, 2017). However, Hinnerich et al. (2015) found that the difference in performance between high school students of Swedish and foreign background increased in the context of non-blind assessment, i.e., that students with a foreign background appear to be discriminated against when their names were disclosed in the assessment of their tests. Sprietsma (2013) found that if an essay had a Turkish name assigned to it, this had a small but significant negative effect on German teachers’ perceived quality of it.

Burgess & Greaves (2013) found teachers over and underassessments to go in different directions depending on the ethnic group. Relative to remotely marked examinations, teachers systematically over assessed British 11-year students with Asian backgrounds, in particular in mathematics and science, but under assessed black Caribbean and African students. In a meta-analysis of 20 experimental studies Malouff & Thorsteinsson (2016) found blinding to benefit minority groups, but the effect was imprecisely estimated with a rather large confidence band.
Theorizing bias in grading

As noted, the blinding of examinations may have different consequences depending on priors of teachers and how these priors are related to the groups of students having their examinations blinded. When a teacher assesses a non-blinded written examination, we may imagine that the answers to be graded to some extent are framed differently depending on the combination of social categories the student belongs to. As emphasized by scholars working in the intersectional tradition, people identify with, and are categorized into, multiple social categories, and suffer or benefit from the status of these categorizations combined. Assessments are affected by culturally and traditionally defined ‘suitable properties’ in relation to the context at hand (Correll & Benard, 2006), and persons exhibiting contextually unusual properties may, consciously or sub-consciously, be assumed to be (contextually) less competent, or held to a harsher standard than groups with contextually expected properties (Batruch, 2018; Hanna & Linden, 2012). Empirically, it is however hard to disentangle mechanisms. Such bias may be produced by cognitive frame filtering, because of in-group biases favouring those with suitable properties, or perhaps because these on average get better test results, and teachers therefore use the short-cut of using group-level information to grade individuals (i.e., statistical discrimination).

The reverse might also be true; teachers wanting to compensate and help groups unaccustomed with higher education, or unprivileged groups generally, by using a more lenient standard for them (Behaghel, et al., 2015; Cornwell, et al., 2013), which might be a particularly salient tendency in a liberal setting such as a university (Klein & Stern, 2005). If such ‘reverse’ discrimination is practiced, it would perhaps not be surprising to find it in Sweden, a country where the population expresses relatively tolerant attitudes towards immigrants (Gorodzeisky & Semyonov, 2016), and has one of the world’s most gender equal attitudes (Brandt, 2011; Inglehart & Norris, 2003). Thus, in an international perspective, to the extent that explicit attitudes correspond with behaviours, we would expect less, or even reversed, discrimination based on gender and ethnicity in this country than in many other countries.

Case and design

It is rather difficult to observe and quantify discrimination. It is a sensitive issue and few would admit that they deliberately discriminate. There is also reason to believe that such biases are practiced at a more or less subconscious level (Cunningham, et al., 2004), making it extremely hard to prove. One strategy that has been used to measure discrimination is to compare differences in outcomes between blinded and non-blinded assessments of performance. In non-blinded testing, a rater has information about the student’s identity, while this information is hidden from a rater in blinded assessments. Any biases are captured by comparing group differences in ratings between blinded and non-blinded assessments. In the present study I exploit a natural experiment to study the effects of such blinding - in this case the blinding of written examinations - that potentially affects the grading of student examinations.

The empirical case is the introduction of blinded examinations at Stockholm University in 2009. Before the introduction of blinded examinations, students wrote their name and a personal identification number (a ten-digit number used in all Swedish public registers, including six digits indicating date of birth) on a cover sheet and at the top of every answer sheet. The examiner therefore had the opportunity to infer the students’ gender, ‘foreignness’ and age from this information. With the introduction of blinded examinations, the students’ names and personal identification numbers were replaced with a serial number, and the examiners’ opportunities to identify the students’ social category placement were consequently much reduced. The serial numbers are linked back to the student by administrative staff when the grade is registered. The treatment group in this study consists of students studying economics, political science and
sociology, whose examinations were assessed non-blinded until the spring of 2009, but were blinded from the fall 2009 and onwards.

