ENDOPHILIA OR EXOPHOBIA: BEYOND DISCRIMINATION

Jan Feld, Nicolás Salamanca and Daniel S. Hamermesh*

ABSTRACT

The discrimination literature treats outcomes as relative. But does a differential arise because agents discriminate against others—exophobia—or because they favor their own kind—endophilia? Using a field experiment that assigned graders randomly to students' exams that did/ did not contain names, on average we find favoritism but no discrimination by nationality, and some evidence of favoritism for the opposite gender. We identify distributions of individuals' preferences for favoritism and discrimination. We show that a changing correlation between them generates perverse changes in market differentials and that their relative importance informs the choice of a base group in adjusting wage differentials.

*Feld and Salamanca: Ph.D. candidates in economics, Maastricht University; Hamermesh: Sue Killam Professor of Economics, University of Texas at Austin, professor in economics, Royal Holloway University of London, and research associate, IZA and NBER. We thank Jeannette Hommes, Ad van Iterson and Caroline Kortbeek for their assistance in making this experiment possible. Eric Bonsang, George Borjas, Deborah Cobb-Clark, , Thomas Dohmen, Hannah Ebin, Matthew Embrey, Andries de Grip, Ilyana Kuziemko, Corinne Low, Arjan Non, Christopher Parsons, Joseph Price, Stephen Trejo, participants in seminars at a number of universities and institutes, and especially Leigh Linden provided very helpful comments. The Board of Examiners of the School of Business and Economics at Maastricht University formally approved this project.

> Although we could not perceive our own in-groups excepting as they contrast to out-groups, still the in-groups are psychologically primary. Hostility toward out-groups helps strengthen our sense of belonging, but it is not required. [Allport, 1954]

I. Introduction

Economists have studied labor-market discrimination at least since Becker (1957). Differences in labor-market and other outcomes by race, gender, ethnicity, religion, weight, height, appearance and other characteristics have been examined in immense detail, over time and in many economies. The focus has, however, been nearly exclusively on measuring differences in outcomes between groups, under the assumption that the "majority" group's outcome is the norm while the "minority" group is discriminated against. But since the only concept that is measured is a difference, it could just as easily be that the majority group is favored while the minority group's outcome is the norm.

The possibility that we are measuring the extent of favoritism rather than discrimination has been pointed out by Goldberg (1982) and by Cain (1986) in his survey; but beyond that the issue appears to have been completely neglected, including by the more recent *Handbook* surveys of the literature on discrimination (Altonji and Blank, 1999; Fryer, 2011). Once we recognize that favoritism need not be the obverse of discrimination, the importance of studying preferences for favoritism/discrimination increases. Although the distribution of discriminating agents' tastes underlay Becker's theory, in most empirical research the demand side—the behavior of discriminatory agents—has not been studied explicitly. Only recently has there been even a small upwelling of interest in examining their behavior and its impacts on outcomes.[1] These studies typically consider how agents' behavior toward those who match them along some dimension differs from their behavior toward those who do not match them, again only estimating relative differences. Even then, most of these studies have looked only at averages, and none has combined this with the analysis of the distributions of preferences.

[1] See Price and Wolfers (2010) and Parsons *et al* (2011) for evidence from professional sports; Fong and Luttmer (2009) on charitable giving; Dee (2005), Lavy (2008), Hinnerich *et al* (2011), and Hanna and Linden (2012) for examinations of education; Cardoso and Winter-Ebmer (2010) and Giuliano *et al* (2011) on wages and hiring; Baguës and Esteve-Volart (2010) on parliamentary elections; and Dillingham *et al* (1994), Donald and Hamermesh (2006) and Abrevaya and Hamermesh (2012) for studies of economists' behavior.

Here we discuss the results of a field experiment that allows us to identify separately the means of favoritism and discrimination, as well as their distributions. The key to doing this that, instead of measuring differences in outcomes *between* groups, we compare outcomes of members of the *same* group with and without visible characteristics that reveal to which group they belong.[2] In the context of our experiment, we do this by randomly revealing or concealing names on students' final exams, and thus randomly allowing or not allowing graders to infer the gender and nationality of the students. Because of the random assignment, students without visible names on their exams have on average the same observable and unobservable characteristics as students with visible names on their exams. Students without visible names thus serve as a neutral baseline to identify discriminatory preferences. Differences from this baseline can be entirely attributed to the presence of the name—and by inference to favoritism/discrimination.[3] Hence, we have evidence for favoritism if members of a group are treated better when their names are visible. Conversely, we can infer the presence of discrimination if members of a group are treated worse when their names are visible. We focus specifically on favoritism/discrimination by gender and nationality, but this method could be applied to any of the groups that have been studied in this immense literature.

To distinguish clearly the *who* and the *how* in discrimination, we introduce four terms: Endophilia, endophobia, exophilia, and exophobia. The prefix *endo* refers to preferences towards people like oneself, the prefix *exo* to people unlike oneself. The suffixes *philia* and *phobia* refer to favoritism to discrimination. Hence, *endophilia* denotes preferences for member of one's own group, while *exophobia* denotes preferences against members of other groups. One can also imagine, however, that some agents

[2] A number of studies (e.g., Goldin and Rouse, 2000, Burgess and Greaves, 2013) have focused on "blindness" in quasi-experimental situations to infer the extent of discrimination (or favoritism, since neither study could distinguish between these).

[3] The only experiments like ours were conducted in laboratories (Fershtman *et al*, 2005; Ahmed, 2007). The latter had artificially-designated in- and out-groups; the former dealt with nationalities but was based on statements by students on how they would behave in a trust game. While laboratory evidence is useful, as discussed by Levitt and List (2007) it suffers from a number of difficulties that can be addressed in field experiments.

prefer members of other groups—are *exophilic*, while other agents are *endophobic*—discriminate against people like themselves.

II. Theoretical and Empirical Motivation

The importance of the distinction between favoritism and discrimination can be seen both theoretically and empirically. Our theoretical work is a generalization of Becker's (1957) theory of discrimination and Goldberg's (1982) alteration of it. Goldberg adapted Becker's model to show that if favoritism toward one's own group drives observed, apparently discriminatory wage differentials, these differentials can persist in a competitive market. He reached this conclusion by assuming that employers have favoring *instead of* the discriminatory preferences as in Becker (1957). Employers can, however, have both discriminatory *and* favoring preferences. We extend Becker's model to show that if both preferences are present, the intergroup wage differential will not only depend on the distributions of favoritism and discrimination but also on the relationship between their distributions.

Assume, as Becker does, that all employers are White and that there is a fixed labor force, some fractions of which are White and Black. Let employers have endophilic and exophobic preferences simultaneously, so that we can characterize a typical employer's utility as:

$$U = Q(L_w + L_b) - W_w L_w - W_b L_b + eL_w - xL_b \ ,$$

where we assume as usual that White and Black workers are perfect substitutes in the production function Q for the good sold at a constant price of unity. This utility function implies that employers obtain or forgo a fixed amount of utility when they hire Whites or Blacks according to their preferences.[4] Analogous to both Becker's and Goldberg's models, employers here will hire Whites or Blacks depending on whether $W_w - W_b$ is smaller or larger than $\delta = e + x$, the relative preference for Whites over Blacks.

In equilibrium, Blacks will be employed by employers with lower values of δ and Whites will be employed by the remaining employers. Knowing this, we can find a simple expression that implicitly

[4] Goldberg and Becker model discrimination and favoritism as wage premia and discounts rather than as a fixed utility gain or loss. Although more complex, the model here and the general intuition do not change if we choose to model discrimination and favoritism as they do.

identifies the preferences of the marginal employer. Assume that the distribution of endophilia is $f(e)$ and of exophobia is $h(x)$ across the population of jobs on offer, and denote the relative preference of the marginal employer, for whom $W_w - W_b = e + x$, as δ^*. Then in long-run equilibrium the share of Blacks in the economy determines δ^*:

$$\frac{L_b}{L_b + L_w} = \int_{-\infty}^{\delta^*} g(\delta) d\delta,$$

where $g(\delta)$ is the density function of δ which, in general, will depend on the densities f and h and their relation. If $e \sim N(\mu_e, \sigma_e^2)$ and $x \sim N(\mu_x, \sigma_x^2)$:

$$\frac{L_b}{L_b + L_w} = \Phi\left(\frac{\delta^* - \mu_e - \mu_x}{\sqrt{\sigma_e^2 + \sigma_x^2 + 2\sigma_e\sigma_x\rho}}\right),$$

where Φ is the cumulative standard normal distribution and ρ is the correlation between endophilic and exophobic preferences in the population of jobs being offered. It is easy to see that, keeping the shares of Blacks and Whites and the means and variances of the densities f and h constant, an increase in ρ will result in a marginal discriminator with weaker relative preferences for Whites over Blacks. The increase in ρ increases the variance of δ, the sum of preferences, while keeping its mean constant. This leads to a distribution of relative preferences for Whites with more extreme values; more employers now have an extremely strong and extremely weak relative preference for Whites. Because Blacks are hired by employers who are most favorable to them, the marginal discriminator now has a lower relative preference for Whites. A more positive correlation of endophilia and exophobia thus leads to a decrease in the absolute value of the equilibrium Black-White wage gap.[5] In other words, for the same means and variances of endophilic and exophobic preferences, and holding constant the share of Black workers, the wage gap is smaller if the most bigoted (against Blacks) employers are also those who favor Whites most.

[5] See Charles and Guryan (2008) for a discussion of the empirical importance of the marginal discriminator.

By allowing agents to have endophilic and exophobic preferences at the same time, our model becomes a more general version of both Becker's and Goldberg's, effectively nesting both cases.[6] The model predicts that the wage gap increases if employers become more exophobic (as in Becker) and as they become more endophilic (as in Goldberg). By allowing endophilia and exophobia to co-exist, however, the model introduces an additional force that can shape market outcomes, the correlation of the two types of preferences.

