PRELIMINARY JANUARY 2013

ENDOPHILIA OR EXOPHOBIA: BEYOND DISCRIMINATION

Jan Feld, Nicolás Salamanca and Daniel S. Hamermesh*

*Feld and Salamanca: Ph.D. candidates in economics, Maastricht University; Hamermesh: Sue Killam Professor of Economics, University of Texas at Austin, professor of economics, Royal Holloway University 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. Hannah Ebin, Matthew Embrey, Christopher Parsons, Stephen Trejo and especially Leigh Linden provided very helpful comments.

ABSTRACT

The immense literature on discrimination treats outcomes as relativistic: One group suffers relative to another. But does a difference arise because agents discriminate against others—are exophobic—or because they favor their own kind—are endophiles? We conduct a field experiment in which graders at one university are randomly assigned students' exams that did or did not contain the students' names. Examining the effects of matches by nationality or gender on exam scores, on average we find favoritism but no discrimination by nationality, and neither favoritism nor discrimination by gender. Favoritism by nationality is due chiefly to behavior by the most experienced graders and those who had been rated as poor teachers in previous courses. We observe heterogeneity in both discrimination and favoritism by nationality and by gender in the distributions of graders' preferences.

> 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 disadvantaged. 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). Obviously if the only purpose of policy in this area is to equalize outcomes in a zero-sum environment with only two groups, this distinction does not matter. But if it is to reduce unfair treatment of one of several groups and/or if the environment is not zero-sum, determining whether we are observing discrimination, favoritism or both, and the extent of each, matters.

That favoritism and discrimination can lead to different outcomes becomes clearer when we examine a situation with more than two groups. Consider a supervisor who distributes a bonus payment of $300 among one Black, one White and one Hispanic worker. The norm is to give $100 to each worker, but a favoring (discriminating) supervisor will give $10 more (less) to a certain worker. A supervisor who favors the White worker will give him $10 extra, leaving $95 each for the Black and Hispanic workers (B = $95, W = $110, H = $95). A supervisor who discriminates against the Black worker gives him $10 less, leaving $105 each for the White and Hispanic workers (B = $90, W = $105, H = $105). In both cases the difference in outcomes between the White and the Black worker is $15, but the distributions resulting

under each alternative clearly differ. The importance of this difference is strengthened if we consider convexity in utility and the likely differences in the initial situations of the workers.

Once we recognize that favoritism need not be the opposite side of the same coin as 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 has been neglected. Only recently has there been even a small upwelling of interest in examining the behavior of discriminatory/favoring agents and their impacts on outcomes, typically considering how agents' behavior toward those who match them along some dimension differs from their behavior toward those who do not match.[1] Even most of these studies, however, have looked only at averages, and none has combined the analysis of the distribution of preferences (i.e., none has examined whether and to what extent some agents prefer their own—are *endophiles*—while some agents dislike others—are *exophobes*). One can even imagine that some agents prefer members of other groups—are *exophilic*, while other agents are *endophobic*—disfavor people like themselves.

Here we discuss the results of a field experiment that allows us to examine both this issue and to characterize the distributions of the tastes of the discriminating/favoring agents. 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. While this kind of "blindness" has been used before in the literature, it has been employed only to examine its effects on relative differences between groups.[2] We focus specifically on endophilia/exophobia by nationality and gender, but the method could be applied to any of the outcomes that have been studied in this immense literature. The crucial

[1]See Price and Wolfers (2010) and Parsons *et al* (2011) for evidence from professional sports; Dee (2005), Lavy (2008) and Hinnerich *et al* (2011) 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.