A comparison of grades before and after the blinding reform may confound discrimination with the influence of other time-varying factors, e.g., unobserved cohort differences. To net out such confounding, I use law students at Stockholm University as a control group because their examinations were blinded both prior to and after 2009, and they were consequently unaffected by this change in policy, the assumption being that effects of such time-varying factors are equal across the treatment and control groups. More specifically, I use a so-called difference-in-differences-in-differences (DDD) design. Generically the treatment effect on the grade gap between students belonging to a certain social category \( (X=1) \) versus students not belonging to this social category \( (X=0) \) can be expressed as:

\[
\text{DDD} = \left( \frac{\bar{Y}_{\text{TREAT}=1,X=1,\text{POST}=1} - \bar{Y}_{\text{TREAT}=1,X=1,\text{POST}=0}}{\bar{Y}_{\text{TREAT}=0,X=1,\text{POST}=1} - \bar{Y}_{\text{TREAT}=0,X=1,\text{POST}=0}} \right) - \left( \frac{\bar{Y}_{\text{TREAT}=0,X=1,\text{POST}=1} - \bar{Y}_{\text{TREAT}=0,X=1,\text{POST}=0}}{\bar{Y}_{\text{TREAT}=0,X=0,\text{POST}=1} - \bar{Y}_{\text{TREAT}=0,X=0,\text{POST}=0}} \right)
\]

where TREAT is a dummy for non-law students, X a social category dummy, and POST a dummy for the post-reform period. This means that I follow the treatment and control groups over time, examine the gaps in the respective groups to observe whether the gap in the treatment group change – relative to the gap in the control group – with the treatment. The DDD model may also be expressed as a regression equation with a saturated model including main effects, two-way interaction effects, and a triple interaction between TREAT, X and POST, where this interaction is identical to the DDD estimate above.

The assumptions we need to make to be able to interpret this coefficient as causal are that in the absence of the reform, the gap in grade averages between students with \( X=0 \) and students with \( X=1 \) is merely a function of: i) a time-invariant difference between the control and treatment group, and ii) a time effect common to both the treatment and the control group. The second assumption is important; the reform must not coincide with other important determinants of the grade gap being estimated. If so, the effect of the blinding of examinations may be either over or underestimated in unknown ways. In this regard, coincidental changes in the selection of students and/or teacher composition constitute potential threats to causal inference, as do changes in curricula and examination style that benefit just one of the groups in one of the treatment conditions.

Lacking detailed information about these potential sources of bias, the plausibility of these assumptions are hard to evaluate for the present case. As law students may be selected to university studies in a pattern distinct from that of social science students, one may also question this choice of control group. However, a rather large portion of students in our data, one-fourth, took law as well as non-law courses during the observation period, indicating that the two groups are to a nontrivial degree overlapping. Exploratory analysis of the data indicated that students taking at least one law course are positively selected (i.e., they receive higher grades) compared to the stable social science students, but there was no indication of change in the strength of this positive selection during the observation period.

**Materials and methods**

I extracted the data from the Swedish national Ladok register and they include all grades recorded for five introductory-level courses at the undergraduate level in law, economics, political science and sociology at Stockholm University, between spring 2005 and fall 2013. Ladok is a national system for the administration of studies in higher education in Sweden and is maintained by a consortium of the universities and the National Board of Study Aid. Transcripts are publicly available through the so-called publicity principle that Swedish authorities, including universities, abide to. I collected data for nine semesters prior to and nine semesters after the
blinding of written examinations. As mentioned above, law courses used blinded examinations throughout the period. To isolate the effect of anonymization, I chose courses with large numbers of students, i.e. courses where the only likely source of information on the students’ background come from the information conveyed in the written examination. All courses are taught in Swedish and the final grading is largely or completely based on the final written examination. A course director usually leads the assessment of examinations on these large, introductory-level courses, but teams of teaching assistants do the bulk of the assessments.