The concepts of endophilia, exophobia, and their correlation can be measured, albeit imperfectly, in the real world. Beginning in 1996, and biennially except in 2002, the U.S. General Social Survey has asked questions, "In general, how close do you feel to Whites [Blacks]?" with answers on a nine-point scale ranging from 9 = very close to 1 = not close at all. Table 1 describes these data, separating answers by Whites and Blacks, and pooling 1996-2000 as an early period, 2004-2006 as a later period. (We exclude the 2008 and 2010 data because the campaign and election of President Obama may have altered expressed preferences.) Several things stand out: 1) Unsurprisingly, expressed closeness to one's own group exceeds that to the other group; 2) While Whites' closeness to other Whites changed little over this period, there was a very large increase in their expressed closeness to Blacks; 3) There are only small changes in Blacks' expressed closeness to either Whites or Blacks; and 4) The correlation between expressed closeness to one's own group and the other is positive and increased (significantly) between the two sub-periods. Implicitly, those who favor members of their own group more disfavor members of the other group less, or, in our terminology, there was an increasing negative correlation between endophilia and exophobia.

To illustrate how thinking about endophilia and exophobia jointly can add to our understanding of discriminatory outcomes, consider the implications of the GSS data for the evolution of the Black-White wage gap. Assume for simplicity that the share of Black workers remained constant and that all employers are White. In Table 1 we can see between 1996-2000 and 2004-2006 Whites' endophilia remained

[6] If we assume that endophilia is non-existent, the model reverts to Becker's; if we assume that exophobia is non-existent, it reverts to Goldberg's.

constant, Whites' exophobia (the negative of the measure in the Table) decreased, and the correlation between endophilia and exophobia decreased (became more negative).[7]

Becker's model (where only exophobia matters) predicts that the decline in exophobia shown in the Table would decrease the wage gap. Goldberg's model (where only endophilia matters) predicts that the wage gap would remain constant. Our model captures one force that tended to decrease the gap—the reduction of exophobic preferences; and one that tended to increase it—the more negative correlation between endophilia and exophobia. Perhaps these opposite forces contributed to the constancy of the black-white earnings ratio over this period, although with so many other shocks over this short period attributing changes is difficult.

III. Constructing the Experiment

A. The Environment

To make the distinction between favoritism and discrimination empirically we set up a field experiment that we carried out during the final exam week in June 2012 at the School of Business and Economics (SBE) of Maastricht University in The Netherlands. The language of instruction throughout the SBE is English. This environment has a number of features that make it particularly appropriate for distinguishing between favoritism and discrimination. Partly because Maastricht is near the German border, the SBE has a large share of German students (51 percent) and academic staff (22 percent) mixed with Dutch and other nationalities. The student population is 36 percent female, and the academic staff is 28 percent female.[8] German students have a reputation for being more hard-working than Dutch and other

[7] Their correlation decreases because Whites' closeness to Blacks, as reported in Table, is measuring exophilia, the opposite of exophobia.

[8] The SBE homepage (http://www.fdewb.unimaas.nl/miso/index.htm) provides these statistics for enrolled students in 2010 for nationality and 2012 for gender. Statistics about staff refer to full-time-equivalent academic staff in 2012 and are taken from the internal information system "Be Involved."

students. These contrasts by nationality could potentially be the basis for discrimination/favoritism, although it is unclear *a priori* in which direction these will be.[9]

The grading of final exams, which we examine here, is a good setting for identifying discrimination/favoritism, because graders do not gain anything from favoring or disfavoring specific groups. Also, until the teaching period that we examine all students were required to write their names on their exams, enabling the graders to identify the students' gender and nationality.[10] Finally, and most important, this experiment has real-world consequences: The grades are important to students; also, much of the graders' jobs revolves around their role in scoring exams.

In the SBE written exams are administered in ten sessions spread over a week, with many courses giving their exams simultaneously. Students in all the courses assigned to each session take their exams together in a large conference hall filled with desks that are arranged in blocks of 5 columns and 10 rows.[11] To prevent cheating the location of each student's desk is predetermined by the Exams Office (the organization responsible for examination procedures). The desk assignment is based on student ID numbers, first by sorting them from lowest to highest within each block, and then filling in sequentially within each column from left to right.[12] Figure 1 illustrates the arrangement of desks in each block.

[9] While it is often found that people favor (discriminate against) groups with same (different) characteristics, there are also situations in which the opposite is the case. One can, for example, think of many situations in which relative outcomes suggest that males are exophilic or endophobic (e.g., Donald and Hamermesh, 2006, although that study cannot distinguish between these two types of preferences).

[10] The grader can infer the nationality and gender of the students when she sees the family name, even if she does not know the student, because Dutch and German names are quite distinct. To test this we asked 9 staff (5 German and 4 Dutch, of whom 5 were female) to guess the nationality and gender of 50 student names from our sample. We selected the student names block-randomly to reflect the nationality mix in our sample (19 German, 17 Dutch and 14 other nationalities, of whom 16 were female). The staff correctly identified the German names in 64 percent and Dutch names in 65 percent of cases, and they correctly guessed gender in 90 percent of the cases. On the other hand, graders may be more able to infer student gender than nationality from handwriting *per se*.

[11] Exams in courses with more than 50 students are written in the same session in multiple blocks. Exams in courses with fewer than 50 students are either kept in one block or are combined with the exams in other courses. There are a few blocks that have as many as 12 rows.

[12] Student IDs are assigned in ascending order based on the moment a prospective student contacts Studielink (the Dutch centralized system for university application; https://app.studielink.nl/front-office/). This means that earlier cohorts have lower-number IDs, and later cohorts and exchange students have higher-number IDs.

B. The Experiment and Data Collection

The students in each session arrive at the exam hall and locate their assigned block based on the course they are taking. Within the block they then locate their assigned desk, which is marked with their student ID number. Once the exam session starts students have three hours to complete their exams. During that time one invigilator (not the same person as the exam grader) supervises each block. We asked the invigilators to place yellow sheets on all desks in the first three rows of each block (see Figure 1), thus ensuring that the recipients were mixed by ID number, and thus were more or less randomly treated by seniority in the University. The sheets stated that the students on whose desks one was placed should *not* write their name but *only* their ID number on the exam sheets (see Figure A1 in the Appendix).[13] Because of the predetermined arrangement of desks this meant that a random sample of students within each course—the "*blind*" group—was asked not to write their names, so that the grader would only observe their ID numbers when grading. For the rest of the students—the *"visible"* group—graders could observe both names and IDs, as in previous teaching periods.

We collected additional information from several other sources. The Exams Office provided us with the nationality and gender of the students, grades in previous courses, and the desk arrangement during the exam. From the seating arrangement we could infer which students were asked not to write their names (yellow sheets, rows 1-3) and which were allowed to do so. To check students' compliance with the experiment's instructions, we manually went through all the exams and noted which students wrote down their names and which students did not.[14]

At the SBE it is common practice to split the grading burden among various graders by letting each one handle all the answers to a particular set of questions on the same exam. The course coordinators

[13] We placed the sheets on entire rows instead of scattered seats within each block for simplicity. We treated rows instead of columns in order to capture students with a variety of high and low ID numbers within each course. The Exams Office informed the course coordinators—who were in charge of organizing the grading of the exams—before the examination period that a new examination procedure was being tested, so that some exams might only have ID numbers. They were asked to have those exams graded as they usually would.

[14] This was done immediately after the exam, before the course coordinators received the exams and started the grading process.

identified the grader of each question and provided us with information on the grading. This information included the score on each question and the maximum possible points per question. They also provided other grades that the student had attained in the course, including on course participation, presentation and any term paper.[15] A survey sent after the grading to all graders and course coordinators provided information on the grader's gender, nationality, teaching experience and grading behavior during the experiment.[16] From the SBE's online tool for course evaluations we gathered the total number of courses in which the grader had been involved at the SBE and the average instructor evaluations provided by students for that grader in all previous courses since the creation of the online tool. Our sample contains 25 out of the 42 courses that had final exams, including 42 different graders and 1,495 exams.[17]

The upper part of Table 2 examines the internal validity of the experiment, testing whether the questions in the treated (Visible) group were answered by students whose characteristics before they entered the examination room differed in measurable dimensions from those in the untreated (Blind) group. We present these results separately for those students whom we intended to treat (ITT) and those who were actually treated.[18] We first examine differences by gender and nationality, the two characteristics on which we focus, and in the students' grades before the final exam. The Blind and Visible groups are balanced in both gender and nationality: The p-values indicate that none of the tests of differences in the means between the Blind and Visible groups along the dimensions that form the focus of this study can reject the hypothesis that they are zero. Indeed, not only are the fractions of men and

[15] Most course coordinators had this information readily available in an Excel file. We manually collected the scores on each exam question for 7 courses.

[16] We manually added the gender and nationality of the graders who did not fill out the survey. Grading behavior includes whether graders looked up any names while grading.

[17] We excluded 8 courses that only used Multiple Choice or Fill-In-The-Blank questions. In 7 out of the 34 eligible courses the coordinators either declined permission to use the data or did not respond to repeated requests for this information. We excluded one course for which the answer sheets did not ask for the students' names but only for their IDs and another course which did not hold the exam in the conference hall.

[18] The blind treatment group had a little over 80-percent effectiveness, and an additional 2 percent of the students got into the blind group but should not have. This latter was most likely due to mistakes by the invigilators when placing the yellow sheets or by students forgetting to write their names.

women, Germans and Dutch, insignificantly different from each other; the absolute differences between the Blind and Visible groups are never greater than two in the second decimal place.

We have additional information on some of the students—other grades that were received before the exams were given, such as prior grade point average (GPA), and classroom participation, presentation in class and term-paper grades in the particular course. We find no significant differences between the Blind and Visible students in GPA and their participation grades. The Visible group performs slightly better in the grades assigned for student presentations. This difference is not quite statistically significant, however; and perhaps more important, grades for classroom presentations were given to only about one-third of the students.

