Ethnic Discrimination in High School Grading: Evidence from a Field Experiment

Björn Tyrefors Hinnerich[1,2,*], Erik Höglin[3] & Magnus Johannesson[4]

[1] Department of Economics, Stockholm University, SE-106 91 Stockholm, Sweden

[2] School of Economics and Management, Aarhus University, DK-8000 Aarhus C, Denmark

[3] Swedish Fiscal Policy Council, Box 3273, SE-103 65 Stockholm, Sweden

[4] Department of Economics, Stockholm School of Economics, Box 6501, SE-113 83 Stockholm Sweden

Latest version: June 2011

Abstract

We rigorously test for ethnic discrimination in high school grading in Sweden. A random sample of the national tests in the Swedish language is graded both non-blind by the student's own teacher and blind without any identifying information. The increase in the test score due to non-blind grading is significantly higher for students with Swedish background compared to students with foreign background. This discrimination effect is sizeable, about 10 % of the mean or 20 % of a standard deviation of the blind test score.

Key words: Discrimination, field experiments, education.

Jel Codes: C93, I20, J15.

*Corresponding author (e-mail: bjorn.hinnerich@ne.su.se).

Acknowledgements: We thank George Farkas and Tore Ellingsen for helpful comments and the Institute for Labour Market Policy Evaluation (IFAU), the Jan Wallander and Tom Hedelius Foundation, the Swedish Research Council, and the Swedish Council for Working Life and Social Research for financial support.

I. Introduction

Discrimination according to ethnic origin is a violation of the Universal Declaration of Human Rights adopted by the United Nations. Yet, several studies have suggested that minority ethnic groups are discriminated against on the labor market (see e.g. Darity Jr and Mason 1998; Szymanski 2000; Bertrand and Mullainathan 2004; and the review in Anderson, Fryer, and Holt 2006).[1] To what extent ethnic discrimination occurs also in the education system is less well established (see the review of Farkas 2003).[2] More recent works show that ethnic achievements gaps can be explained by the minority-majority match of the student and the teacher (Dee 2004; Dee 2005; Dee 2007). However, discrimination in the direct or active form (e.g. same student performance but different grades) is only one potential explanation. Another is that the interaction between the majority teacher and minority students make the students perform worse, for example by adjusting their effort in line with the stereotype of the teacher.[3] As Dee (2005) states: "the exact design of … policies also require a clear understanding of the underlying structural mechanisms that make these student–teacher interactions relevant in the first place". The major contribution of this paper is to unbundle

[1] See also the paper by Fershtman and Gneezy (2001) who detects ethnic discrimination using standard experimental games and the paper by List (2004) who finds evidence of ethnic discrimination in the sportscard market. For a discussion of the methodological problems in identifying discrimination on the labor market, see Heckman (1998).

[2] Ethnicity is by nature multidimensional (e.g. shared heritage including dimensions such as ancestry, history, religion, language, country, nationality, set of attitudes, behaviors, physical appearance etc.). In this paper we define ethnicity based only on country of origin.

[3] See Ouazad and Page (2011) for some new experimental evidence.

some effects by investigating whether the direct form of ethnic discrimination exist in high school grading by comparing non-blind and blind grading of the same test.[4] We find that students with foreign background are discriminated in its active form.

The methodology of varying the degree of "blindness" to detect discrimination has previously been used in economics by for instance Blank (1991), Goldin and Rouse (2000), and Lavy (2008). Of particular interest here is the study by Lavy (2008), who used a large data set from high school in Israel and compared two different test scores for the same individuals: one school score based on a non-blind grading of a school exam by the student's own teacher and one test score on a similar test graded blindly by an external examiner. He found a statistically significant discrimination of men in all the examined tests. A limitation of the Lavy study is that it does not involve a comparison of blind and non-blind grading of the exact same tests.[5,6] Hence, our way of grading the same test twice is an improvement making our method closer to the ideal test setting.

Our study is carried out in the Swedish high school, where national tests are given in Swedish, English and math. These tests are graded non-blindly by the student's own teacher based on written guidelines stating the prerequisites for each grade. To test for discrimination we draw a random sample of the written tests in Swedish from the academic year 2005/06. The students in our sample are classified into Swedish background (n=1423) or foreign

[4] Using the terminology of Harrison and List (2004) our study is a natural field experiment.

[5] 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).

[6] In a related paper using the same dataset as in the present paper we fail to find any significant gender discrimination in grading, with a point estimate close to zero (Hinnerich, Höglin, and Johannesson 2011).

background (n=199) according to the classification used by Statistics Sweden (Statistics Sweden 2002). After rewriting the tests on a word processor and removing the student identities, but nothing else, the tests were re-graded blindly by a group of teachers hired from a teachers' agency. The re-grading teachers were not given any information regarding the purpose of the study.

The difference between the non-blind grade and the blind grade is a measure of the bias induced by non-blind grading (Blank 1991; Goldin and Rouse 2000). In the absence of ethnic discrimination this bias should be the same for different ethnic groups. However, we find a sizeable and significant difference between students with Swedish background and students with foreign background. Non-blind grading increase the test score by 16% for students with Swedish background, but for students with foreign background this effect is only 4%.

In a further analysis we subdivide students with foreign background into students with European background (n=84) and students with non-European background (n=115). This analysis reveals that the discrimination effect is consistently stronger for students with non-European background.