[2]Blindness has been used both in a field experiment (Blank, 1991) and non-experimentally (Goldin and Rouse, 2000), but only to measure net impacts on a particular group—not to distinguish favoritism from discrimination, since neither study linked the characteristics of potentially discriminatory agents' to those of the suppliers of the characteristics.

contribution—what makes this exercise possible—is that, instead of having only two groups, the construction of the experiment allows us to create a third group, the no-name or blind group, that is arguably neither discriminated against nor favored by any of the agents. This group provides a baseline against which we can measure endophilia/exophobia by the agents on the demand side.[3]

II. The Experimental Environment and Some Initial Impressions

A. The Environment

We conducted the experiment 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 this study. First, 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 in with Dutch and other nationalities. The student population is 36 percent female, and the academic staff is 28 percent female.[4] German students have a reputation for being more hard-working than Dutch and other students. These contrasts by nationality could potentially be the basis for discrimination/favoritism, although it is unclear *a priori* in which direction this will be.[5] Second, the grading of final exams, which we examine here, is a good setting for distinguishing the effects of tastes for discrimination from those of statistical discrimination, because graders do not gain anything from favoring or disfavoring specific groups. Third, 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

[3]The only studies 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.

[4]Statistics about the student population are taken from the SBE homepage (http://www.fdewb.unimaas.nl/miso/index.htm) and refer to 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."

[5]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 toward women (e.g., Donald and Hamermesh, 2006, although that study cannot distinguish between these two types of behavior).

and nationality.[6] Finally, and most important, this experiment has real-world consequences: The grades are important to students, and much of the graders' job revolves around their role in scoring exams.

B. Initial Pre-Experiment Impressions

Before we organized the experiment we had obtained data for earlier teaching periods (used by Feld and Salamanca, 2013). These data provide some initial impressions that allow us essentially to replicate the studies contained in the sparse prior literature that has examined how agents treat those who do or do not match their characteristics (see Footnote 1). Table 1 reports the results of regressions describing standardized exam scores on nearly 2000 answers written by nearly 400 students. The results show that graders on average scored students of the same nationality higher, but that on average they gave neither higher nor lower grades to students of the same gender. Disaggregating these results by nationality and gender, we can see that German graders gave significantly higher grades to German students.

Interpreting these estimates has several problems. The differential treatment that graders accorded certain students could be due to unobserved differences in ability. If the German students have higher ability on average, as their reputation suggests, this could explain the result that German graders gave higher relative grades to German students, and that Dutch graders gave lower relative grades to Dutch students. Also, we cannot determine whether and to what extent the differential treatment was a result of endophilia or exophobia—the outcome is merely relative. In spite of these problems, Table 1 suggests that nationality could play a role in grading, and these results are sufficiently encouraging to justify an experiment that examines graders' preferences in much more detail.

[6]In many cases the grader knows at least some of the students from tutorial meetings and lectures, and therefore can perfectly observe their nationality and gender. Furthermore, the grader can infer the nationality and gender of the students, even if he/she does not know them, 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 mixture in our sample (19 German, 17 Dutch and 14 other nationalities, of which 16 were female). The staff correctly identified the German names in 64 percent, the Dutch names in 65 percent and the gender in 90 percent of the cases.

III. Constructing the Experiment

To make the distinction between favoritism and discrimination we set up a field experiment that we carried out during the final exam week in June 2012. In the SBE written exams are given 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.[7] To prevent cheating the desk for each student 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 smallest to largest within each block, and then filling in sequentially first the columns and then the rows.[8] Figure 1 illustrates the arrangement of desks in each block.

The students in each session arrive at the exam hall and locate their assigned block based on the course they are taking. They then go to their block and 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). The sheets asked students on whose desk one was placed *not* to write their name but *only* their ID number on the exam sheets.[9] 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,

[7]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.

[8]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.

[9]We blocked entire rows instead of scattered seats within each 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 grade those exams as they usually would.

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, 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 wrote down the ID numbers of the students who did not write their names.[10]

At the SBE it is common practice to split the grading burden among various graders by letting each one handle different questions on the same exam. The course coordinators identified the grader of each question and provided us with information on the scores given. This information included the score on each question, the maximum possible points per question and other available grades that the student had attained in the course, including on course participation and any term paper.[11] 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.[12] 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 the students for that grader in all previous courses since the creation of the online tool. Overall we obtained complete information on 27 out of the 42 courses that had final exams, including 48 graders and 2463 exams that were scored.[13]

[10]This was done immediately after the exam, before the course coordinators received the exams to arrange the grading.