We also have grades from Multiple Choice and Fill-In-The-Blank questions that were included in a minority of the final exams. We can thus test whether, despite the apparent randomness of assignment, outcomes differed between the two groups on questions on which the grading was unambiguous and could not have been affected by the mechanisms we study here. As the bottom part of Table 2 shows, the Blind group did have marginally higher scores on the Multiple Choice questions, but here too the differences are not quite statistically significant. These results confirm that the research design created equivalent groups of students.[19]

IV. Inferring Average Outcomes and Distributions of Preferences

Let a student, denoted by *s*, answer an exam with several questions, and let the grader of each question be denoted by *g*. We index each answer by the pair (*s*, *g*).[20] We also know the pair (C(*s*),C(*g*)), where C is either some student-invariant bivariate characteristic, such as gender, or some characteristic vector, such as nationality. Finally, we know whether a particular answer by a particular student was

[19] Considering that we tested several separate characteristics, it is not unlikely that some of those tests will reject the null hypothesis at the 10 percent level purely by chance. If we correct the p-values for multiple testing (using the Bonferroni, Šidák, or Holm adjustments), we find no significant differences between Blind and Visible students in any of the characteristics, even at the 10 percent level of significance.

[20] We ignore course identifiers for simplicity, since all graders except one were uniquely assigned to one course.

graded blind or visible, so that each pair (C(s),C(g)) can be expanded to the triplet (C(s),C(g),v), where v=1 if the grading is visible and 0 if not.[21]

Consider the score function S(C(s),C(g),v) for each exam question, where we are especially interested in examining how S varies between cases when s and g match (i.e. share a common characteristic) and when they do not, and how that variation is affected by v. Define the following indicators:

(1a) I1{(C(s),C(g), v)} = 1, if C(s)=C(g) and v=1, 0 if not;

(1b) I2{(C(s),C(g),v)} = 1, if C(s)=C(g) and v=0, 0 if not;

(1c) I3{(C(s),C(g),v)} = 1, if C(s)≠C(g) and v=1, 0 if not;

and

(1d) I4{(C(s),C(g),v)} = 1, if C(s) ≠C(g) and v=0, 0 if not.

The average score of all students is:

(2) $T = \theta_1 S^*(\mathrm{I1}) + \theta_2 S^*(\mathrm{I2}) + \theta_3 S^*(\mathrm{I3}) + [1- \theta_1- \theta_2 - \theta_3] S^*(\mathrm{I4})$,

where the weights θ_i are the shares of answers graded under each regime, and the (*) denotes an average over those answers.[22] Because we created the neutral categories with blind grading, we can estimate the average treatment effect on students for whom C(s) = C(g) (i.e., grader and student "match" on characteristic C) as:

(3a) $e* = [S^*(\mathrm{I1}) - S^*(\mathrm{I2})]$;

and the treatment of students for whom C(i) ≠ C(g) (who do not "match" on C) as:

(3b) $x*= [S^*(\mathrm{I4}) - S^*(\mathrm{I3})]$.

If graders are endophilic and exophobic, $e*$, $x* > 0$. Identifying endophilia and exophobia as $e*$ and $x*$ relies on the assumption that graders are neutral towards blind exams. In Section V we present estimates

[21] Presumably all particular (s, g) combinations are either blind or visible (although we investigate the extent of blindness in the blind grading in Section VI).

[22] While the same average would apply for a n-fold characteristic if we focus only on whether or not C(s)=C(g), we could analogously and generally calculate n^2 average treatment effects, one for each of the n aspects of the characteristic compared to itself and each other aspect.

of each of the effects as discussed here. We discuss the implications of alternative behavioral assumptions in Section VI.

From Equation (2) we can also recover the average "total" effect of the characteristic C(*s*) for a particular value, C(*s*) = C'. This is particularly important if we want to address the question of whether disclosing certain information (such as gender or nationality) affects an outcome, given a distribution of preferences and graders. Consider a variant of (2):

(4) $T_C = \eta_1 S^*(\mathrm{I1}|\mathrm{C}(s)=\mathrm{C'}) + \eta_2 S^*(\mathrm{I2}|\mathrm{C}(s)=\mathrm{C'}) + \eta_3 S^*(\mathrm{I3}|\mathrm{C}(s)=\mathrm{C'}) + [1- \eta_1- \eta_2 - \eta_3] S^*(\mathrm{I4}|\mathrm{C}(s)=\mathrm{C'})$,

where the weights η represent the shares of answers graded under each regime for all students with characteristic C(*s*) = C'. The total treatment effect of a particular characteristic C' being observable is the weighted average of the treatments when C(*g*) = C' and when C(*g*) ≠ C'. Thus:

(5) $M_{C'} = (\eta_1 + \eta_2)[S^*(\mathrm{I1}|\mathrm{C}(g)=\mathrm{C'}) - S^*(\mathrm{I2}|\mathrm{C}(g)=\mathrm{C'})] - (1- \eta_1 - \eta_2)[S^*(\mathrm{I4}|\mathrm{C}(g)=\mathrm{C'}) - S^*(\mathrm{I3}|\mathrm{C}(g)=\mathrm{C'})]$.

Equation (5) shows that the average treatment effect of a characteristic will depend on two factors: 1) The degree of endophilia and exophobia (the two bracketed expressions); and 2) The share of questions that are graded by graders with matching characteristics ($\eta_1 + \eta_2$) versus non-matching characteristics ($1- \eta_1 - \eta_2$).

We can also observe the behavior of individual graders toward the student groups as defined by C(*s*). Each grader scores answers written by many different students, some with characteristics that match hers, others with characteristics that do not match, some of whom are graded Blind, others graded Visible. Then for a grader *g* we can calculate her average treatment of students, T^g, in a manner analogous to the average effect in (2) and obtain a distribution over all graders. More interesting for our purposes, we can estimate each grader's preferences for students who do and do not match their characteristics as:

(6a) $e^g = S^{*g}(\mathrm{I1}) - S^{*g}(\mathrm{I2})$;

and

(6b) $x^g = S^{*g}(\mathrm{I4}) - S^{*g}(\mathrm{I3})$,

where $S^{*g}(\mathrm{I}j)$, j=1,2,3,4, is the average over all students whose exams are scored by grader *g* under each regime I*j*. Using these grader-specific average treatments, we can then obtain non-parametric estimates of the distributions of endophilia and exophobia, *f(e)* and *h(x)*, as discussed in Section II. Thus, in addition to being able to distinguish the average extent of favoritism toward one's own group from the average extent of discrimination against other group(s), the data allow us to obtain complete distributions of agents' implicit preferences.

V. Empirical Strategy and Basic Results

To estimate the impacts of nationality and gender matches on the points that graders assigned to students' answers, and to infer the differences discussed above, we estimate the regression:

$$(7) \quad S = \beta_1 MATCH*VISIBLE + \beta_2 MATCH*BLIND + \beta_3 NON\text{-}MATCH*VISIBLE + \beta_4 NON\text{-}MATCH*BLIND + \gamma' Z + \varepsilon,$$

where here *S* is a unit normal deviate calculated for each exam question, and the other variable names are self-explanatory.[23] The matrix Z includes nationality or gender indicators for both students and graders, ε is a zero-mean error term and the regression is estimated without a constant. From this equation the estimates of the average extent of endophilia and exophobia are:

$$(8a) \quad e* = S^*(\mathrm{I1}) - S^*(\mathrm{I2}) = \beta_1 - \beta_2,$$

and:

$$(8b) \quad x* = S^*(\mathrm{I4}) - S^*(\mathrm{I3}) = \beta_4 - \beta_3.$$

Thus the estimates of (7) provide direct analogs to the concepts we seek to measure. Note that these calculations mean that endophilia (exophobia) is indicated by a positive *e** (*x**).

One special benefit that we obtain from our setting is that we can be sure that the implied preferences on matching are not being driven by confounding factors like unobserved heterogeneity. In our experimental setting we are comparing arguably identical groups whose only difference—because the treatment was random—is that the graders observed the names of some but not of other students. The experiment allows us explicitly to compare e.g., Visible to Blind German students. This means that anything specifically German, such as writing style in English or particular calligraphic patterns, washes

[23] The distribution of the standardized question scores is roughly normal and slightly negatively skewed, but it is the same for all four groups defined by *VISIBLE*, *BLIND*, *MATCH*, and *NON-MATCH*.

out in this comparison. This framework also makes it easy to expand Equation (7) to include interactions with some of the graders' measurable characteristics and thus to examine how e^* and x^* vary with them. We deal with these extensions in Section VI.

The first two columns of Table 3 present the estimated β and their standard errors for the basic equations describing matches/non-matches along the criteria of nationality and gender. Since the experimental design randomized by blocks of students within each course, we cluster the standard errors at the Intention-To-Treat and course (ITT-course) level, allowing for two clusters per course. We focus throughout on the estimates of e^* and x^* and their statistical significance.

It is clear that there is substantial endophilia by nationality in the grading. A student who matches the grader's nationality receives a score that is 0.17 standard deviations higher when her name is visible than when it is not. This addition to a matched student's grade is statistically significant at conventional levels. This effect is also economically important: Given that all the scores have been unit-normalized, this effect is equivalent to moving from the median score to the 57th percentile of the distribution of scores. Its magnitude is similar to that of the effect of large differences in teacher quality on students' test scores that was found by Rivkin *et al* (2005). While favoritism by nationality exists in grading, there is no apparent exophobia by nationality: The estimated impact of being visible when not matching by nationality is small and positive.