At the level of the individual student, I have information on the student’s first name, surname, gender and year of birth. This is also the only information about the student’s identity that the examiner receives in non-blinded written examinations. I define an observation as a student-grade at an examination. In total, there are 25,077 student-grade observations, for 17,235 unique students. A student can appear multiple times in the data, because s/he either has failed an examination at one point in time and has then re-taken the same examination, or received a grade from at least two different courses during the period examined. If a student fails, s/he has the opportunity to re-take the examination until s/he receives a passing grade. For the large introductory courses in the data, there is at least one additional opportunity per semester for a failed student to re-take an examination. As a test of robustness I also estimated individual panel models with individual fixed effects where I in effect only exploit within-individual grade variation from individuals who took an examination both prior to and after the introduction of blinded examinations.

**Dependent variable**

The outcome variable is a measure of the student’s course-grade, which I recoded to be consistent across time and courses. I used the lowest common denominator, which is a three-point scale: Pass with distinction, Pass, and Fail. During the period examined, the social sciences courses (the treatment group) used a three-point grading scale until spring 2007, and a seven-point grading scale from fall 2007 and onwards. The law courses (the control group) used a three-point scale, a four-point scale and a seven-point scale. I use the lowest common denominator of these scales, which is a three-point scale: Pass with distinction (2), Pass (1), and Fail (0). The highest grade on the four-point scale, and the two highest grades on the seven-point scale were defined as Pass with distinction. I collapsed the mid-range categories on both scales (two categories for the four-point scale and three categories for the seven-point scale) into Pass, whereas I retained the Fail categories throughout.

**Independent variables**

‘Foreignness’ was coded based on the students’ surnames, and is a binary variable distinguishing between native-sounding and foreign-sounding names, the details of which I give below. Woman has the value 1 if the student is a woman and 0 if the student is a man. This variable derives from gender as recorded in Ladok. Treatment group has the value 1 for economics, political science and sociology, and 0 for law. Post is a dummy variable with the value 0 for all observations between fall 2005 and spring 2009, and the value 1 for observations from fall 2009 to fall 2013. I measure Age in integer years.

**The classification of names as foreign**

Two research assistants classified student surnames as ‘foreign’ or ‘native’. They coded names linguistically belonging to the Swedish, Norwegian, Danish, Icelandic, English, Dutch, German, French or Finnish languages as ‘native’. This broader conceptualisation of the native category...
was used as there is reason to believe that Western European names signal a higher degree of social integration and status compared to other non-Scandinavian sounding names, since the economic integration of immigrants from these countries is more or less on a par with that of the Swedish-born population (Le Grand & Szulkin, 2002). Sweden has been belonged to the Western European sphere through trade and migration at least since medieval times. For example, many names originally carried by German immigrants are for all practical purposes nowadays considered as native by Swedes today (Brylla, 2011).

I ran a robustness check with Scandinavian names versus all other names categorized as ‘foreign’, and the results turned out to be extremely similar to those reported below. I do not report this result but it is available upon request. We categorized names interpreted as linguistically originating from languages other than those of the countries listed above as belonging to the ‘foreign’ name category. For the ‘native’ category, the three most common names in the dataset are Andersson, Eriksson and Johansson. These are also the most common surnames in Sweden. The three most common ‘foreign’ surnames in the data are Ali, Wang and Khan. The inter-coder agreement was 98.4%, indicating the classifications to be highly consistent. We discussed and jointly decided the classification of all off-diagonal names. I report descriptives for the all variables in the analytical sample in Table 1.

### Results

To give an overview of the patterns in the data, I report grade differences between groups, before and after the blinding of examinations, in Table 2.

Regardless of whether students are in the treatment or control group, ‘natives’ tend to receive better grades than ‘foreigners’. Males receive somewhat higher grades than females, but the gender difference is much smaller than the foreign-native difference. It can also be noted that average grades decreased over the period, and more so in the control group. It became more common to fail an examination, and less common to receive the highest grade. In the treatment group, the decline in average grades was larger in the ‘foreign’ group than in the ‘native’ group.