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

[12]We manually added the gender and nationality of the graders who did not fill out the survey.

[13]We excluded 7 courses which only used multiple-choice questions and another small course which did not have the exam in the conference hall. In 7 out of the 34 eligible courses the course coordinators either declined permission to use the data or did not respond to repeated requests for this information.

Table 2 examines the internal validity of the experiment, testing whether the questions in the treated (blind) group were answered by students who differed in measurable dimensions from those in the untreated (visible) group. We present these results separately for those students whom we intended to treat (ITT) and those who were actually treated.[14] We first examine differences by gender and nationality, the two ascriptive characteristics on which we focus, and in the students' achievements on other graded components of the courses. 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 women, Germans and Dutch, insignificantly different from each other; the absolute differences are never greater than one in the second decimal place between the visible and blind groups. Moreover, those whom we intended to treat but who were not treated do not differ significantly along these dimensions from those actually treated.

We have additional information on some of the students—other grades that were awarded before the exams were given, such as classroom participation, presentation in class and term-paper grades, and we also have grades from the multiple-choice and fill-in parts that were included in some of the exams. We find no significant differences between the blind and visible students in their participation grades, the only activity that was included in most courses. 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, presentation grades were given to less than one-third of the students. On the multiple-choice and fill-in questions the grading was unambiguous and should not be affected by the mechanisms we study here. The blind group has marginally higher scores on the multiple-choice

[14]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.

questions, but here too the differences are not quite statistically significant. These results confirm that the research design created equivalent groups of students.[15]

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 can index each answer by the pair (*s*, *g*).[16] 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 question was graded blind or visible, so that each pair (C(*s*),C(*g*)) can be expanded to the triplet (C(*s*),C(*g*),*b*), where *b*=1 if the grading is blind and 0 if not.[17]

Consider the score function S(C(*s*),C(*g*),*b*) 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 *b*. Define the following indicators:

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

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

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

and

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

The average score of students is:

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

[15]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.

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

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

where the weights θ_i are the shares of answers graded under each regime, and the (*) denotes an average over those answers.[18] 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) $M_M = [S^*(I1) - S^*(I2)]$;

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

(3b) $M_N = [S^*(I4) - S^*(I3)]$.

If graders are endophilic and exophobic, M_M , $M_N > 0$. In Section V we present estimates of each of the effects discussed here.

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 regardless of who observes it. Consider a variant of (2):

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

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 observed is the weighted average of the treatments when C(*g*)=C' and when C(*g*)≠C'. Thus:

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

Equation (5) shows that the average treatment effect of a characteristic will depend on two factors: 1) The degree of endophilia and exophobia (the expressions [S*(I1|C(*g*)=C') - S*(I2|C(*g*)=C')] and [S*(I4|C(*g*)=C') - S*(I3|C(*g*)=C')]); and 2) The share of questions that are graded by graders with matching characteristics ($\eta_1 + \eta_2$) versus non-matching characteristics ($1- \eta_1 - \eta_2$). In Section VI we

[18] 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 $2n^2$ average treatment effects, one for each of the n aspects of the characteristic compared to itself and each other aspect.

explore how these two factors contribute to the estimated effect of "being visible" as German or Dutch, female or male.

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 his/hers, others with characteristics that do not match. Then for each grader *g* we can calculate his/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) $m_M^g = S^{*g}(I1)-S^{*g}(I2)$;

and

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

where the differences $S^{*g}(Ij)-S^{*g}(Ik)$, j=1,4, k=2,3, are the averages over all students who are graded by grader *g*. Using these grader-specific average treatments we can then obtain the distributions of endophilia and exophobia as $f(m_M)$ and $h(m_N)$. 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' expressed preferences.

One special benefit that we obtain from our setting is that we can be sure that the implied preferences on matching that are found from (3a) and (3b) are being driven by taste-based discrimination and not by other common 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. This means, e.g., that the experiment allows us to compare "visible" to "blind" German students. The advantage of this approach becomes clearer if we recall Table 1, which suggests that in previous teaching periods both German and Dutch graders favored German students/discriminated against Dutch students. That pattern could also have been caused by ability differences between German and Dutch students, a difficulty that cannot affect the results of our experiment.