Our paper is closely related to recent work of Hanna and Linden (2009). They test for discrimination with respect to age, gender and caste in India. They invite children to participate in an exam competition with a cash prize and recruit local teachers to grade the exam. Before the grading, child characteristics (age, gender, and caste) are randomly assigned to the exams to identify any discrimination effect. They find that the teachers assign lower scores to exams assigned to the lower caste, but they find no evidence of discrimination with respect to age or gender. Our results are consistent with their findings in the sense that both caste and foreign background may be indicators of social status. Lately, Ewijk (2010) let 113 Dutch teachers grade the same 10 essays written by 11 years old students, but varying the names indicating different ethnic group membership of the author. Although finding a small discriminatory effect on grading of the ethnical minority group, it is not statistically significant. Except for the different methodology used in our case, our study also differs in terms of the students' age, where our study uses students of age around 18 years, which may influence the results.[7]

There is also a literature on teacher bias in sociology testing the effects on grades of variables like ethnicity and gender after controlling for some measure of performance or cognitive ability. The evidence on ethnic discrimination from this literature is mixed (Farkas, Sheehan, and Grobe 1990; Farkas et al. 1990; Leiter and Brown 1985; Farkas 1996; Rosenbaum 2001). However, it is difficult to fully control for performance to rigorously test for discrimination with this approach. A related line of work focuses specifically on whether teachers' perceptions are biased by racial stereotypes, which could be a mechanism leading to ethnic discrimination. Ferguson (2003) reviews this literature and tentatively concludes that "... teachers' perceptions, expectations, and behaviors probably do help to sustain, and perhaps even to expand, the Black-White test score gap." (pp. 495).

It is common in economics to make a distinction between statistical discrimination (Arrow 1972a,b; Phelps 1972) and taste based discrimination (Becker 1957). In statistical discrimination individuals form beliefs about important unobserved characteristics based on observable characteristics like ethnic background. In a strict sense it should not be possible to observe statistical discrimination in our setting as the outcome to be evaluated (the test) is fully observable to the teacher. However, one cannot rule out that the grading is still affected

[7] Evidence from Swedish Schools Inspectorate (2011) using a random sample of blind and non-blind graded essays in Swedish for students of the age 10 shows very little difference in the blind and non-blind grade. However, for older students, especially at high school, the difference is remarkable.

by beliefs about ability or performance correlated with ethnic background.[8] Teachers who are unsure how to grade the test may for instance be affected by the perceived ability of the student. Taste based discrimination involves having a preference for one group versus another group, and this is a potential mechanism for our observed effect. To discriminate based on taste would not be costly to the teachers in our setting, which is an important difference compared to the labor market or other markets. It is also possible that our observed discrimination effect could operate through other more indirect channels if ethnic background is correlated with other factors biasing the non-blind grading. However, the discrimination effect is robustly significant after controlling for a host of factors such as gender, year of birth, and school fixed effects.

The remainder of our paper is structured as follows. Below we first outline the design of our study and the methodology and data used to test for discrimination. Thereafter we present our findings, and then we present extensions and further robustness followed by some concluding remarks.

II. Design of the Study

II. A. 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

[8] This is consistent with the literature on teachers’ perceptions and racial stereotypes reviewed by Ferguson (2003).

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.[9] 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 should be set according to absolute knowledge criteria and the core subjects have nationally stipulated prerequisites for each grade. 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.[10]

II. B. The Test

[9] 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.

[10] Recently, the Swedish Schools Inspectorate has begun to partly evaluate schools grading but except for monitoring, the actions that could be taken against schools are absent. See, for example, Swedish Schools Inspectorate (2011)

Compulsory national tests in the Swedish high school system are given in the core subjects Swedish, English and math. We focused on the test in Swedish, since we posited that grading a Swedish test allows for more arbitrariness than, for example, math. The tests have three parts, one oral and two written. We used data from the second, more extensive, written test for the academic year 2005/06. In this test, students were asked to write an essay based on one out of nine topics within a common theme.[11] Students choose their topic with full discretion.

The written part of the national test is graded on the same scale as the subjects: Fail (0), Pass (10), Pass with Distinction (15), and Excellent (20). The national tests are graded by the student's own teacher. The teachers are given written guidelines stating the prerequisites for each grade, but they have great discretion in the actual grading. No measures are taken by the national authorities to ensure that the guidelines are followed, and no evaluations of the schools are conducted.

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 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 variation proving the fact that the test is only one of the determinants for the final grade in the subject (Lindahl 2007). 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. Every academic year, two national tests in Swedish are constructed by the National Agency of

[11] 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.

Education in conjunction with the Department of Scandinavian Languages at Uppsala University.

II. C. The Sample

We drew a random sample of 2880 students eligible to take the test (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 enrolment data for the schools from the National Agency of Education. Based on this data, we used a two-step procedure to ensure that each eligible student was 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 randomly selected 100 schools, where the probability of each school being chosen corresponded 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 any citizen, 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 eligible students in the population had the same probability of being sampled.

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 (a response rate of 59%). The main reason for the non-response is absenteeism of the test (i.e. these students did not actually take the test), but some tests were also missing due to inferior administrative routines at the schools (the National Agency of Education requires that all tests and test results should be properly filed). Two out of the 96 participating schools did not have proper filing procedures in line with the guidelines of the National Agency of Education and did not deliver the required material.

As compared to the annual collection of test scores by the National Agency of Education for 200 representative high schools, we had approximately the same response rate. For Swedish B, their total response rate for 2006 was about 62%, as compared to 59% in our study, and about 10% of their missing values were due to administrative causes (Swedish National Agency for Education 2006). The rest was due to the fact that eligible students were absent from the test.

II. D. The Blind Grading

To estimate the effect of ethnic discrimination we estimated the difference between the non-blind and the blind grade. The non-blind grade is the grading of the test of the student's own teacher. To obtain the blind grade we first rewrote all tests on a word processor and the student identities as well as their teachers' notes were deleted (the rewriters were not given any information regarding the purpose of the study). Nothing else was changed (i.e. any spelling or grammatical mistakes were carried over to the typed text). The tests were then re-graded blindly by 42 teachers hired from a teachers' agency (each teacher re-graded 35-50 tests). The re-grading teachers had no information regarding the purpose of the study. We required re-graders to have been grading national tests in Swedish before. The teachers were provided the official written guidelines stating the prerequisites for each grade and topic.