The results of estimating the regression examining gender matching are shown in Column (2) of Table 3. Although the point estimate suggests the existence of a small degree of endophilia, we cannot reject the hypothesis that it is zero. For non-matches there is exophilia, but here too the impact is statistically insignificant and also minute. On average grading seems gender-neutral in all dimensions.[24]

Going behind the information in Columns (1) and (2), we can ask whether, for examples, endophilia by nationality is the same for Dutch and German graders, and whether the absence of

[24] The results are also essentially the same when we include additional controls for seat number (see Figure 1) and the student's prior GPA.

endophilia or exophobia exists for both male and female graders. We do this by expanding Equation (7) to include interactions of student nationality or gender with *MATCH*VISIBLE*, *MATCH*BLIND*, *NON-MATCH*VISIBLE*, and *NON-MATCH*BLIND.* Columns (3) of Table 3 show the estimates of this expanded specification by nationality. A comparison of the results suggests that endophilia by nationality arises more from the behavior of Dutch than of German graders, although the difference between the two point estimates is not statistically significant.

Columns (4) of Table 3 show estimates of expanding Equation (7) by gender. The results look very much like those in Column (2): Neither male nor female graders exhibit significant endophilia or exophobia, and for both men and women the absolute impacts are small. Again, there is no sign of either statistically significant or important differences in behavior depending on the match or non-match of the grader's and student's gender.

VI. Robustness and Extensions

A. Treatment Failures

In interpreting these main results it is important to note that there are two potential sources of slippage in our treatment: Some students did not comply with the experimental instructions shown in the Appendix and mistakenly wrote their names on the exam sheets; and some graders may have looked up at least some of the students' names.[25] To account for the first source of slippage we re-estimated the models described in the first two columns of Table 3 using intention to treat (ITT) as an instrument for *VISIBLE*. As the first two columns of Table 4 show, the results are qualitatively identical to the ones in Table 3.[26]

To account for the second source of slippage—that the grader may have been able to identify the characteristic of the Blind group—in the post-grading survey we asked graders whether they looked up any names on the exams that only contained ID numbers. Six of the thirty-three graders who responded to the survey acknowledged having done this. When we re-estimated (7) including only those graders who

[25] Evidence on the magnitude of the first type of slippage can be seen in Table 2 in the differences between ITT and Treatment.

explicitly stated that they did not look up names, the estimated endophilia by nationality is the same and is even more precisely estimated. The last column of Table 4 shows there is no significant endophilia but significant exophilia by gender among those graders who did not look up names. The results of both slippages suggest that, if anything, our results understate the true extent of favoritism by nationality and gender.

B. Alternative Behavioral Assumptions

So far we have implicitly assumed that the graders are indifferent toward "blind" exams and treat these groups as a neutral baseline against which we measure endophilia and exophobia. Graders, however, might also form rational expectations about the "blind" exams, considering the underlying *distribution* of characteristics of students who wrote those exams and might score them accordingly. Let A_g be the share of students in the course who match grader g on the characteristic of interest, and let e^{re} and x^{re} be the grader's latent endophilic and exophobic preferences. Under rational expectations we can rewrite Equation (7) as:

(7') $S = e^{re} \cdot MATCH{*}VISIBLE + x^{re} \cdot NON\text{-}MATCH{*}VISIBLE - [e^{re} \cdot A_g + x^{re} \cdot (1 - A_g)] + \gamma_2{'}Z + \varepsilon,$

where, from the grader's perspective, the students can either visibly match him, visibly not match him, or be in the Blind group (the omitted category). Equation (7') specifies that the grader will treat the students in the Blind group as the weighted average of how he would have treated students who matched him or not, with weights based on the characteristics of the Visible groups.

To determine whether assuming rational expectations about the Blind group's students can alter our results, we estimate (7') by non-linear least squares. The results confirm our main findings: We again find endophilia by nationality, although of slightly lesser but still statistically significant magnitude (0.133 standard deviations, p=0.025). We find no endophilia by gender and no exophobia by either gender or nationality. Moreover, the root mean square error of estimating (7') exceeds that of the estimate of (7). Because the blind-as-neutral assumption fits the data better, we continue defining endophilia and exophobia as discussed in Section IV throughout the rest of the study.

C. Prior Grader-Student Contact, and Exam Type

The graders and exams differ along several dimensions on which we have information and which might affect their ability or interest in favoring/discriminating for/against students. We first look at whether the graders knew the students they graded, and thus whether endophilia/exophobia is present towards anonymous and familiar students alike.[27] We have no specific hypothesis on this possibility. On the one hand, it could be that prejudices are overridden by personal experience with the students. If so, discriminatory preferences will be stronger toward unknown students. On the other hand, it might not be the characteristic *per se* that the graders pay attention to, but something that graders can only observe on students with whom they interact. In this case discriminatory preferences will be stronger toward and against students whom the grader knows.

We construct an indicator of whether the grader may know a student based on whether the grader also taught him or her. Most of the teaching at the SBE is done in tutorials of 10 to 15 students for about 10 sessions in each seven-week block, so teachers have a fair chance to get to know their students. Some graders taught none of the students they graded, others taught all of the students they graded. By this measure the median grader knew 47 percent of the students graded (although obviously in most cases the grader could not identify individual students in the Blind group).

The first two columns of Table 5 present re-estimates of Equation (7), expanded to include interactions of the Know indicator with the four Match/Visible variables. The results show that endophilia by nationality is only present when graders did not know the students. This effect is twice as large as the mean effect in the baseline model. There is no evidence of exophobia by nationality regardless of whether the grader knew the student or not. There is evidence of endophilia and exophilia by gender, but again only when the grader did not know the student.

[27] The assignment of students and teachers to classes within a course is done by the Scheduling Department of the SBE, which does not consider students' preferences for particular teacher or teachers' preferences for a particular class. (See Feld and Zölitz (2014) for a detailed explanation on the assignment of students and teachers to classes at the SBE.) Also, the students have no way of knowing *ex ante* who their grader will be.

The exams at the SBE differ in the extent to which they have mathematical questions, depending mostly on the nature of the courses. Answers on the more mathematical exams are arguably less ambiguous, so that showing favoritism/discrimination on them might be more difficult. To separate the more from the less mathematical exams we asked three raters (from the SBE's pool of potential graders) to rate the exams as mathematical or not. Two of the three agreed in their categorizations of all the exams, while the third agreed in 80 percent of the cases. We thus created an indicator for Mathematical when at least two of the three raters designated an exam as such, which occurred for 9 out of 25 exams.

The third and fourth columns of Table 5 present estimates of Equation (7), expanded to include interactions of the Mathematical indicator with the main variables. The point estimates suggest that endophilia by nationality is stronger for less mathematical exams. The point estimates for exophilia by nationality and endophilia by gender are also significant for the more mathematical exams. This latter result is surprising, as one might expect that Blind exams might be less likely to be assignable to nationality or gender based on handwriting styles if the exam is more mathematical. None of the other results in the two columns is statistically significant.

D. *Distinguishing by Graders' Other Characteristics*

We also examine whether discrimination or favoritism varies with grader experience and grader quality. We measure grader experience at this University as the number of separate courses taught or tutored during the grader's tenure. We have no hypotheses about how university-specific experience might mitigate or exacerbate endophilia/exophobia. On the one hand, the set of more experienced graders may exclude those whose behavior was so egregiously unfair that the University did not renew their contracts. On the other hand, more experienced graders may be secure in their positions and feel able to indulge their preferences for students who match their characteristics and/or against those who do not.

The total number of courses taught/tutored at the University since the online data became available (including the courses we are using here) ranges from 1 to 94; the 5th, 50th and 95th percentiles, for which we present estimation results, are 1, 8 and 59 courses.[28] Figure 2 shows the kernel density of courses taught by grader, which demonstrates the distribution's very long right tail. The first and second columns of Table 6 present re-estimates of Equation (7), expanded to include interactions of grader experience with the four match/visible variables.

While the point estimate of the extent of endophilia by nationality is almost identical to the estimate in Table 3 at the median value of grader experience, it is not quite significantly nonzero. Rather, the significant average endophilia shown in Table 3 results disproportionately from the behavior of the more experienced graders. By inference, they feel less inhibited about indulging their preferences for students who match their nationality. Inexperienced graders, perhaps because they feel themselves to be under greater scrutiny, show no significant endophilia (although the point estimate of their behavior is 60 percent of that of highly experienced graders). As with the basic estimates, there is no evidence of exophobia by nationality at any level of grader experience. The results by gender remain very similar: Just as at the sample means, so too at various levels of grader experience the parameter estimates show no sign of any significant endophilia or exophobia. The exception is the evidence of exophilia by gender for the most experienced graders.

We measure grader quality as the average of all the evaluations that the instructor received from students during her career at the University. Evaluations are given on a ten-point scale. In our sample the averages range from 6.5 to 9.2, with the 5th percentile being 7.1, the median being 8.0, and the 95th percentile equaling 8.8. As Figure 3 shows, the distribution of average evaluations is quite close to symmetric.

We interact the grader's average instructional evaluation with all the variables in Equation (7) and present the results in Columns (3) and (4) of Table 6. Our finding of endophilia by nationality at the mean demonstrated in Table 3 arose from behavior that varies sharply with the regard in which graders have been held by students. Those graders/instructors who have been rated highest by students show no

[28] 59 and 94 might seem outlandishly large; but at this University there are 6 teaching blocks in each academic year, so it is not difficult to accumulate 50 or more courses of experience.

significant endophilia, and the point estimate of this effect is small. An instructor whose teaching has been rated at the median of this measure behaves much like the mean instructor—substantially favoring those who match her nationality, unsurprisingly given the symmetry in the distribution of teaching evaluations. The worst-rated instructors, however, favor those students who match their nationality much more strongly than does the median or average instructor. Implicitly a poorly rated instructor raises the score of the median student who matches her nationality from the mean to the 61st percentile of the distribution of scores. There is no evidence of exophobia by nationality. In a similar fashion, the little evidence there was of exophilia by gender seems to be driven by the worst-rated teachers. In sum, worse teachers behave differently from better ones, favoring students of their own nationality and, to a lesser extent, the other gender.