V. Basic Results and the Effects of Graders' Characteristics

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

(7) $S = \beta_0 + \beta_1 VISIBLE + \beta_2 MATCH + \beta_3 VISIBLE \cdot MATCH$,

where here S is a unit normal deviate calculated for each exam question, and the other variable names are self-explanatory. With this equation the estimates of endophilia and exophobia are: [19]

(8a) $M_M = \beta_1 + \beta_3$

and:

(8b) $M_N = -\beta_1$.

Note that these calculations mean that endophilia (exophobia) is indicated by a positive M_M (M_N). This framework makes it easy to expand Equation (7) to include interactions with some of the graders' characteristics and thus to examine how examine how M_M and M_N vary with them.

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 student within each course, we cluster the standard errors at the student-course level. We focus throughout on the estimates of M_M and M_N 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 will receive a score that is 0.17 standard deviations higher when his/her name is visible than when it is not. This addition to a matched student's grade is statistically significant at conventional levels. Moreover, it is economically important: Given that all the scores have been unit-normalized, it is equivalent to moving from the median score to the 57^{th} percentile of the distribution of scores. While favoritism by nationality exists in grading, there is no apparent exophobia by nationality:

[19] M_M is equal to the difference between the average grade for students with matching characteristics when the name is visible ($\beta_0 + \beta_1 + \beta_2 + \beta_3$) and when the name is not visible ($\beta_0 + \beta_2$). M_N is calculated analogously.

The estimated impact of being visible when not matching by nationality is small and positive. If anything there is evidence of exophilia.

The results for the regression on gender matching are shown in the second column of Table 3. None of the estimated parameters is statistically significant. More important, although the point estimate suggests the existence of endophilia, we cannot reject the hypothesis that it is zero. For non-matches there is again exophilia, but here too the impact is statistically insignificant (and also minute). On average grading is gender-neutral in all dimensions.

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 endophilia or exophobia exists for both male and female graders. We do this by expanding Equation (7) to include the main effects of student nationality or gender and their interactions with *VISIBLE*, *MATCH*, and *VISIBLE·MATCH*. Columns (3) of Table 3 present estimates of Equation (7) when we add these terms. A comparison of the results suggests that endophilia by nationality arises more from the behavior of Dutch than of German graders. While the point estimate for German graders does suggest that they favor German students (by about 1/7 of a standard deviation in the score), the effect is not statistically significant. Dutch graders, on the other hand, advantage Dutch students by about 1/5 of a standard deviation, an advantage that is statistically significant.

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

It is interesting to contrast these results to those in Table 1. They confirm the notion that nationality matters for grading while gender does not. Table 3, however, shows two features that Table 1 could not. First, it demonstrates that the effect of matching on nationality is due to endophilia, not exophobia. Second, Table 3 shows that this effect of matching is mainly due to the behavior of Dutch

graders, whereas Table 1 suggested it was mainly due to German graders. Since the results of Table 3 not only allow us to separate endophilia from exophobia but also to control for unobserved heterogeneity, we can conclude that both German and Dutch graders exhibit endophilia and that the results from Table 1 were driven by unobserved differences in students' ability.

The graders differ along several dimensions on which we have information and that might affect their ability or interest in favoring/discriminating for or against students. The first is grader experience at this University—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.

In our data 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 5^{th}, 50^{th} and 95^{th} centiles, for which we present estimation results, are 1, 8 and 59 courses.[20] Figure 2a presents the kernel density of courses taught by grader, which makes it clear that the distribution has a very long right tail. The second and third columns of Table 4 present re-estimates of Equation (7), expanded to include a main effect in grader experience and that variable interacted with *VISIBLE*, *MATCH* and *VISIBLE·MATCH*. While the point estimate of the extent of endophilia by nationality is almost identical at the median value of grader experience to the estimate in Table 3, it is not quite significantly nonzero. Rather, the significant average endophilia shown in Table 3 results from the behavior of the more experienced graders. By inference, they feel less inhibited about indulging their preferences for students who match their nationality. Obversely, 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

[20] 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.

of highly experienced graders). As with the basic estimates, there is no evidence of exophobia by nationality at any level of grader experience.