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

II. E. Ethnic Background

For each student we collected data on the country of birth of the student and both parents. This data was obtained from the Swedish Tax Agency. To classify students into students with Swedish background and students with foreign background we used the classification used by Statistics Sweden (Statistics Sweden 2002). If the student was born in Sweden and at least one of the parents was born in Sweden, the student was classified into Swedish background. If the student was born abroad or if both parents were born abroad, the student was classified into foreign background. We obtained data about ethnic background for 1622 out of the 1713 students in our sample; for the remaining observations the country of birth of the student or the parents was not known by the Swedish Tax Agency or this information was classified. Our analyses were performed for the sample of 1622 students of which 1423 were of Swedish background and 199 were of foreign background. We are aware of the weakness that our study share with other studies carried out on the labor market of not observing other factors than country background determining ethnicity such as ancestry, history, religion, language, set of attitudes, behaviors, physical appearance etc. In fact this is another unbundling problem in the literature. For example, a large immigrant group in Sweden is originating from former Yugoslavia, and within this group we cannot distinguish whether a student is Christian or Muslim or any other religion.

II. F. Identification and Hypothesis

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

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

where *Foreign* is an indicator taking the value of 1 if student *i* has a foreign background and 0 otherwise. We define discrimination as differences in the test results across groups, conditional on ability. To put it differently: If grades are not discriminatory, then students of different background producing the same quality of the test should get the same grade.[12] If not, one group is discriminated. Thus, we interpret β as a discrimination effect. If negative, then students with foreign background are discriminated and if positive, students with Swedish background are discriminated. The classical problem with this formulation is that we do not observe ability. If ability is correlated with foreign background, then estimating this equation without conditioning on ability would bias β and we could falsely conclude that

[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 foreign background. But even if this is the case, it would still result in discrimination of students with a foreign background. It is very difficult to separate such 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).

students with foreign background are discriminated, when in fact students with a Swedish background are more able, or vice versa.

Given our set up of the study, this endogeneity problem can be taken care of. Consider an examiner that has no information about the student background (*B* for ‘blind’). Then, we simply have β = 0 and

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

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

$$\Delta Testscore_i = \alpha + \beta Foreign_i + u_i\,, \qquad (3)$$

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

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.[13] 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

[13] This can be investigated by adding controls to equation (3), which we do in Section III.

create a problem when estimating the discrimination effect, we need their potential difference in test scores to be related to ethnical background. It is not a problem for our identification strategy that this group would perform differently from the students taking the test.

Our design identify the causal effect of non-blind grading on test scores (the bias induced by non-blind grading). This is the experimental manipulation. This implies that we can correctly estimate how this bias varies between students with Swedish background and students with foreign background (or any other sub-groups identified in the data).

We test the hypothesis that students with foreign background are discriminated, by testing whether the increase in the grade due to non-blind grading is larger for students with Swedish background than for students with foreign background using the cardinal scale of the grades. We use both parametric and non-parametric tests and all reported t-values are two-sided. Moreover, we present regression results in order to further investigate robustness and the mechanism explaining our result. In the regression section we follow the convention in the test score literature by using scores standardized to a distribution with zero mean and a unit standard deviation, meaning that the discrimination effect should be interpreted as the share of a standard deviation of the blind test score.

III. Results

III. A. Descriptive Results and Tests

The distribution of the non-blind and blind test scores for students with Swedish background and students with foreign background is shown in Figure 1 and the mean test scores are shown in Table 1.[14] As is evident from the Figure, students with Swedish background have higher non-blind test scores. On average the non-blind test score is 8% higher for students with Swedish background compared to students with foreign background and this difference is significant at the 5% level. The average rank in the non-blind grade is 801 for students with Swedish background and 889 for students with foreign background (in a total sample of 1622). However, when we look at the blind test scores the entire difference between the two student groups disappears; the blind test score is even somewhat higher for students with foreign background.[15] The average rank in the blind grade is 813 for students with Swedish background and 798 for students with foreign background, i.e. an improvement in the average rank with about 6 percentiles for students with foreign background.

Non-blind grading thus favors students with Swedish background. As illustrated in Figure 2, non-blind grading increase the test score by 16% for students with Swedish background compared to only 4% for students with foreign background.[16] This difference is significant at the 1% level with both an independent samples t-test and a non-parametric

[14] In this table, 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.

[15] The correlation between the non-blind and the blind grade is 0.41 (for both the Pearson and the Spearman correlation). The scores are identical for about 46% of the students.

[16] That the non-blind grading leads to higher test scores in both ethnic groups is consistent with grade inflation and may depend on teacher incentives and competition between high schools (Jacob and Levitt 2003; Wikström and Wikström 2005).

Mann-Whitney test.[17] The mean size of this discrimination effect is 1.24 on the 0-20 grading scale; this corresponds to 11.8 % of the average blind test score of 10.49 in the sample.

[Figure 1 and 2 and Table 1 about here]

The students with foreign background are from a vast array of countries. Figure 3 shows a histogram for the different countries.

III. B. Main Regression Results

To examine our results further we run OLS regressions with the difference between the non-blind and the blind grade as the dependent variable. We follow the convention in the literature and use test scores standardized to a distribution with zero mean and a unit standard deviation. Hence, the discrimination effect should be interpreted as the share of a standard deviation of the blind test score. These results are shown in Table 2. In the first model we only include a dummy variable for foreign background to measure the discrimination effect. To account for a possible correlation in observations for different schools and different regraders we estimate standard errors based on two-way clustered standard errors at the school and regrader level (Cameron, Gelbach, and Miller 2006; Thompson 2010). Students with a foreign background are discriminated against and the point estimate shows a discrimination effect of about 0.24 of a standard deviation of the blind score distribution. The

[17] We can reject the hypothesis that the difference in the non-blind and blind test score is normally distributed (Shapiro-Wilk W test, W=0.995, n=1622, P<0.001), and we therefore report results for both parametric and non-parametric tests.

estimate is significant at the 1 % significance level. In comparison to the study by Lavy (2008) on gender discrimination in grading, our estimate is at least twice as large.

To check for robustness of our result we add controls sequentially. In columns 2-3, we add a dummy for gender (being one if the student is a boy) and student year of birth. Any of those might potentially be correlated with foreign background. For example, if immigrants are older when they are at high school and discrimination is at work against older students in class, then our discrimination effects should be sensitive of including age as control. However, table 2 column 2-3 shows that, gender and student year of birth are not significant and our discrimination effect is left unchanged.[18]

Our experimental manipulation is to rewrite and regrade the test. Thus, it is natural to include fixed effects for the rewriter and the regraders. As shown in column 4-5, our results are robust to this inclusion. Moreover, if students with foreign background to a larger extent choose academic/vocational training, this might influence the results if there is a relatively more liberal way of grading at one of the programs. Adding a fixed effect for academic or vocational program in column 6, however, does not change the results.