In Figure 1, I report the development over time in the treatment-control group difference for the ‘foreign-native’ gap in the upper panel. Each point estimate stands for the treatment to control group difference in the average between-group probability gap to receive a certain grade. Positive values indicate that ‘foreigners’ have over-risks relative to ‘natives’ to receive the specific

| Table 1. Descriptives (N = 26,475). |
|-----------------------------------|
| **Grade**                        |
| Fail                             | 24.14 |
| Pass                             | 53.34 |
| Pass with distinction            | 22.52 |
| **Foreignness**                  |
| ‘Native’ sounding name           | 75.52 |
| ‘Foreign’ sounding name          | 24.48 |
| **Gender**                       |
| Man                              | 43.83 |
| Woman                            | 56.17 |
| **Treatment group**              |
| Economics                        | 58.97 |
| Political science                | 32.85 |
| Sociology                        | 11.58 |
| **Control group**                |
| Law                              | 43.27 |
| Post                             | 56.73 |
| Age (mean, sd)                   | 23.94, 5.22 |
grade in the treatment group compared to the control group, and negative values indicate the reverse of this.

Not much happens to the treatment group ‘foreigners’ relative risks of being failed after entering the treatment period. The confidence bands are overlapping and point estimates are

---

**Table 2.** Grade distribution before and after blinding of examinations (percentages).

|                  | Control Group | Treatment Group |
|------------------|---------------|-----------------|
|                  | ‘Foreign’ name | ‘Native’ name   | ‘Foreign’ name | ‘Native’ name   |
| **Pre-treatment period** |               |                 |               |                 |
| Fail             | 30.0          | 13.6            | 35.9          | 19.1            |
| Pass             | 53.7          | 54.6            | 48.4          | 50.7            |
| Pass with distinction | 16.2        | 31.7            | 15.8          | 30.2            |
| N                | 906           | 3597            | 1553          | 4882            |
| **Treatment period** |               |                 |               |                 |
| Fail             | 43.9          | 21.2            | 42.8          | 20.6            |
| Pass             | 46.4          | 60.0            | 46.9          | 54.8            |
| Pass with distinction | 9.7         | 18.8            | 10.4          | 24.5            |
| N                | 1511          | 4849            | 2509          | 6668            |
| Women            | 16.6          | 17.3            | 24.2          | 21.7            |
| Men              | 54.6          | 54.2            | 50.5          | 49.8            |
| Pass             | 28.8          | 28.4            | 25.3          | 28.5            |
| N                | 2549          | 1954            | 3646          | 2789            |
| Women            | 27.2          | 25.6            | 28.1          | 25.1            |
| Men              | 55.0          | 59.4            | 50.6          | 55.0            |
| Pass with distinction | 17.8        | 15.0            | 21.3          | 19.9            |
| N                | 3785          | 2575            | 4892          | 4285            |

---

**Figure 1.** Point estimates and confidence intervals of the difference between the treatment group and control group with regard to the ‘foreign’ to ‘native’ grade gaps (upper panel), with regard to the female to male grade gaps (lower panel), before (pre-treat) and after (post-treat) the blinding of examinations in the treatment group.
close to zero both before and after the blinding of examinations. However, the treatment group ‘foreigners’ relative chances of receiving pass increased with the blinding of examinations. All the same, their relative chances of receiving a pass with distinction decreased, implying that with the blinding of examinations, ‘foreigners’ experienced a decline in their average passing grades. Judging by the degree of confidence interval overlap, this change appears to be statistically significant when we compare the pre-treatment period to the treatment period.

With regard to the development of the female to male gap, reported in the lower panel in Figure 1, there is no evidence pointing in the direction of an effect of blinding the examinations on female to male differences in grades. The development of this gap is very similar across the treatment and control groups, as confidence bands are heavily overlapping. That is, there are no significant changes with regard to female advantage, or disadvantage, because of blinded examinations.