VII. The Average Treatment Effect of Visible Student Characteristics

To evaluate whether the visibility of names differentially favors or disadvantages certain groups of students, and also to see how these students would be affected by the introduction of anonymous grading, we calculate the average treatment effect (ATE) of each characteristic's visibility. Recall from Equation (5) that the ATE can be calculated as the difference between endophilia and exophobia, each weighted by the share of questions that was graded by graders with matching and non-matching characteristics. Table 7 shows the ATE of being seen as German, Dutch, or any other nationality, and of being seen as female or male. The point estimates for German and Dutch students are similar in size and (marginally) significantly positive, demonstrating that both German and Dutch students benefit from visible grading. The point estimates further suggest that other nationalities are disadvantaged by it, although the ATE is not statistically significant. Even if they are not suffering from an absolute disadvantage, however, the notion that other nationalities are disadvantaged becomes straightforward for situations in which they compete with German and Dutch students. An example is the allocation of student exchange positions at popular universities abroad, which is done based on relative grades. The difference between Germans and others is significant (p=0.004), as is the difference between Dutch and others (p=

0.025). Consistent with our previous results, the point estimates for females and males are positive but smaller in size.

Columns (1) to (4) of Table 7 decompose the ATE by showing endophilia and exophobia (Columns (2) and (4)) and the share of students with the given characteristic that was graded under each regime (Columns (1) and (3)). (The estimated effects of endophilia and exophobia are taken from Table 3.) The ATE for German and Dutch students is small because of the relatively small shares of questions that are graded by graders of the same nationality. It is easy to simulate the sizes of these effects for a situation in which a large share of the students in either category were matched to the graders. Notice also that the mix of graders is not always the most important determinant of the ATE: The difference between the effects when matched and not matched for females is rather small, so that the ATE will be small regardless of the gender mix of graders.

VIII. Heterogeneity in the Distributions of Preferences

The results thus far describe either average responses of endophilia or exophobia by nationality or gender over all graders, or examine how this behavior differs in relation to a few of the graders' specific characteristics. This parallels but expands upon the focus in the literature on average differences between groups. In this section we move to a different dimension, the distribution of implicit tastes for favoritism and discrimination, first considering the shapes of the entire distributions of graders' preferences and then calculating their correlations, as suggested by the theoretical discussion.

To obtain a feel for why examining heterogeneity in preferences might be interesting, consider the kernel density estimates of the graders' endophilia and exophobia by nationality, shown in Figure 4, and their kernel densities by gender, shown in Figure 5. Each kernel is based on those graders for whom we could infer the extent of both endophilia and exophobia (for nationality, 24 graders, for gender, 38 graders).[29] The estimates along the criterion of nationality suggest that preferences are distributed fairly

[29] We derive the shape of the graders' preferences based on the estimates of e^g and x^g calculated as in equations (6a) and (6b). We infer these two measures for each grader based on how each scores the student who does or does not match them under the blind and visible regimes.

symmetrically, in the case of endophilia around a positive mean, and around zero in the case of exophobia. Both densities are consistent with our inferences in Table 3 about the mean effects. A similar conclusion is suggested by the kernels of endophilia and exophobia by gender, although there are a few outliers.[30]

By observing the entire distribution of preferences we can also test two hypotheses: 1) There is evidence of endophilia or exophobia in the overall distribution (not just at the mean), and 2) There is heterogeneity in endophilia or exophobia among graders. Testing these two hypotheses is equivalent to testing whether e^g=0 (x^g=0) for all *g*, and whether the e^g (x^g) are equal to each other for all *g*, respectively. The F-tests of these hypotheses (eight in total) all reject the null hypothesis at conventional significance levels, showing that endophilia and exophobia in both nationality and gender are real phenomena (even though at the means only endophilia by nationality seems to matter), and that there is significant heterogeneity in these preferences across graders.[31]

As we showed in Section II, the impact of the interaction of endophilia and exophobia depends on their correlation across potentially discriminating agents. In our data the correlations are -0.36 for preferences on nationality, and -0.16 for preferences on gender (weighting each grader by the number of students graded). Those who are more endophilic are less exophobic. Interestingly, and remarkably, in the GSS data summarized in Table 1, the correlations are in the same direction: Those Whites who feel closer to Whites also feel closer to Blacks, and to roughly the same extent as implied by behavior in our sample.

IX. Conclusions and Implications

We have demonstrated that what is called discrimination—a relative difference in outcomes between two groups—is composed of differential treatment of the in-group and the out-group, and that it

[30] We can examine whether extreme values in the distributions of preferences for nationality or gender are driving our mean effects. We trim those graders with the most extreme preferences from the samples, dropping the two most extremely endophilic/endophobic and exophobic/exophilic graders in each case. Despite the small amounts of asymmetry in some of the distributions, trimming does not qualitatively alter the conclusions about the absence of endophilia or exophobia by gender on average, nor does it alter the conclusions about these outcomes by nationality.

[31] The demonstrated heterogeneity of preferences should reduce any concern about the absence of exophobia at the mean because the possibly large (psychological) costs of giving lower grades. Also, it should vitiate concerns about our behavioral assumption on how graders treat Blind exams, since it shows that Visible non-matches are treated differently from Blind ones by most graders.

is possible in real-world situations to measure the sizes of these two components simultaneously. In our example we find that most of the apparent discrimination by nationality results from substantial endophilia and that there is no evidence on average of exophobia. We find some evidence of graders favoring the opposite gender on average, though it is less definitive.

These are average effects. At least as interesting is the heterogeneity in the demonstrated preferences of the individuals deciding how to treat those who match or do not match them. We have further shown that apparently discriminatory outcomes can be vitiated in a variety of ways, operating both on the endophilic and exophobic preferences of the discriminating agents and the share of matching and non-matching characteristics.

We also show the importance of measuring the relation between endophilia and exophobia in the labor market: Their joint distribution will influence market-based measures of discrimination. This result makes it even clearer that they are non-redundant measures. It also forces us to reconsider what we know about the effectiveness of anti-discrimination policies and the advances against discrimination in the labor market. The change over time in measures of discrimination, such as the market discrimination coefficient, may not only reflect a change in the means and variances of the distributions of expressed preferences. It may also reflect a change in the correlation between endophilia and exophobia. This changing correlation might explain the unchanging Black-White wage gap over a period when racial attitudes appear to have become more tolerant.

Assuming that the dominance of endophilia over exophobia that we have demonstrated for nationality is ubiquitous in labor markets, the fact has important implications for the measurement of "discrimination" in labor markets. Decompositions that adjust a gross wage differential into parts due to different characteristics or different treatments in the labor market can be made using either the majority or the minority wage as the base case. In the literature (e.g., Neumark, 1988; Booth *et al*, 2007; Elder *et al,* 2010) that discusses these decompositions of wage differentials (by race, gender, and many others) a crucial question has been which group's actual wage to treat as the baseline. Endophilia dominating exophobia would suggest using the minority group's wage as the baseline and adjusting the wages of the majority. More generally, if we knew the relative importance of each type of behavior, the appropriate treatment would be a weighted average of the different methods of decomposition.

Having shown that we can distinguish endophilia from exophobia, it is also worth considering how policy might be tailored to reduce relative differences arising from prejudice. Assume that our results carry over to the labor and other markets, and that endophilia is the main source of apparently discriminatory outcomes. If so, we can infer, for example, that moral suasion that stresses to members of the majority group that minority-group members are not “bad” might be ineffective.

Can the distinctions that we have defined and measured here be inferred in the still more important labor-market context using actual wage and/or employment outcomes? One might imagine cases where a majority group deals with several minority groups, about one of which it feels demonstrably neutral. In that case too endophilia and exophobia (toward the other minorities) are identifiable. So too, one might link differences in economic outcomes to information on attitudes in a population about one’s own and other groups. The main point is that these preferences generate different outcomes with different distributions of welfare, so that determining their relative sizes is economically important and, as we have shown, possible.

References

Jason Abrevaya and Daniel Hamermesh, “Charity and Favoritism in the Field: Are Female Economists Nicer (to Each Other)?” *Review of Economics and Statistics*, 94 (Feb. 2012): 202-7.

Ali M. Ahmed, “Group Identity, Social Distance and Intergroup Bias,” *Journal of Economic Psychology*, 28 (2007): 324-37.

Gordon Allport, *The Nature of Prejudice*. Cambridge, MA: Addison-Wesley, 1954.

Joseph Altonji and Rebecca Blank, “Race and Gender in the Labor Market,” in Orley Ashenfelter and David Card, eds., *Handbook of Labor Economics, Vol 3C*. Amsterdam: North-Holland, 1999, pp. 3143-3259.

Wiji Arulampalm, Alison Booth and Mark Bryan, “Is There a Glass Ceiling over Europe? Exploring the Gender Pay Gap across the Wage Distribution,” *Industrial and Labor Relations Review*, 60 (Jan. 2007): 163-86.

Manuel Baguës and Berta Esteve-Volart, “Can Gender Parity Break the Glass Ceiling? Evidence from a Repeated Randomized Experiment,” *Review of Economic Studies*, 77 (Oct. 2010): 1301-28.

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

Simon Burgess and Ellen Greaves, “Test Scores, Subjective Assessment, and Stereotyping of Ethnic Minorities,” *Journal of Labor Economics*, 31 (July 2013): 535-76.

Glen Cain, “The Economic Analysis of Labor Market Discrimination: A Survey,” in Orley Ashenfelter and Richard Layard, eds., *Handbook of Labor Economics, Vol. 2*. Amsterdam: North-Holland, 1986, pp. 693-785.

Ana Rute Cardoso and Rudolf Winter-Ebmer, “Female-Led Firms and Gender Wage Policies,” *Industrial and Labor Relations Review*, 64 (Oct. 2010): 143-63.

Kerwin Charles and Jonathan Guryan, “Prejudice and Wages: An Empirical Assessment of Becker’s *The Economics of Discrimination*,” *Journal of Political Economy*, 116 (Oct. 2008): 773-809.

Thomas Dee, “A Teacher Like Me: Does Race, Ethnicity or Gender Matter?” American Economic Association, *Papers and Proceedings*, 95 (May 2005): 158-65.