In general, a major concern with any non-blind/blind set up is that the blind assessor also can at least partly observe the variable that is supposed to be non-observable (the foreign background is absent in equation (2)). It is possible that the re-grading teachers may be able to infer the background of the student based on the text of the test. This could lead to a bias towards zero in our estimated discrimination effect, making our estimate too small in absolute terms. As the students choose among different topics, the choice of topic may reveal some

[18] However, there is very little variation in the year of birth (84% were born in 1987) limiting the statistical power to detect any discrimination effect of age.

information about ethnicity in line with the reasoning of Ewijk (2010). [19] Thus, including a control for the topic of the student should increase the discrimination effect if the topic has informational value of the background of the student. In table 2, column 7, we include topic specific effects, but the estimate is left rather unchanged.

Lastly, since the 1990's the students in Sweden has the right to choose schools and the schools are getting paid for each student via a governmental voucher system. Thus, schools compete for students. It is likely that grading can depend on the incentives for teachers and the competition between schools (Jacob & Levitt, 2003) 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. This is also one of the major candidates explaining why external graders do set lower grades in general. Moreover, the right to choose schools and competition has led to an increasing socioeconomic and ethnic sorting (Böhlmark ann Holmlund, 2011). For example in 2007, 70 percents of the students with immigrant background went to schools that had more than the mean share of immigrants. Moreover, choice of residence also influence which school a person goes to and Sweden is segregated in terms of residence. Lindbom and Almgren (2007) attribute a large share of the segregation in schools to segregation of residence. If segregation and grade inflation are related then adding school fixed effects should decrease our estimate. The last column presents the results when including all controls and a fix effect for the schools. Indeed, including school fixed effects results in a drop in the discrimination effect of about 4 percentage points. This suggests that part of the discrimination effect may be an indirect effect of that the grading levels differ between schools with different fractions of students with

[19] The student could choose among nine topics. Since we have data from two test occasions (fall 2005 and spring 2006) we observe 18 topics. Moreover, as some students failed to indicate the chosen topic, we added a category for unknown topics.

foreign background (i.e. that schools with a higher fraction of students with Swedish background may be more prone to grade inflation). However, the point estimate of the discrimination effect with school fixed effects is not significantly different from the point estimate without school fixed effects. Moreover, it still indicates discrimination as it is significant on the 5 percent level.

[Table 2 about here]

III. C. Extensions

There are some concerns regarding bias of our discrimination effect. One is the censoring due to the discreteness of the data. If students with foreign background have lower ability, then they will earn the lowest original grade more often than the Swedish students. If external graders are more conservative, then they may thus be able to down-grade Swedish students to a larger extent. If this happens, it will falsely make us conclude that there is discrimination against students with foreign background. The same argument holds for increases if the external graders are more liberal. One way to investigate this is to see whether our results change when excluding the students with either the lowest or the highest grade in both the non-blind and the blind grading. In column 1 in table 3 we exclude the former sub group and in column 2 the latter is excluded. Table 3 always includes the full set of controls from table 2, so the estimates should be compared to column 8 in table 2. The results in column 1 and 2 in table 3 are almost identical to those in column 8 in table 2, indicating robustness to this data exclusion.

Due to the nature of the field experiment, we cannot investigate the mechanisms driving the results on a very thorough level. However, adopted children with two parents born in

Sweden are more likely to share the “foreign” physical attribute with other students with foreign background but not other attributes such as religion, language, set of attitudes, behaviors, etc. We define adopted children (n=38) as those born outside Sweden, but whose parents are both born in Sweden. We separate this group from students with foreign background, including them as its own category in the regression. The point estimates in Column 3 of Table 3 shows little evidence of discrimination of this group. This supports the hypothesis that the discrimination is more complex than just defined by physical appearance.

Moreover, our control group, students with Swedish background, includes students born in Sweden, but were only one of the parents are born abroad (n=115). The regression presented in column 4 in table 3 includes for a dummy for this group. Although, imprecisely measured, the point estimate is in line with a sizeable discrimination effect in this group as well.

[Table 3 about here]

Bias may arise if the re-graders set their test scores randomly with respect to ability. To see this, suppose that Swedish students are more able than students with foreign background. Further, assume that the original teachers grade tests without bias, while the re-graders assign grades randomly to tests. This would make our discrimination effects biased upwards (in absolute terms), making us falsely detect discrimination against students with foreign background. Several facts suggest that this is not a major problem. First, the correlation between the original grading and the re-grading is high and significant at the 1 % - level. The correlation between the non-blind and the blind grade is around 0.41 (for both the Pearson and the Spearman correlation). Second, the scores are identical for about 46% of the students, making it unlikely that re-graders grade randomly. Finally, the fact that we failed to find that boys were discriminated in a related paper using the same dataset (Hinnerich, Höglin, and Johannesson 2011) also suggests that random regrading is unlikely. If re-graders graded randomly we would have found such discrimination in that paper as boys have significantly lower non-blind grades than girls.

Other studies in the labor market of Sweden provide some evidence that some groups (Arabs and Africans) face more discrimination than if the person is from, for example, a EU15 country (e.g. Arai and Skogman Thoursie 2009). In our study the treatment group consists of only 199 students with foreign background. This rules out dividing the treatment group country-by-country. Moreover, we only observe the country where the student, the mother, and the father were born, and not other variables determining ethnicity. For example, a student born in Yugoslavia or Bosnia-Herzegovina might be Muslim or Christian, and similar reasoning applies to some Asian and African born students. We avoid being speculative and arbitrary by tying our hands to the definitions of Statistics Sweden, separating the students with foreign background into European (n= 87) and non-European (n=112).[20,21]

[20] According to Statistics Sweden the following countries in our sample are classified as European: Finland, Norway, Denmark, Iceland, Austria, France, Germany, Greece, Italy, Netherlands, UK, Spain, Switzerland, Croatia, Czech Republic, Hungary, Poland, Romania, Bosnia-Herzegovina, Macedonia, Yugoslavia, Russia and Ukraine. The following are classified as being non-European: Algeria, Turkey, Egypt, Morocco, Tunisia, Afghanistan, Argentina, Bangladesh, Bolivia, Brazil, Chile, China, Korea, Colombia, Cuba, El Salvador, Ethiopia, Gambia, India, Indonesia, Iran, Iraq, Israel, Kenya, Kuwait, Laos, Lebanon, Liberia, Malaysia, Pakistan, Palestine, Peru, Somalia, Sri Lanka, Syria, Tanzania, Taiwan, Thailand, Uganda, Venezuela, Vietnam and USA.