As a robustness check, I estimated logistic regression models instead of linear probability models. With regard to the direction of effects and statistical significance, the results from these models were very similar to those reported here.

In Table 3, I report the difference-in-difference-in-differences estimates of the treatment effect on the gaps. In columns 1 and 4, these are identical to the differences between the point estimates reported in Figure 1. There are two statistically significant estimates: a positive one at .072 for ‘foreigners’ to receive pass, and a negative one at .062 for ‘foreigners’ to receive pass with distinction. Relative to the control group, ‘foreigners’ in the treatment group thus decreased their probability of receiving the highest grade and increased their probability of receiving a passing grade by roughly the same amount. This is a rather substantial effect given that the baseline chance of a treatment group ‘foreigner’ to receive the highest grade in the pre-treatment period was 16 percent. In columns 2 and 5, I report estimates conditioned on age, year dummies and treatment group discipline, but these are very close to the unconditional estimates.

In order to conduct a more stringent test of whether the blinding of examinations affected the grade gaps, I additionally made use of having for some students observations both before and after the introduction of blinded examinations. However, the number of observations used for estimating these effects is much smaller; such longitudinal information was available for just 392 students in the control group and 510 students in the treatment group. Therefore, I cannot estimate the effect of blincling examinations with a great deal of precision, and the confidence intervals of the estimated treatment effects overlapped zero by large margins (see columns 3 and 6 in Table 3).

Table 3. Difference-in-difference-in-differences estimates, linear probability models.

|                  | X = ‘Foreign’ name |                  | X = Woman |
|------------------|-------------------|------------------|----------|
|                  | (1) (2) (3) (4) (5) (6) |
| Outcome: Fail    |                   |                   |          |
| Treatment*Post*X | -.010 (.033)      | -.015 (.032)      | -.001 (.055) | -.018 (.025) | -.012 (.025) | .025 (.047) |
| Outcome: Pass    |                   |                   |          |
| Treatment*Post*X | .072* (.032)      | .077* (.032)      | .030 (.064) | -.004 (.027) | -.012 (.026) | .064 (.056) |
| Outcome: Pass with Distinction | | | | | | |
| Treatment*Post*X | -.062** (.023)    | -.062** (.023)    | -.029 (.037) | .022 (.023) | .024 (.023) | .038 (.032) |
| Controls         | NO YES YES        | NO YES YES        | NO YES YES | NO NO NO YES      |
| Individual fixed effects | NO NO YES | NO NO NO YES |
| N                | 26475 26475 26475 | 26475 26475 26475 | 26475 26475 26475 |

*p<.05, **p<.01.
Notes: Control variables include age, treatment group discipline (economics/political science/sociology), and year. Additionally, gender is a control variable in models 1–3 and ‘foreignness’ is a control variable in models 4–6. All standard error estimates are robust to heteroscedasticity.
To check for any pre-treatment trends in the data, I estimated the effects reported in Figure 1 for each semester in the data (not reported). With regard to the foreign to native gaps, pre-treatment trends in the data appeared to be present for the pass and the pass with distinction outcome, which casts a certain amount of doubt as to whether I can treat the statistically significant effect reported in Table 3 as causal. For the gender differences, no trends were discernible in the data.

Discussion

In this study, I asked whether written examinations in higher education are assessed differently depending on the ‘foreignness’ and gender of the assessed students. I did not find any support for the hypothesis that students with ‘foreign’ names suffer from negative grading discrimination. On the contrary, I found some support for the conjecture that students with ‘foreign’ names are advantaged if their names are disclosed in connection with the evaluation of their examinations. I could however not exclude the possibility that this effect was spuriously generated by a pre-treatment trend, and when individual student fixed effects were conditioned, all effects of blinding the examinations turned statistically insignificant (but the statistical power of this analysis was substantially lower, increasing the risk of a type-II error). The direction of the effect as well as its rather unstable impression deviate from the rule in the literature on effects of blinding examinations, which usually reports blinding to benefit minority groups (Burgess & Greaves, 2013; Hinnerich, et al., 2015; Spietsma, 2013; van Ewijk, 2011; Malouff & Thorsteinsson, 2016).