Alan Dillingham, Marianne Ferber and Daniel Hamermesh, “Gender Discrimination by Gender: Voting in a Professional Society,” *Industrial and Labor Relations Review*, 47 (July 1994): 622-33.

Stephen Donald and Daniel Hamermesh, “What Is Discrimination? Gender in the American Economic Association, 1935-2004,” *American Economic Review*, 96 (Sept. 2006): 1283-92.

Todd Elder, John Goddeeris and Steven Haider, “Unexplained Gaps and Oaxaca-Blinder Decompositions,” *Labour Economics*, 17 (Jan. 2010): 284-90.

Jan Feld and Ulf Zölitz, “On the Nature of Peer Effects in Academic Achievement,” Unpublished paper, SBE, Maastricht University, 2014.

Chaim Fershtman, Uri Gneezy and Frank Verboven, “Discrimination and Nepotism: The Efficiency of the Anonymity Rule,” *Journal of Legal Studies*, 34 (June 2005): 371-96.

Christina Fong and Erzo Luttmer, “What Determines Giving to Hurricane Katrina Victims? Experimental Evidence on Racial Group Loyalty,” *American Economic Journal: Applied Economics,* 1 (April 2009): 64-87.

Roland Fryer, “Racial Inequality in the 21st Century: The Declining Significance of Discrimination,” in Orley Ashenfelter and David Card, eds., *Handbook of Labor Economics, Vol. 4B,* 2011, Amsterdam: Elsevier, pp. 855-971.

Laura Giuliano, David Levine and Jonathan Leonard, “Racial Bias in the Manager-Employee Relationship: An Analysis of Quits, Dismissals and Promotions at a Large Retail Firm,” *Journal of Human Resources,* 46 (Winter 2011): 26-52.

Matthew Goldberg, “Discrimination, Nepotism and Long-Run Wage Differentials,” *Quarterly Journal of Economics,* 97 (May 1982): 307-19.

Claudia Goldin and Cecilia Rouse, “Orchestrating Impartiality: The Impact of “Blind” Auditions on Female Musicians,” *American Economic Review,* 90 (Sept. 2000): 715-41.

Rema Hanna and Leigh Linden, “Discrimination in Grading,” *American Economic Journal: Economic Policy,* 4 (Nov. 2012): 146-68.

Björn Tyrefors Hinnerich, Erik Höglin and Magnus Johannesson, “Are Boys Discriminated in Swedish High Schools?” *Economics of Education Review,* 30 (Aug. 2011): 682-90.

Victor Lavy, “Do Gender Stereotypes Reduce Girls’ or Boys’ Human Capital Outcomes? Evidence from a Natural Experiment” *Journal of Public Economics,* 92 (Oct. 2008): 2083-105.

Steven Levitt and John List, “What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World?” *Journal of Economic Perspectives,* 21 (Spring 2007): 153-74.

David Neumark, “Employers’ Discriminatory Behavior and the Estimation of Wage Discrimination,” *Journal of Human Resources,* 23 (Summer 1988): 279-95.

Christopher Parsons, Johan Sulaeman, Michael Yates and Daniel Hamermesh, “Strike Three: Discrimination, Incentives and Evaluation,” *American Economic Review,* 101 (June 2011): 1410-35.

Joseph Price and Justin Wolfers, “Racial Discrimination among NBA Referees,” *Quarterly Journal of Economics,* 125 (Nov. 2010): 1859-87.

Steven Rivkin, Eric Hanushek and John Kain, “Teachers, Schools and Academic Achievement,” *Econometrica,* 73 (March 2005): 417-58.

Table 1. Endophilia and Exophobia in the U.S. General Social Survey, 1996-2006, 9-point scale*

Time period:	1996-2000	2004-2006
WHITES		
Feel Close to Whites	7.060	6.966
	(0.031)	(0.038)
Feel Close to Blacks	5.121	5.494
	(0.032)	(0.039)
N	3,550	2,174
P	0.146	0.226
BLACKS		
Feel Close to Whites	5.799	5.945
	(0.084)	(0.106)
Feel Close to Blacks	7.547	7.685
	(0.079)	(0.093)
N	651	387
P	0.242	0.318

*In general, how close do you feel to …? not close at all = 1; very close = 9.

Table 2. Student Characteristics by Intended and Actual Treatment Status*

		Internal validity: Pre-experiment						
		(1)			(2)			
		Blind			Visible			p-value of difference
		Mean	SD	N	Mean	SD	N	Blind-Visible
Female	ITT	0.369	0.483	452	0.352	0.478	1,043	[0.502]
	Treatment	0.363	0.482	399	0.355	0.479	1,096	[0.758]
German	ITT	0.374	0.484	452	0.353	0.478	1,043	[0.420]
	Treatment	0.373	0.484	399	0.354	0.478	1,096	[0.486]
Dutch	ITT	0.363	0.481	452	0.343	0.475	1,043	[0.452]
	Treatment	0.351	0.478	399	0.349	0.477	1,096	[0.932]
GPA	ITT	7.197	0.628	443	7.215	0.665	1,021	[0.607]
	Treatment	7.178	0.618	389	7.221	0.667	1,075	[0.241]
Participation	ITT	7.690	0.986	306	7.633	1.031	706	[0.386]
	Treatment	7.612	0.968	263	7.664	1.035	749	[0.452]
Presentation	ITT	7.795	1.164	191	7.930	1.059	436	[0.179]
	Treatment	7.758	1.172	181	7.942	1.055	446	[0.070]
Term paper	ITT	7.870	0.665	109	7.743	0.898	281	[0.126]
	Treatment	7.870	0.697	97	7.748	0.882	293	[0.166]
		Internal validity: Within-experiment						
		(1)			(2)			
		Blind			Visible			p-value of difference
		Mean	SD	N	Mean	SD	N	Blind-Visible
Multiple Choice	ITT	5.829	1.972	277	6.043	1.942	661	[0.128]
	Treatment	5.792	2.009	253	6.049	1.928	685	[0.078]
Fill-In-The-Blank	ITT	5.325	2.208	152	5.555	1.996	382	[0.264]
	Treatment	5.367	2.167	148	5.536	2.016	386	[0.411]

*The pre-experiment validity only includes students in the estimation sample. The within-experiment validity uses information on students who participated in the experiment, but the information on these answers is not part of our analysis. The p-values of differences between the Visible and Blind groups are calculated with clustered standard errors by student.

Table 3. Basic Estimates of the Extent of Favoritism and Discrimination by Nationality and Gender (N = 9330)*

	(1)	(2)		(3)		(4)	
	Nationality	Gender		Nationality		Gender	
Interaction with:	-	-	*German*	*Dutch*	*Other*	*Female*	*Male*
*(1) MATCH*VISIBLE*	0.287	-0.039	0.306	-0.012	-	0.156	-0.039
	(0.038)	(0.025)	(0.021)	(0.099)	-	(0.028)	(0.027)
*(2) MATCH*BLIND*	0.115	-0.099	0.165	-0.204	-	0.101	-0.101
	(0.081)	(0.039)	(0.101)	(0.106)	-	(0.075)	(0.042)
*(3) NON-MATCH*VISIBLE*	0.177	-0.076	0.148	-0.048	-0.123	0.150	-0.101
	(0.050)	(0.040)	(0.070)	(0.053)	(0.067)	(0.047)	(0.046)
*(4) NON-MATCH*BLIND*	0.172	-0.101	0.060	-0.095	-0.035	0.053	-0.071
	(0.057)	(0.047)	(0.080)	(0.077)	(0.072)	(0.038)	(0.079)
Endophilia [(1)-(2)]	0.172	0.060	0.140	0.193	-	0.055	0.062
p =	[0.028]	[0.140]	[0.171]	[0.049]	-	[0.471]	[0.188]
Exophobia [(4)-(3)]	-0.005	-0.025	-0.088	-0.047	0.088	-0.097	0.030
p =	[0.904]	[0.700]	[0.095]	[0.528]	[0.174]	[0.124]	[0.740]
Adj. R2	0.015	0.009		0.016		0.010	

*Standard errors in parentheses and p-values in square brackets. Both are clustered by ITT-Course. Columns (1) and (2) present the estimates of Equation (7) without a constant. Columns (3) and (4) are based on Equation (7), with the main variables interacted with CHARACTERISTIC, where CHARACTERISTIC are indicators for nationality in (3) and for gender in (4). MATCH*Other interactions in (3) are empty because we define MATCH = 1 only for German and Dutch students. Other nationalities almost never matched. Main effects are included throughout, when not perfectly collinear with the main coefficients.

Table 4. The Effects of Treatment Slippage by Students and Graders on Estimates of Endophilia and Exophobia*

	(1) Nationality	(2) Gender	(3) Nationality	(4) Gender
Regression:	IV		Did Not Look Up Names	
Endophilia	0.193	0.090	0.174	0.009
p =	[0.034]	[0.190]	[0.009]	[0.856]
Exophobia	-0.033	-0.039	-0.008	-0.119
p =	[0.538]	[0.595]	[0.878]	[0.010]
N	9,330	9,330	5,108	5,108
Adj. R2	0.015	-0.001	0.015	0.007

*p-values in squared brackets, based on standard errors clustered at the ITT-Course level. We report linear combinations based on extensions of Equation (7). Columns (1) and (2) are based on an instrumental variable regression (IV) estimated by 2SLS, where we use the intention to treat (ITT) to instrument for the treatment. The F-tests for the instruments always strongly reject the null. Columns (3) and (4) are based on Equation (7) using only graders who did not look up any of the names in the Blind group of exams. Main effects are included throughout.