We would expect the non-Europeans to be more discriminated than the Europeans – in line with the results from the labor market – although the *difference* of the point estimates between the two groups with foreign background would be biased towards zero according to potential misclassification of ethnicity discussed above. Moreover, given the reduction in sample sizes of the groups the estimates will be less precisely estimated.

Table 4 shows the OLS regression results with the same procedure of adding controls as in table 2. The point estimate of the discrimination effect is consistently higher for students with non-European background in all specifications. For students with non-European background the point estimate varies between 0.219 - 0.302 standard deviations and all estimates are significant at the 5 % level. As expected, students with European background are less discriminated and the point estimates are insignificant. Thus, there is evidence that non-Europeans are more discriminated than Europeans not only in the labor market but also in the Swedish schools. However, we cannot reject the null hypothesis of no difference between students with European and non-European background, which is not very surprising given the small sample sizes of these two groups.

[21] A student with foreign background could have two different classifications, being both non-European and European if for example the mother is born outside Europe, but the father is born in Europe (but not Sweden). We define the groups by first using information on if the student with foreign background was born in Sweden, Europe, or outside of Europe. If she was not born in Sweden, she is classified according to her country of birth. If she is born in Sweden, we base our presented results on the country of birth of the mother. However, the results are similar using the country of birth of the father (it only changes the classification of two students from the European to the non-European group).

[Table 4 about here]

IV. Conclusions

We find a sizeable and robust discrimination effect of students with foreign background. The estimate is based on a representative sample of Swedish high school students for one of the core subjects (Swedish) in the Swedish High School system. National tests are carried out also in math and English and further work is needed to test if discrimination occurs also for these tests. It is also important to investigate the mechanism of the discrimination further. From a policy perspective it would be relatively easy to eliminate the discrimination from the national test by introducing a blind grading system. However, as the final subject grades are not based solely on the national tests there would still be scope for discrimination in the final subject grades. To fully ensure against discrimination would necessitate a system where the final grades are based solely on blindly graded tests.

References

Andersson, Lisa, Roland Fryer, and Charles Holt. 2006. “Discrimination: Experimental Evidence from Psychology and Economics.” In *Handbook on the Economics of Discrimination*, ed. William M. Rodgers III, 97-118. Northampton, MA: Edward Elgar.

Mahmood Arai and Peter Skogman Thoursie. 2009. "Renouncing Personal Names: An Empirical Examination of Surname Change and Earnings", *Journal of Labor Economics*, vol. 27, 2009.

Arrow, Kenneth J. 1972a. “Models of Job Discrimination.” In *Racial Discrimination in Economic Life*, ed. Anthony H. Pascal, 83-102. Lexington, MA: D.C. Heath.

Arrow, Kenneth J. 1972b. “Some Mathematical Models of Race Discrimination in the Labor Market.” In *Racial Discrimination in Economic Life*, ed. Anthony H. Pascal, 187-204. Lexington, MA: D.C. Heath.

Becker, Gary. 1957. *The Economics of Discrimination*. Chicago: University of Chicago Press.

Bertrand, Marianne, and Sendhil Mullainathan. 2004. “Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination.” *American Economic Review*, 94, 991-1013.

Blank, Rebecca M. 1991. “The Effects of Double-Blind Versus Single-Blind Reviewing: Experimental Evidence from the American Economic Review.” *American Economic Review*, 81, 1041-1067.

Böhlmark, Anders, and Helena Holmlund. 2011. *20 år med förändringar i skolan. Vad har hänt med likvärdigheten*, Stockholm: SNS.

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller. 2006. “Robust Inference with Multi-Way Clustering.” National Bureau of Economic Research Working Paper T0327.

Darity Jr, William A., and Patrick L. Mason. 1998. “Evidence on Discrimination in Employment: Codes of Color, Codes of Gender.” *Journal of Economic Perspectives*, 12, 63-90.

Dee, Thomas S. 2004. Teachers, Race, and Student Achievement in a Randomized Experiment, Review of Economics and Statistics, 86(1), 195–210.

Dee, Thomas S. 2005. A Teacher Like Me: Does Race, Ethnicity, or Gender Matter?, American Economic Review, 95(2),158–65.

Dee, Thomas S. 2007. Teachers and the Gender Gaps in Student Achievement. Journal of Human *Resources, 42*(3): 528-554.

Farkas, George. 1996. *Human Capital or Cultural Capital? Ethnicity and Poverty Groups in an Urban School District*. New York: Aldine de Gruyter.

Farkas, George. 2003. “Racial Disparities and Discrimination in Education: What Do We Know, How Do We Know it, and What Do We Need to Know?” *Teachers College Record*, 105, 1119-1146.

Farkas, George, Daniel Sheehan, and Robert P. Grobe. 1990. “Coursework Mastery and School Success: Gender, Ethnicity, and Poverty Groups Within an Urban School District.” *American Educational Research Journal*, 27, 807-827.

Farkas, George, Daniel Sheehan, Robert P. Grobe, and Yuan Shuan. 1990. “Cultural Resources and School Success: Gender, Ethnicity, and Poverty Groups Within an Urban School District.” *American Sociological Review*, 55, 127-142.

Ferguson, Ronald F. 2003. “Teachers’ Perceptions and Expectations and the Black-White Test Score Gap.” *Urban Education*, 38, 460-507.