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 by gender. Whether inexperienced or experienced, graders on average appear to ignore students' gender when scoring exams.

Another characteristic on which we have data on graders is the average of all the instructor's evaluations received from students during his/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 5^{th} percentile being 7.1, the median being 8.0, and the 95^{th} percentile equaling 8.8. As Figure 2b shows, while the distribution of average evaluations is not perfectly symmetric, it is not far from that.

We interact the grader's average instructional evaluation with all the variables in Equation (7) and present the results in Columns (4) and (5) of Table 4. As with all the other results in this Section, there is no evidence of endophilia or exophobia by gender, regardless of how badly or well the grader's teaching is rated by students. There is also no evidence of exophobia by nationality. Our finding of endophilia by nationality at the mean that we demonstrated in Table 3 arises 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 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—significantly and substantially favoring those who match his/her nationality, an unsurprising result since the teaching evaluations are distributed nearly symmetrically around their mean. 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 his/her nationality from the mean to the 61^{st} percentile of the distribution of scores. In sum, worse teachers behave differently from better ones, favoring students along the dimension of their nationality.

VI. The Average Treatment Effect of Visibility of a Student Characteristic

To evaluate whether the visibility of names under the current grading procedure 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 is the sum of endophilia and exophobia weighted by the share of questions that was graded by graders with matching and non-matching characteristics. Table 5 shows the ATE of being seen as German, Dutch, or any other nationality, and of being seen as female or male. Even though none of the point estimates is significantly different from zero, they suggest that on average Dutch and German students benefit from visible grading, while students of other nationalities are disadvantaged by it. The difference between Germans and others is marginally significant (p=0.062) as is the difference between Dutch and others (p= 0.076). Consistent with our previous results, the point estimates for females and males are positive but smaller in size.

Columns (1) to (4) of Table 5 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 effects of endophilia and exophobia are taken from Table 3.)

The ATE for German and Dutch students is small because of the relatively small share of questions that are graded by graders of the same nationality. This outcome results from the heterogeneous mix of nationalities of 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.

VII. Robustness and Responses to Changes in the Prices of Favoritism and Discrimination

Some of those whom we intended to treat were not actually treated, as we saw in Table 2. To account for potential problems induced by this discrepancy, we re-estimated the models described in the first two columns of Table 3 using intention to treat as an instrument for *VISIBLE* (actually, for *1 – VISIBLE*). The results are very similar to those depicted there. For matches by nationality the average extent of endophilia is 0.197 standard deviations (p = 0.028), nearly identical to that estimated without

instrumenting; and the estimated extent of exophobia is -0.024 (p=0.679), tiny and not statistically significant, as was the estimate in Table 3 (showing, as in Table 3, that there is weak evidence of exophilia). For matches by gender the measured extent of endophilia is 0.090 standard deviations (p = 0.175), and the extent of exophobia is -0.039 (p = 0.579), both quite close to the estimates in Table 3. The slight leakage between the intention to treat and actual treatment did not bias the basic results.[21]

Another concern is that some of the graders looked up students' names before grading the exams. Because our identification relies on the distinction between blind and visible grading, graders who have looked up names will dilute our findings. To address this concern we take advantage of the post-experiment survey, in which graders were asked, "Before grading each question, did you look up the names of the students?" Six out of the 32 graders who responded to the survey indicated that they did so.

Columns (1) and (2) of Table 6 show re-estimates of Equation (7) adding a main effect for an indicator that the grader looked-up some or all of the names, and interactions of that indicator with the three independent variables. The results for nationality are as expected: The estimated endophilia of graders who did not look up any names exceeds the average, while that of graders who did look up names is small and insignificant. There is also no evidence of exophobia by nationality or endophilia by gender, whether or not the grader looked up any names. We do, however, find marginally significant exophilia by graders who did not look up any names. This suggests that there may be exophilia in our sample, but that we may not be able to discern its presence because of some dilution of the treatment due to graders looking up the names of students.