These results are similar to those in Behaghel, Crépon and Le Barbanchon’s (2015) experiment where blinded job applications lowered foreign applicant job chances, suggesting that blind assessment might go in different directions depending on the context. University lecturers might be a select group, prone to using a relatively lenient evaluation standard for students with foreign-sounding names, and when blinding is introduced in this setting, these students consequently lose out in terms of final grades.

With the caveat that I am far from certain that the estimated effect is causal, its size is nontrivial. In the pre-treatment period, the baseline chance for ‘foreigners’ of receiving a pass with distinction was 16%. The blinding of examinations was on average followed by a six percentage point decrease in the chance of students with ‘foreign’ names receiving a high grade, i.e., the chance was reduced by more than one-third after their names had been hidden from the examiner.

Regarding gender, I did not find any pattern supporting the notion that female students, or male students, were discriminated in the grading of their examinations. The (small) gender gap in grades remained unaffected by the blinding of examinations. Again, this contrasts with the main tendency in the literature on effects of blinding examinations, which usually finds this to benefit male students, whom – as a rule – are assessed with a harsher standard compared to women (Burgess & Greaves, 2013; Cornwell, et al., 2013; Breda & Ly, 2012; Robinson & Lubienski, 2011; Lavy, 2008; Krawczyk, 2018). A speculation is that this literature is often based on tests by secondary level students, where stereotypes of ‘immature boys’ may come into play, but in a setting with adult students, such as the one in the present study, such stereotyping may no longer be practiced.

The mechanisms that are likely to drive the results in this particular context are hard to identify. I find tentative evidence pointing to a weak form of positive discrimination favouring students with ‘foreign’ names, but no evidence of bias when I make the most stringent test. Our results therefore speak against theories of status-based discrimination and statistical discrimination, at least conventionally formulated.

Statistically based discrimination would occur if the examiner, in order to save time, used readily observable characteristics, such as names, for the assessment of student achievements, inferring individual performance from group-level averages. Assuming the grader holds negative prejudices towards students with foreign backgrounds, the results do not support this model. Before the introduction of blind grading, students with ‘foreign’ names on average got lower
grades. If we block the examiners ability to statistically discriminate, we would expect the gap to decrease, since examiners could not then use their prior knowledge of average performances to set individual grades. The gap increases when we block this ability, which suggests that examiners do not practice this kind of discrimination.

The results also speak against status-based discrimination models. Applied to the present case, these would assume that nominal individual characteristics, such as ‘foreignness’, are attributed low status in a context such as higher education, resulting in low expectations of academic achievement. Since the blinding of examinations blocks information on student origins, opportunities to attribute different levels of status and achievement expectations to the students are greatly constrained, and the expectation would therefore be for the grade gap to decrease. Because I find the opposite to be the case, the model seems inadequate to explain the patterns observed.

A plausible alternative explanation of the observed pattern is rather that examiners have a (weak) tendency to practice reverse discrimination. In relative terms, students with ‘native’ names benefited somewhat from the blinding of examinations. It seems as though this group was subject to a somewhat harsher assessment prior to the introduction of blinded examinations, and that examiners applied a relatively lenient standard in their assessments of students with ‘foreign’ names. The results imply that discrimination to some extent is contextual. In a setting such as a university, where impartiality and tolerance is valued, discrimination seems to be absent or even ‘reversed’ compared to what is at hand in the surrounding society.

**Funding**

This study was supported by Swedish Research Council for Health, Working Life and Welfare [grant number 2018-00594] and Swedish Research Council [grant number 2016-02559].