Fershtman, Chaim, and Uri Gneezy. 2001. “Discrimination in a Segmented Society: An Experimental Approach.” *Quarterly Journal of Economics*, 116, 351-377.

Goldin, Claudia, and Cecilia Rouse. 2000. “Orchestrating Impartiality: The Impact of Blind Auditions on Female Musicians.” *American Economic Review*, 90, 715-741.

Hanna, Rema, and Leigh Linden. 2009. “Measuring Discrimination in Education.” National Bureau of Economic Research Working Paper 15057.

Harrison, Glenn W., and John A. List. 2004. “Field Experiments.” *Journal of Economic Literature*, 42, 1009-1055.

Heckman, James J. 1998. “Detecting Discrimination.” *Journal of Economic Perspectives*, 12(Spring), 101-116.

Hinnerich, Björn T., Erik Höglin, and Magnus Johannesson. 2011. “Are Boys Discriminated in Swedish High Schools?” Economics of Education Review, 30, 682-690.

Jacob, Brian A., and Steven D. Levitt. 2003. “Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating.” *Quarterly Journal of Economics*, 118, 843-877.

Lavy, Victor. 2008. “Do Gender Stereotypes Reduce Girls’ Human Capital Outcomes? Evidence From a Natural Experiment.” *Journal of Public Economics*, 92, 2083-2105.

Leiter, Jeffrey, and James S. Brown. 1985. “Determinants of Elementary School Grading.” *Sociology of Education*, 58, 166-180.

Lindbom, A & Almgren, E (2007) Valfrihetens efftekter på skolornas elevsammansättning: Skolsegregationen i Sverige. In: Anders Lindbom (Ed) *Friskolorna och framtiden – segregation, kostnader och effektivitet.* Stockholm: Institutet för framtidsstudier.

List, John A. 2004. “The Nature and Extent of Discrimination in the Marketplace: Evidence from the Field.” *Quarterly Journal of Economics*, 119, 49-89.

Ouazad, Amine and Page, Lionel. 2011 “Estimating Perceptions of Discrimination: Experimental Economics in Schools” INSEAD Working Paper No. 2011/34/EPS.

Phelps, Edmund S. 1972. “The Statistical Theory of Racism and Sexism.” *American Economic Review*, 62, 659-661.

Rosenbaum, James. 2001. *Beyond College for All: Career Paths for the Forgotten Half*. New York: Russel Sage Foundation.

Statistics Sweden. 2002. *Statistics on Persons with Foreign Background: Guidelines and Recommendations*. Reports on Statistical Co-ordination for the Official Statistics of Sweden 2002:3. Örebro: Statistics Sweden.

Swedish National Agency for Education. 2006. *Gymnasieskolans Kursprov vt 2006: En Resultatredovisning* (in Swedish). Stockholm: Swedish National Agency for Education.

Swedish Schools Inspectorate. 2011. *Lika eller olika? Omrättning av nationella prov i grundskolan och gymnasieskolan*(in Swedish). Stockholm: Swedish Schools Inspectorate.

Szymanski, Stefan. 2000. “A Market Test for Discrimination in the English Professional Soccer Leagues.” *Journal of Political Economy*, 108, 590-603.

Thompson, Samuel B. 2010. “Simple Formulas for Standard Errors that Cluster by Both Firm and Time.” *Journal of Financial Economics*, in press.

van Ewijk, Reyn. 2010. "Same Work, Lower Grade? Student Ethnicity and Teachers' Subjective Assessments." Tinbergen Institute Discussion Paper 127(3).

Wikström, Christina, and Magnus Wikström. 2005. “Grade Inflation and School Competition: An Empirical Analysis Based on the Swedish Upper Secondary Schools.” *Economics of Education Review*, 24, 309-322.

Tables

Table 1. Test scores and differences in test scores.

Sample statistics	N	Mean	Std. Dev
Non-blind test score:			
Students with Swedish background	1423	12.122	4.975
Students with foreign background	199	11.181	5.099
Difference		.941	
p-value of diff. (t-test)		0.015	
p-value of diff. (Mann-Whitney test)		0.008	
Blind test score:			
Students with Swedish background	1423	10.457	5.509
Students with foreign background	199	10.754	5.383
Difference		-.297	
p-value of diff. (paired t-test)		0.478	
p-value of diff. (Mann-Whitney test)		0.651	
Non-blind test score – Blind test score:			
Students with Swedish background	1423	1.665	5.743
Students with foreign background	199	.427	5.500
Difference		1.238	
p-value of diff. (t-test)		0.003	
p-value of diff. (Mann-Whitney test)		0.003	

Table 2. OLS Regression results on the difference between the non-blind and the blind test scores. The "Foreign background" variable measures the effect of discrimination.

Variables	1	2	3	4	5	6	7	8
Foreign background	-0.242***	-0.245***	-0.230**	-0.222***	-0.233***	-.224***	-0.211***	-0.168**
	(0.069)	(0.069)	(0.066)	(0.062)	(0.060)	(0.060)	(0.065)	(0.079)
Male	No	-0.034	-0.030	0.002	0.002	0.002	0.002	0.002
		(0.051)	(0.051)	(0.047)	(0.046)	(0.046)	(0.046)	(0.046)
Student birth year	No	No	0.146	0.163	0.163	0.175	-0.015	-0.022
			(0.071)	(0.141)	(0.120)	(0.108)	(0.047)	(0.056)
Re-grader fix effect	No	No	No	Yes	Yes	Yes	Yes	Yes
Re-writer fix effect	No	No	No	No	Yes	Yes	Yes	Yes
Program fix effect	No	No	No	No	No	Yes	Yes	Yes
Topic fix effect	No	No	No	No	No	No	Yes	Yes
School fix effect	No	No	No	No	No	No	No	Yes
R^2	0.005	0.006	0.009	0.102	0.115	0.116	0.133	0.240

Notes: Dependent variables are standardized scores. 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: Thompson 2010). ****,**,* denotes significance at the 1, 5, and 10% levels. N = 1622 in all specifications.