Table 5. Endophilia and Exophobia When Graders Know the Students They Grade, and When the Exams are Mathematical (N = 9,330)*

		(1)	(2)			(3)	(4)
		Nationality	Gender			Nationality	Gender
Grader knows the student?:				*Exam was mathematical?:*			
Endophilia	*No*	0.320	0.120	Endophilia	*No*	0.228	0.034
	p =	[0.003]	[0.042]		p =	[0.039]	[0.578]
	Yes	0.052	-0.001		*Yes*	0.060	0.094
	p =	[0.580]	[0.980]		p =	[0.375]	[0.042]
Exophobia	*No*	-0.070	-0.112	Exophobia	*No*	0.055	0.002
	p =	[0.120]	[0.040]		p =	[0.345]	[0.975]
	Yes	0.070	0.085		*Yes*	-0.095	-0.081
	p =	[0.427]	[0.396]		p =	[0.020]	[0.197]
F-test differences:		[0.066]	[0.166]			[0.024]	[0.584]

*p-values in squared brackets, based on standard errors clustered at the ITT-Course level. We report linear combinations based on extensions of Equation (7). Columns (1) and (2) report interactions of the main variables with GRADERKNOWSSTUDENT, Columns (3) and (4) of the main variables with MATHEMATICALEXAM. F-test differences reports the p-values from testing the null hypothesis that Endophilia and Exophobia are equal for the groups defined by GRADERKNOWSSTUDENT and MATHEMATICALEXAM, respectively. Main effects are included throughout.

Table 6. Effects of Grader Experience and Grader Teaching Quality on Outcomes (N = 9197)*

		(1)	(2)	(3)	(4)
	Percentile:	Nationality	Gender	Nationality	Gender
At the m^{th} percentile of:		*Experience*		*Teacher Quality*	
Endophilia	5th	0.154	0.076	0.307	0.039
	p =	[0.162]	[0.141]	[0.020]	[0.575]
	50th	0.166	0.073	0.170	0.068
	p =	[0.097]	[0.118]	[0.130]	[0.108]
	95th	0.248	0.045	0.048	0.093
	p =	[0.001]	[0.517]	[0.651]	[0.158]
Exophobia	5th	-0.024	0.008	-0.014	-0.140
	p =	[0.639]	[0.919]	[0.859]	[0.048]
	50th	-0.016	-0.007	-0.005	-0.013
	p =	[0.718]	[0.920]	[0.896]	[0.842]
	95th	0.045	-0.121	0.002	0.100
	p =	[0.635]	[0.053]	[0.962]	[0.289]
F-test interactions:		[0.547]	[0.261]	[0.356]	[0.075]

*p-values in square brackets, based on standard errors clustered at the ITT-Course level. We report linear combinations based on extensions of Equation (7). Columns (1) and (2) interact the main variables with TEACHEREXPERIENCE and evaluate the linear combinations at different percentiles. Columns (3) and (4) do the same with TEACHERQUALITY. F-test interactions reports the p-values from testing the joint significance of interactions of Endophilia and Exophobia with TEACHEREXPERIENCE and TEACHERQUALITY, respectively. Main effects are included throughout.

Table 7. The Average Treatment Effect (ATE) of the Visibility of Student Characteristics[*]

			(1)	(2)	(3)	(4)
Student	Total ATE	p-value	Share matched	Endophilia	Share not matched	Exophobia
			(η1+η2)		*(1-η1-η2)*	
German	0.103	[0.049]	0.29	0.140	0.71	-0.088
Dutch	0.107	[0.067]	0.41	0.193	0.59	-0.047
Other	-0.088	[0.174]	-	-	1	0.088
Female	0.078	[0.142]	0.45	0.055	0.55	-0.097
Male	0.027	[0.548]	0.62	0.062	0.38	0.030

*The ATE is calculated as shown in Equation (5). The p-values are based on standard errors clustered at the ITT-Course level. Columns (1) and (3) show the share of questions, for a given characteristic, which were graded by graders with matching and non-matching characteristics. Columns (2) and (4) show the ATE on the treated, as reported in Table 3.

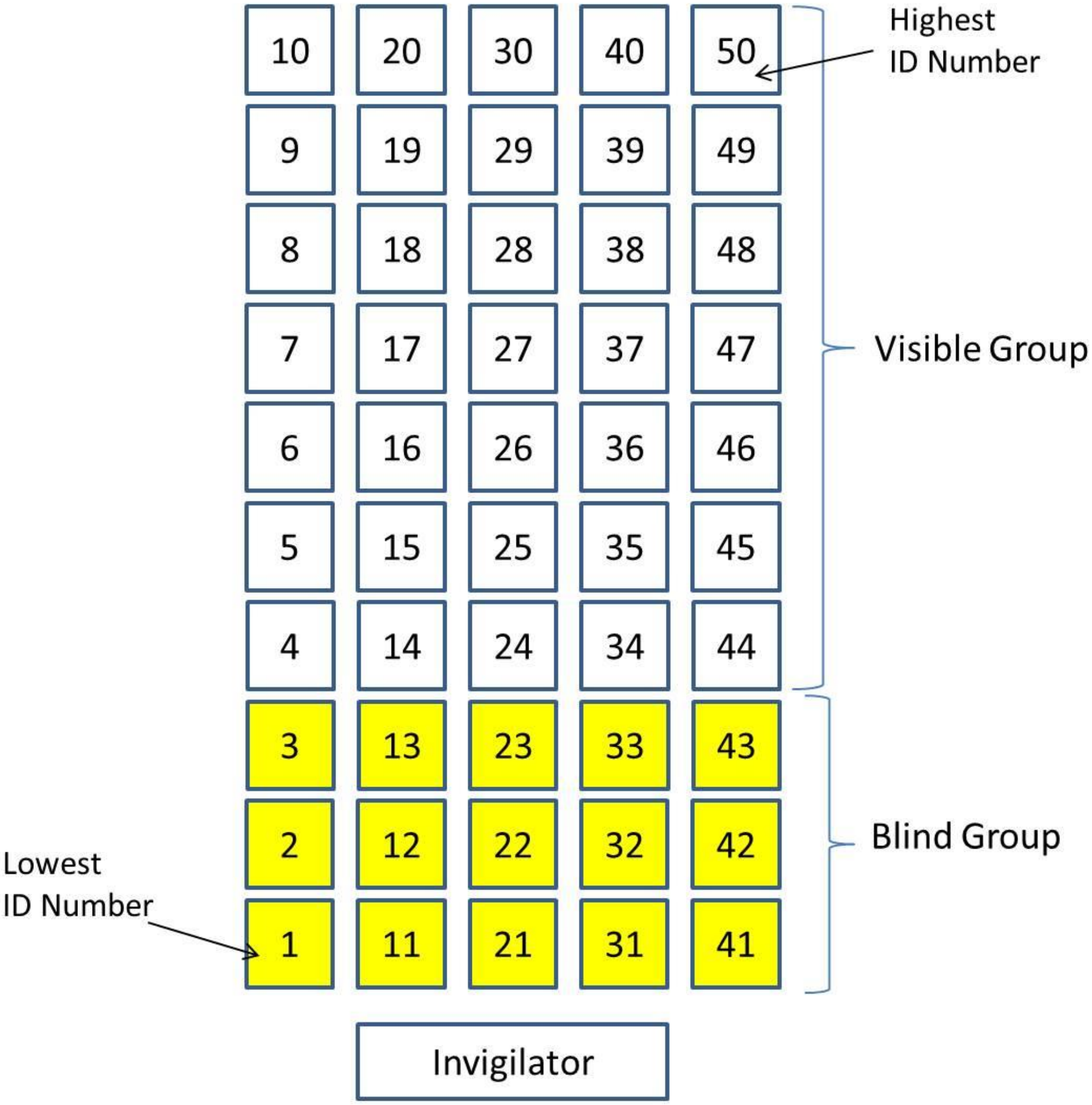


Figure 1: Seating Arrangement for the Experiment[*]

[*] One square represents one desk. Students were seated in order of their ID numbers. Each number indicates the order of student ID numbers in each block. The student with the lowest ID number sat in desk 1, the one with the highest ID in desk 50. Rows 1-3 had yellow sheets on the desks with instructions not to write their name, thus creating the Blind group. Rows 4-10 had no extra sheets. In these rows students were expected to write their name to create the Visible group.