Another concern is that some graders may have been aware that an experiment was being conducted, which could bias our inferences. Graders who realized that they were being "monitored" might be less likely to exhibit endophilia or exophobia, since such behavior might seem costly to them. As in Parsons *et al* (2011), the grader may have been unwilling to indulge his/her tastes, fearing that unfair

[21]The results are essentially the same when we include additional controls for seat number (see Figure 1) and participation grade.

grading might reduce the chances of continued employment (for student graders) or lead to professional disrepute (for faculty graders).

To address this concern we created an indicator of graders' awareness of being monitored and interacted it with the independent variables in Equation (7).[22] The results, contained in the final two columns of Table 6, look remarkably like those in the first two columns: Endophilia by nationality is only detectable among those graders who did not suspect that an experiment was being conducted. There is no sign of exophobia by nationality in either group of graders, and there is no evidence for endophilia by gender. In line with the findings for the graders who did not look up any names, we find marginally significant evidence for exophilia by gender.[23]

VIII. Heterogeneity in the Distribution of Preferences

The results thus far describe either the average responses over all graders of endophilia or exophobia by nationality or gender, or examine how this behavior differs in relation to a few of the graders' specific characteristics. In this section we first consider the shapes of the entire distributions of graders' preferences and discuss the implications of our findings for the original Becker model of taste-based discrimination. We then consider the importance of the distribution of the graders' characteristics C(g) for determining the ATE of a student characteristic C(s) being observed.

To obtain a feel for why examining heterogeneity in preferences might be interesting, consider the kernel density estimates of endophilia and exophobia by nationality, shown in Figure 3, and their kernel density estimates by gender, shown in Figure 4. Each kernel is based on those graders for whom we could infer the extent of both endophilia and exophobia (for nationality, 21 graders, for gender, 30 graders).[24]

[22]We constructed this measure by manually marking the graders who we believe suspected an experiment. This belief is based on what we heard during the implementation of the experiment.

[23]Specifications that include main effects and interactions with both the "look-up" and "suspected" variables yield conclusions very similar to those of the specifications based on each separately.

[24]We derive the shape of the graders' preferences based on the estimates of $m_M{}^g$ and $m_N{}^g$. We infer these two measures for each grader based on how each scores students who do or do not match them under the blind and visible regimes.

The estimates along the criterion of nationality suggest that preferences are distributed fairly 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. The same conclusion is suggested by the kernel of exophobia by gender. The kernel of endophilia by gender is completely different. While the estimates suggest endophilia by the median grader, a few graders are apparently highly endophobic. This asymmetry generated the estimated absence of endophilia by gender that we showed in Table 3, but that inference at the means hides a substantial skewness in preferences.

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 $m_M^g=0$ ($m_N^g=0$) for all g, and whether the m_M^g (m_N^g) are equal to each other for all g, respectively. The F-tests of these hypotheses (eight in total) always reject the null hypothesis at all conventional significance levels, showing that endophilia and exophobia in both nationality and gender are real phenomena (even though at the mean only endophilia by nationality seems to matter), and that there is significant heterogeneity in these preferences across graders.

Observing the entire distribution of preferences also has implications for the taste-based model of discrimination, which states that wage differentials are driven by the discriminatory preferences of the marginal employer and not by the average level of discrimination in the population.[25] In his model Becker implicitly assumes that employers only have preferences against one group and are indifferent to the other group. Goldberg (1982) extended this model to include favoritism, but he assumed that employers are indifferent to the group they are not favoring. We have shown that graders have endophilic and exophobic preferences for both major nationalities and hence are not indifferent to either group. This additional complexity makes the identification of the marginal discriminator less straightforward than in Becker's case. Here the identity of the marginal discriminator will differ depending on whether market power lies

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

on the demand side (employers\graders) or the supply side (employees\students). If market power lies on the demand side, employers will choose employees based on their *relative* preferences for one group over the other. If market power lies on the supply side, the employees will choose their employer based on the employer's *absolute* preference for their characteristic. Each sorting mechanism will result in a different marginal discriminator.