Table 3. OLS Regression results on the difference between the non-blind and the blind test scores. The "Foreign background" variable measures the effect of discrimination. Extensions.

Variables	1	2	3	4
	Limited sample for censoring check		Full sample. Subdivision of treatment & control group	
Foreign background	-0.174**	-0.173**	-0.185**	-0.198**
	(0.082)	(0.084)	(0.089)	(0.087)
Adopted	No	No	-0.012	-0.028
			(0.280)	(0.281)
Swede with one foreign parent	No	No	No	-0.137
				(0.128)
N	1,552	1,576	1,622	1,622
R^2	0.250	0.244	0.240	0.241

Notes: Dependent variables are standardized scores. A constant and the full set of controls in table 2 are always included. Two-way clustered standard errors reported in parentheses at the school and re-grader level (Cameron, Gelbach and Miller 2006: Thompson 2010). ****,**,* denotes significance at the 1, 5, and 10% levels.

Table 4. OLS Regression results; dividing foreign background into European background and non-European background.

Variables	1	2	3	4	5	6	7	8
Non-European	-0.287***	-0.290***	-0.266***	-0.286***	-0.302***	-0.292***	-0.273***	-0.219*
background(β1)	(0.097)	(0.098)	(0.096)	(0.094)	(0.090)	(0.091)	(0.095)	(0.119)
European	-0.181	-0.182	-0.183	-0.136	-0.139	-0.132	-0.129	-0.104
background (β2)	(0.116)	(0.117)	(0.113)	(0.115)	(0.118)	(0.118)	(0.126)	(0.123)
Male	No	-.035	-0.031	0.001	0.001	0.001	-0.016	-0.024
		(.051)	(0.051)	(0.047)	(0.046)	(0.046)	(0.047)	(0.055)
Student birth year	No	No	0.144	0.159	0.160	0.171	0.160	0.126
			(0.071)	(0.140)	(0.127)	(0.105)	(0.114)	(0.163)
Re-grader fix effect	No	No	No	Yes	Yes	Yes	Yes	Yes
Re-writer fix effect	No	No	No	No	Yes	Yes	Yes	Yes
Program fix effect	No	No	No	No	No	Yes	Yes	Yes
Topic fix effect	No	No	No	No	No	No	Yes	Yes
School fix effect	No	No	No	No	No	No	No	Yes
Test of if β1=β2	0.518	0.510	0.608	0.362	0.322	0.328	0.400	0.509
(p-value)								
R^2	0.006	0.006	0.010	0.103	0.116	0.117	0.134	0.240

Notes: Dependent variables are standardized scores. 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: Thompson 2010). ****,**,* denotes significance at the 1, 5, and 10% levels. N = 1622 in all specifications.

Figures

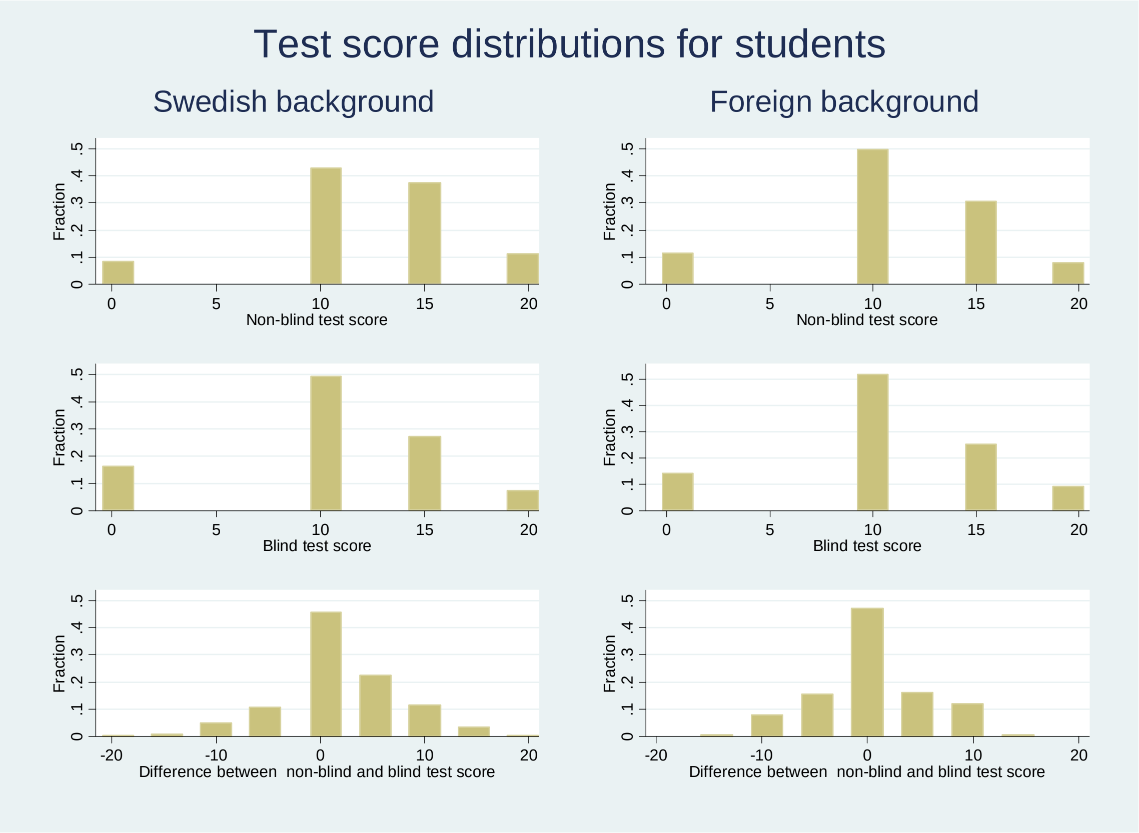

Figure 1. The distribution of the test scores for the non-blind and the blind grading procedures for students with Swedish background and students with foreign background.