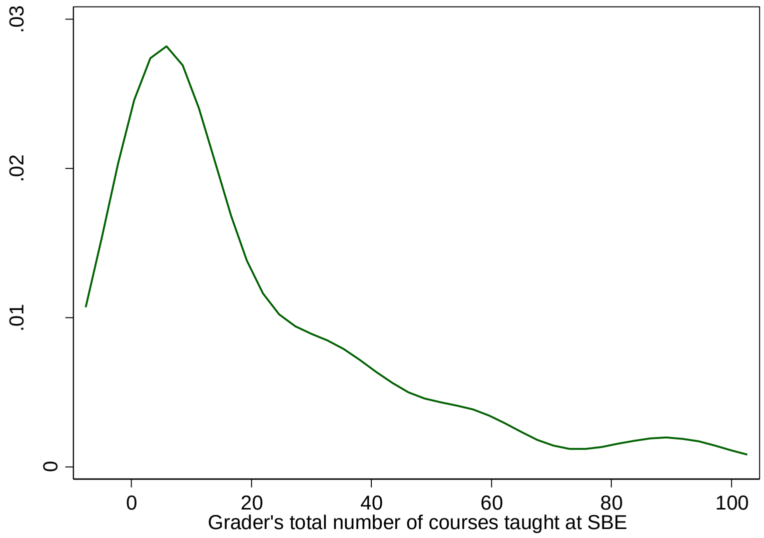


Figure 2. Kernel Density of the Distribution of Grader Experience

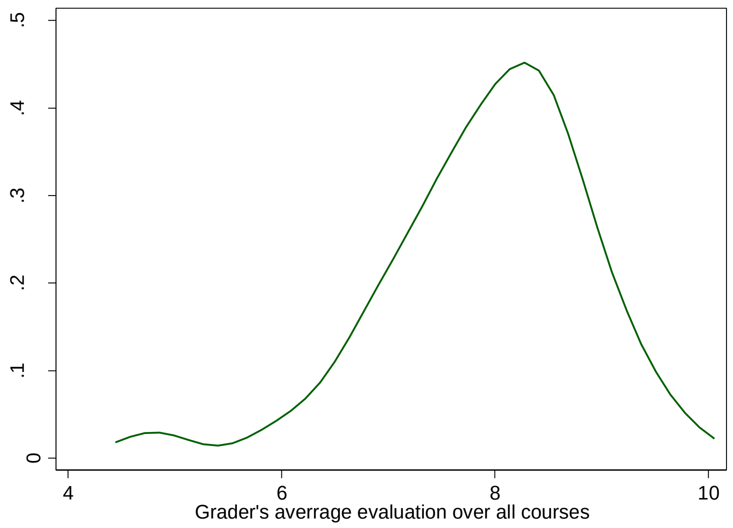


Figure 3. Kernel Density of the Distribution of Student Evaluations of Graders

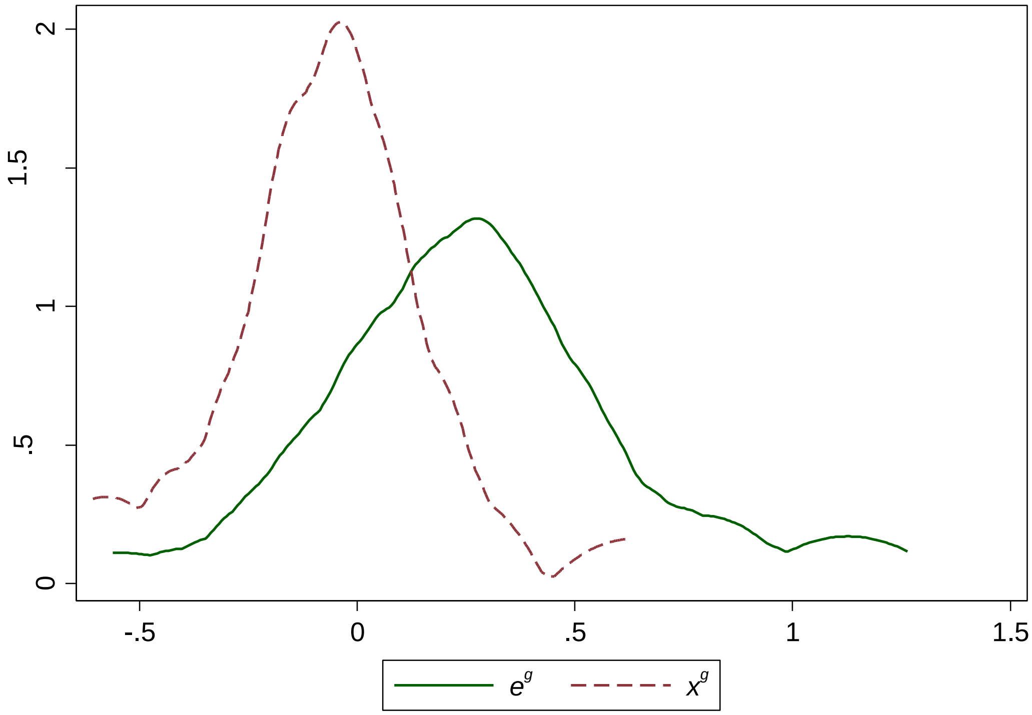


Figure 4. Kernel Density Estimates of Graders' Preferences by Nationality

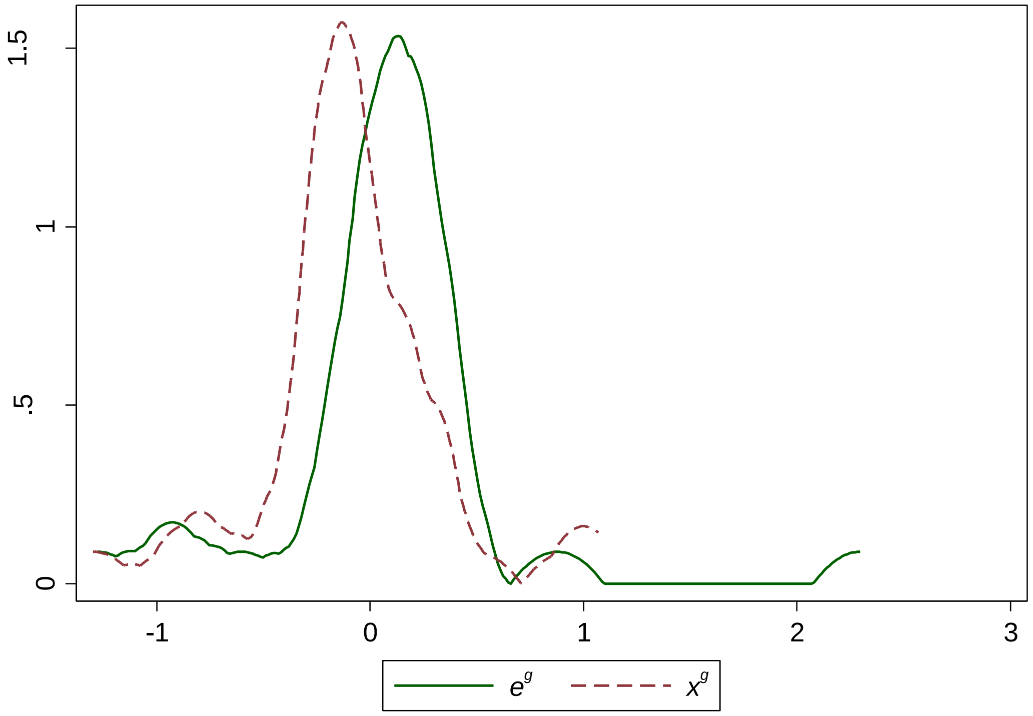
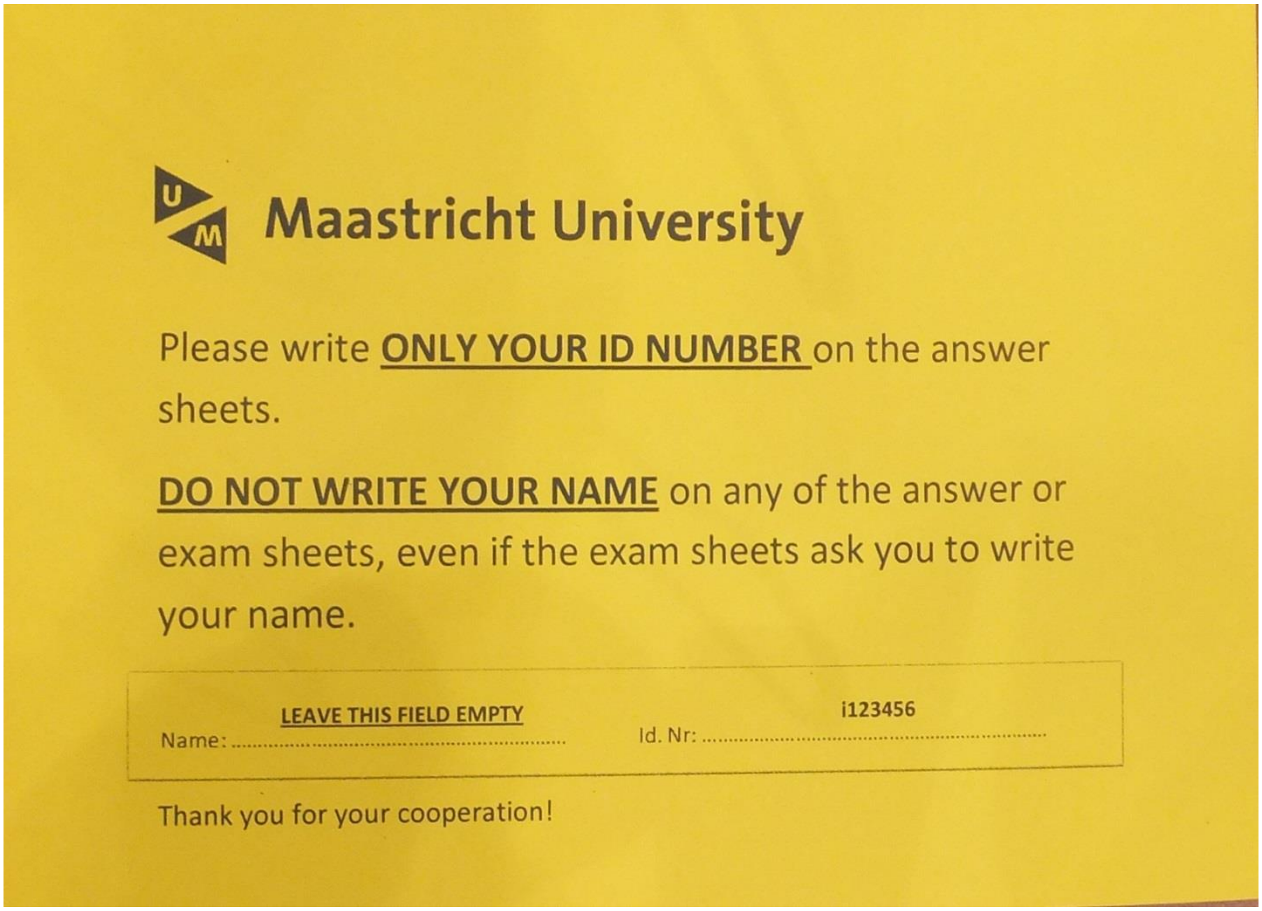


Figure 5. Kernel Density Estimates of Graders' Preferences by Gender

Appendix



Maastricht University

Please write **ONLY YOUR ID NUMBER** on the answer sheets.

DO NOT WRITE YOUR NAME on any of the answer or exam sheets, even if the exam sheets ask you to write your name.

Name: **LEAVE THIS FIELD EMPTY** Id. Nr: **i123456**

Thank you for your cooperation!

Figure A1. Yellow Sheet Placed on Some Students' Desks Before the Exam.