The intuition behind this distinction is simple. In our example, if students can choose their grader (i.e., there is supply-side market power), they will prefer the grader with the strongest preference for their nationality, regardless of the grader's own nationality or preference for other nationalities. If, on the other hand, graders can choose the students whom they judge (i.e., there is demand-side market power), they will choose students of the nationality that they like most (or dislike least). In each of these cases the marginal discriminator will likely end up being a different grader. Substituting employers for graders and worker for students, the discussion carries over to the labor market.

To illustrate the importance of heterogeneity in grader characteristics C(g), consider a hypothetical change in which 90 percent of all exam questions for both German and Dutch students were graded by Dutch graders and only 10 percent by German graders (i.e., $[\eta_1 + \eta_2] = 0.1$ for German, but $[\eta_1 + \eta_2] = 0.9$ for Dutch graders). Assuming that the distribution of graders' tastes remained unchanged, and using the estimated effects in Table 5, the ATE of being visible and German would have been 0.092 standard deviations, whereas the ATE of being visible and Dutch would have been 0.178 standard deviations. This change in the composition of graders by nationality has only a tiny effect on outcomes for German students, but a more substantial effect (+0.07 standard deviations) on those for Dutch students. More generally, this calculation shows, as the original theory of discrimination suggests and recent work (Charles and Guryan, 2008) demonstrates, that the distribution of agents' characteristics will affect the differentials observed in the market.

Finally, we can examine whether extreme values in the distributions of preferences are driving our mean effects. We tackle this problem by trimming those graders with the most extreme preferences from the samples. We do this for both nationality and gender, dropping the two most extremely

endophilic/endophobic and exophobic/exophilic graders in each case. Despite the asymmetry of the distribution of m_N in Figure 4, even there trimming does not qualitatively alter the conclusions about the absence of endophilia or exophobia by gender on average. In the other three cases too the conclusions are not greatly modified by this trimming exercise. Removing outliers does not change the results that we observe significant endophilia on nationality with no other apparent distinctions on average.

IX. Conclusions and Inferences

We have demonstrated that what would be 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 is possible in real-world situations to measure the sizes of these two components of the net amount of discrimination. 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—indeed, the "other" is treated slightly better (exophilia) when the other is identifiable. We find no evidence of differential treatment by gender on average, whether or not the discriminating agents match their subjects.

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 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 their characteristics. Not surprisingly, a neutral outcome can also be achieved in a variety of ways.

We noted in the Introduction that the distributions of welfare in otherwise identical situations characterized by favoritism toward a majority versus discrimination against a minority can differ. Having shown that we can distinguish endophilia from exophobia, it is worth considering how policy might be tailored to reduce relative differences arising from prejudice. Assuming that our results carry over to the labor and other markets, and that endophilia is the main source of apparently discriminatory outcomes, for example, moral suasion that stresses to majority-group members 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? One possibility is a carefully constructed audit study. While we have some doubts about the burden such studies impose on unwitting participants, perhaps the importance of measuring endophilia and exophobia in this context might outweigh those doubts. Going still further toward a general real-world context, 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. The main point is that these preferences generate different outcomes with different distributions of welfare, so that determining their relative size is economically important and, as we have shown, possible.

References

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.

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.

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.

Rebecca Blank, "The Effects of Double-Blind versus Single-Blind Refereeing: Experimental Evidence from the *American Economic Review," American Economic Review,* 81 (Dec. 1991): 1041-67.

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.

Jan Feld and Nicolás Salamanca, "Grading Expectations," Unpublished paper, SBE, Maastricht University, 2013.

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.

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.* 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.

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.

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.