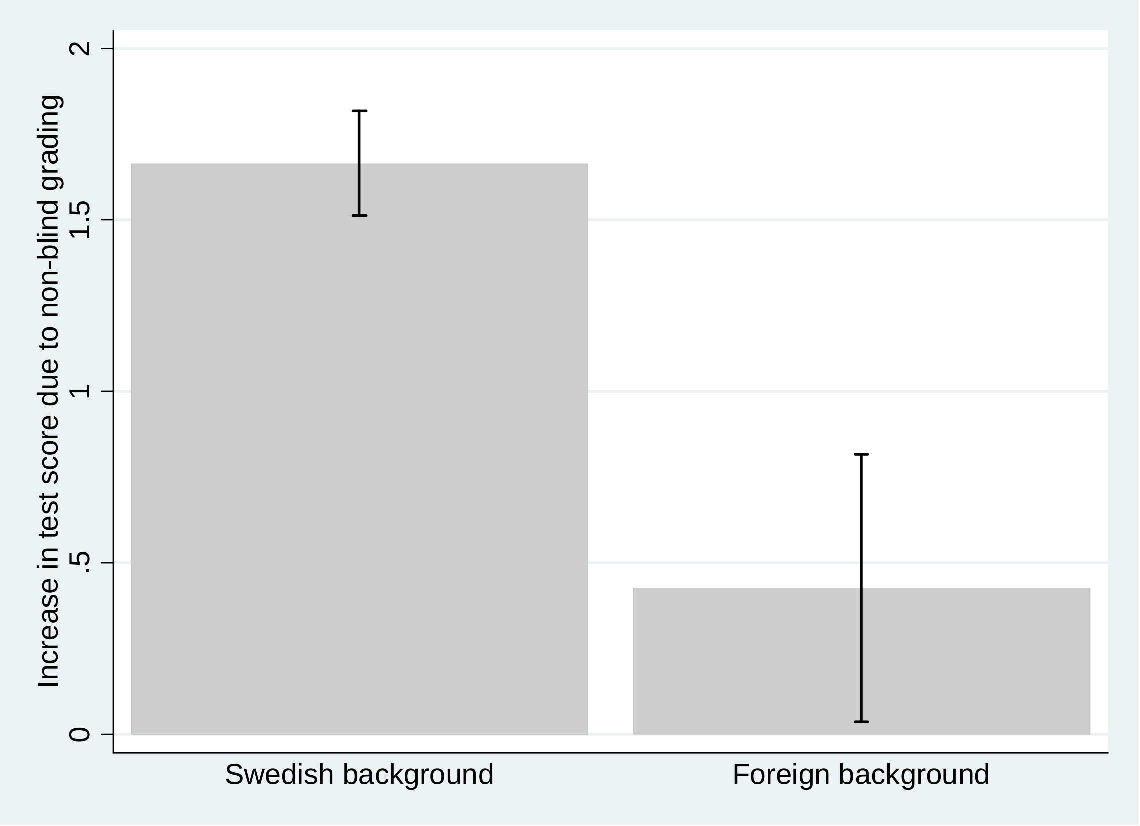

Figure 2. The increase in the test score due to non-blind grading for students with Swedish background (n=1423) and students with foreign background (n=199). Error bars denote ± one standard error.

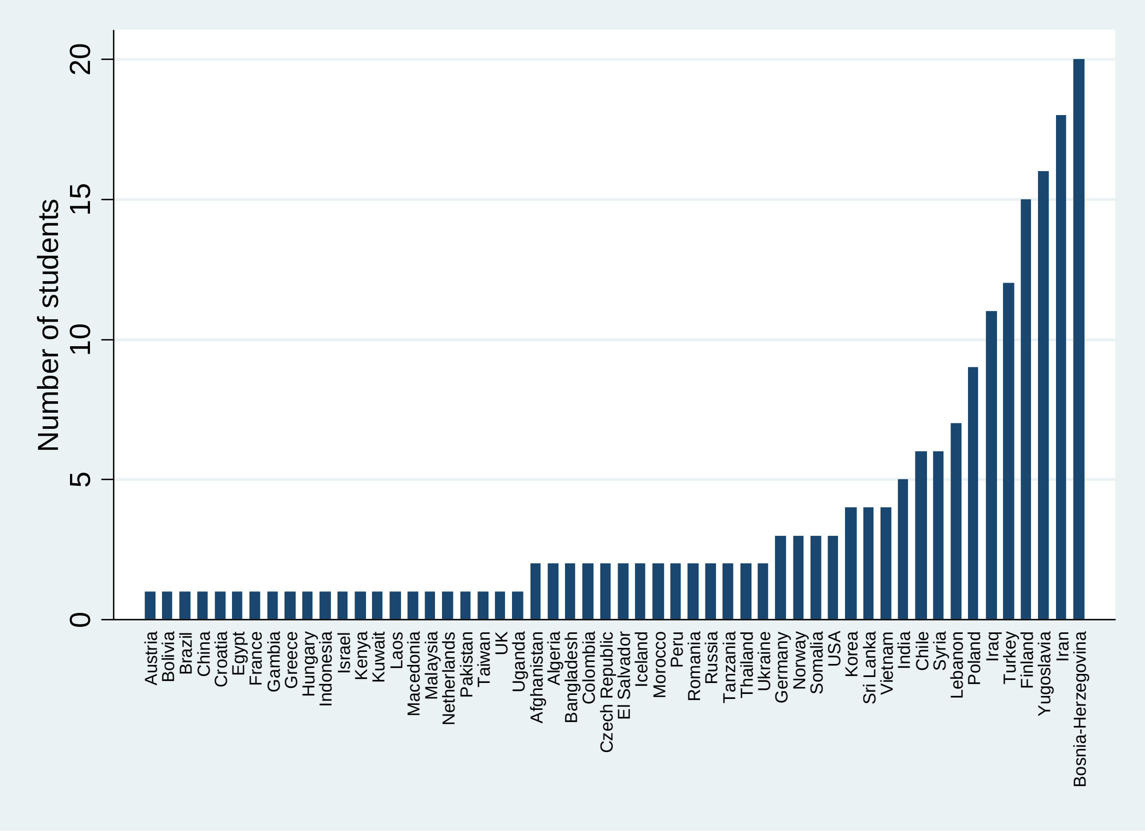

Figure 3. Students' national background