**ORCID**

Magnus Bygren [http://orcid.org/0000-0003-4398-5033](http://orcid.org/0000-0003-4398-5033)

**References**

Batruch, A. 2018. *Reproduction of Social Class Inequalities at School: Experimental Study of Structural Barriers to Educational Equality*. Lausanne: University of Lausanne.

Behaghel, L., B. Crépon, and T. Le Barbanchon. 2015. "Unintended Effects of Anonymous Résumés." *American Economic Journal: Applied Economics* 7(3):1–27. doi:10.1257/app.20140185.

Berger, J., M. Fisek, R. Norman, and M. Zelditch. 1977. *Status Characteristics and Social Interaction: An Expectation States Approach*. New York: Elsevier.

Brandt, M. J. 2011. “Sexism and Gender Inequality across 57 Societies.” *Psychological Science* 22(11):1413–1418. Volym doi:10.1177/0956797611420445.

Breda, T., and S. T. Ly. 2012. Do Professors Really Perpetuate the Gender Gap in Science? Evidence from a Natural Experiment in a French Higher Education Institution., London: Centre for the Economics of Education. London School of Economics and Political Science.

Brennan, D. J. 2008. "University Student Anonymity in the Summative Assessment of Written Work.” *Higher Education Research & Development* 27(1):43–54. doi:10.1080/07294360701658724.

Brylla, E. 2011. Scandinavian Surname Geography: East Scandinavia. In: R. Heuser, D. Nübling & M. Schmuck, eds. *Familienamengeographie: Ergebnisse Und Perspektiven Europäischer Forschung*. Berlin/New York: Walter de Gruyter, pp. 13–20.

Burgess, S., and E. Greaves. 2013. “Test Scores, Subjective Assessment, and Stereotyping of Ethnic Minorities.” *Journal of Labor Economics* 31(3):535–576. Volume doi:10.1086/669340.

Ceobanu, A. M., and X. Escandell. 2010. ”Comparative Analyses of Public Attitudes toward Immigrants and Immigration Using Multinational Survey Data: A Review of Theories and Research.” *Annual Review of Sociology* 36(1):309–328. Volume doi:10.1146/annurev.soc.012809.102651.
Correll, S. J. 2004. “Constraints into Preferences: Gender, Status, and Emerging Career Aspirations.” American Sociological Review 69(1):93–113. doi:10.1177/000312240406900106.

Correll, S. J, Benard, S. 2006. Biased Estimators? Comparing Status and Statistical Theories of Gender Discrimination. In: S. R. Thye & E. J. Lawler, eds. Social Psychology of the Workplace, Vol 23, Advances in Group Processes. New York: Elsevier, pp. 89–116.

Cunningham, W. A., et al. 2004. “Separable Neural Components in the Processing of Black and White Faces.” Psychological Science 15:806–816. doi:10.1111/j.0956-7976.2004.00760.x.

Feagin, J. R., and D. L. Eckberg. 1980. “Discrimination: Motivation, Action, Effects, and Context.” Annual Review of Sociology 6(1):1–20. Volume doi:10.1146/annurev.so.06.080180.000245.

Foschi, M. 1996. “Double Standards in the Evaluation of Men and Women.” Social Psychology Quarterly 59(3): 237–254. doi:10.2307/2787021.

Gorodzeisky, A., and M. Semyonov. 2016. “Not Only Competitive Threat but Also Racial Prejudice: Sources of anti-Immigrant Attitudes in European Societies.” International Journal of Public Opinion Research 28(3):331–354. doi:10.1093/ijpor/edv024.

Hainmueller, J., and M. J. Hiscox. 2007. “Educated Preferences: Explaining Attitudes toward Immigration in Europe.” International Organization 61(2):399–442. doi:10.1017/S0020818307070142.

Hanna, R. N., and L. L. Linden. 2012. “Discrimination in Grading.” American Economic Journal: Economic Policy 4(4): 146–168. doi:10.1257/pol.4.4.146.

Hinnerich, B. T., E. Höglin, and M. Johannesson. 2011. “Are Boys Discriminated in Swedish High Schools?” Economics of Education Review 30(4):682–690. doi:10.1016/j.econedurev.2011.02.007.

Hinnerich, B. T., E. Höglin, and M. Johannesson. 2015. “Discrimination against Students with Foreign Backgrounds: Evidence from Grading in Swedish Public High Schools.” Education Economics 23(6):660–676. Volume doi:10.1080/09645292.2014.899562.

Hinton, D. P., and H. Higson. 2017. “A Large-Scale Examination of the Effectiveness of Anonymous Marking in Reducing Group Performance Differences in Higher Education Assessment.” PLoS ONE 12(8):e0182711. doi:10.1371/journal.pone.0182711.

Inglehart, R., and P. Norris. 2003. Rising Tide: Gender Equality and Cultural Change around the World. Cambridge: Cambridge University Press.

Klein, D. B., and C. Stern. 2005. “Professors and Their Politics: The Policy Views of Social Scientists.” Critical Review: A Journal of Politics and Society 17(3-4):257–303. doi:10.1080/08913810508443640.

Krawczyk, M. 2018. “Do Gender and Physical Attractiveness Affect Grades?” Assessment & Evaluation in Higher Education 43(1):151–161. doi:10.1080/02602938.2017.1307320.

Lavy, V. 2008. “Do Gender Stereotypes Reduce Girls’ or Boys’ Human Capital Outcomes? Evidence from a Natural Experiment.” Journal of Public Economics 92(10-11):2083–2105. doi:10.1016/j.jpubecon.2008.02.009.

Le Grand, C., and R. Szulkin. 2002. “Permanent Disadvantage or Gradual Integration: Explaining the Immigrant–Native Earnings Gap in Sweden.” Labour 16(1):37–64. doi:10.1111/1467-9914.00186.

Malouff, J. M., and E. B. Thorsteinsson. 2016. “Bias in Grading: A Meta-Analysis of Experimental Research Findings.” Australian Journal of Education 60(3):245–256. doi:10.1177/0004944116664618.

Merton, R. K. 1973. [1942]. The Sociology of Science: Theoretical and Empirical Investigations. Chicago: University of Chicago Press.

Pagare, D., and H. Shepherd. 2008. “The Sociology of Discrimination: Racial Discrimination in Employment, Housing, Credit, and Consumer Markets.” Annual Review of Sociology 34(1):181–209. doi:10.1146/annurev.soc.33.040406.131740.

Pitt, E., and N. Winstone. 2018. “The Impact of Anonymity Marking on Students’ Perceptions of Fairness, Feedback and Relationships with Lecturers.” Assessment & Evaluation in Higher Education 43(7):1183–1193. doi:10.1080/02602938.2018.1437594.

Ridgeway, C. L., and L. Smith-Lovin. 1999. “The Gender System and Interaction.” Annual Review of Sociology 25(1): 191–216. Volume doi:10.1146/annurev.soc.25.1.191.

Robinson, J. P., and S. T. Lubinski. 2011. “The Development of Gender Achievement Gaps in Mathematics and Reading during Elementary and Middle School.” American Educational Research Journal 48(2):268–302. Volume doi:10.3102/0021965410382249.

Spreng, M. 2013. “Discrimination in Grading: Experimental Evidence from Primary School Teachers.” Empirical Economics 45(1):523–538. Volume doi:10.1007/s00181-012-0609-x.

Steele, C. 2003. Stereotype Threat and African-American Student Achievement. In: Young, Gifted and Black: Promoting High Achievement among African-American Students. Boston: Beacon Press, pp. 109–177.

Wagner, D. G., and J. Berger. 1997. “Gender and Interpersonal Task Behaviors: Status Expectation Accounts.” Sociological Perspectives 40(1):1–32. doi:10.2307/1389491.

van Ewijk, R. 2011. “Same Work, Lower Grade? Student Ethnicity and Teachers’ Subjective Assessments.” Economics of Education Review 30(5):1045–1058. doi:10.1016/j.econedurev.2011.05.008.