Table 1. Nationality and Gender Matching, Previous Teaching Periods*

	(1)	(2)	(3)			(4)	
	Nationality	Gender	Nationality			Gender	
Interaction with:	-	-	*German*	*Dutch (base)*	*Other*	*Female*	*Male (base)*
CONSTANT	-0.035	-0.015	0.495	-0.258	-0.006	0.177	-0.088
	(0.042)	(0.046)	(0.085)	(0.063)	(0.104)	(0.091)	(0.063)
MATCH	0.156	0.034	-0.128	0.113	-	0.006	0.019
	(0.062)	(0.050)	(0.116)	(0.092)	-	(0.098)	(0.068)
MEAN GRADE IF MATCH	0.122	0.018	0.223	-0.144	-	0.113	-0.069
p =	[0.048]	[0.690]	[0.003]	[0.148]	-	[0.101]	[0.277]
Observations	1,993	1,993	1,993			1,993	
Adj. R^2	0.004	-0.001	0.057			0.007	

*Standard errors in parentheses and p-values in square brackets. Both are clustered by student-course. (1) and (2) are based on regressions of S on MATCH; (3) and (4) are based on a regression of S on MATCH, CHARACTERISTIC dummies, and their interactions, where CHARACTERISTIC are dummies for German and Other in (3) and for Male in (4). MATCH*Other interactions in (3) are empty because we define MATCH = 1 only for German and Dutch students. Because the student and grader population was dominated by German and Dutch, other nationalities almost never matched.

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

		(1) Blind			(2) Visible			
		Mean	*SD*	*N*	*Mean*	*SD*	*N*	*p-value of difference*
Female	ITT	0.367	0.482	732	0.375	0.484	1731	[0.722]
	Treatment	0.374	0.484	653	0.372	0.484	1810	[0.953]
German	ITT	0.485	0.500	732	0.487	0.500	1731	[0.925]
	Treatment	0.490	0.500	653	0.485	0.500	1810	[0.827]
Dutch	ITT	0.301	0.459	732	0.278	0.448	1731	[0.246]
	Treatment	0.294	0.456	653	0.281	0.450	1810	[0.528]
Participation	ITT	7.489	1.263	585	7.387	1.590	1391	[0.128]
	Treatment	7.451	1.279	516	7.406	1.572	1460	[0.512]
Presentation	ITT	7.794	1.161	192	7.930	1.057	438	[0.170]
	Treatment	7.757	1.169	182	7.942	1.053	448	[0.066]
Term paper	ITT	7.867	0.663	110	7.744	0.897	282	[0.137]
	Treatment	7.866	0.694	98	7.749	0.881	294	[0.179]
Multiple Choice exams	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 exams	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]

*p-value of differences calculated with clustered standard errors by student.

Table 3. Basic Estimates of the Extent of Favoritism and Discrimination by Nationality and Gender*

	(1) Nationality	(2) Gender	(3) Nationality			(4) Gender	
Interaction with:	-	-	*German*	*Dutch (base)*	*Other*	*Female*	*Male (base)*
CONSTANT	-0.022	-0.018	0.151	-0.092	0.057	0.125	-0.071
	(0.041)	(0.054)	(0.108)	(0.089)	(0.110)	(0.105)	(0.080)
MATCH	-0.041	-0.027	0.225	-0.119	-	0.077	-0.029
	(0.080)	(0.067)	(0.171)	(0.106)		(0.137)	(0.092)
VISIBLE	0.005	0.025	0.041	0.047	-0.135	0.127	-0.030
	(0.050)	(0.064)	(0.127)	(0.105)	(0.131)	(0.124)	(0.095)
MATCH x VISIBLE	0.163	0.035	-0.093	0.146	-	-0.133	0.092
	(0.092)	(0.080)	(0.196)	(0.128)		(0.166)	(0.109)
Endophilia	0.168	0.060	0.140	0.193	-	0.055	0.062
p =	[0.037]	[0.269]	[0.283]	[0.030]	-	[0.599]	[0.318]
Exophobia	-0.005	-0.025	-0.087	-0.047	-	-0.097	0.030
p =	[0.921]	[0.692]	[0.224]	[0.655]	-	[0.217]	[0.755]
N	9,330	9,330		9,330		9,330	
Adj. R^2	0.002	0.001		0.017		0.010	

*Standard errors in parenthesis and p-values in square brackets. Both are clustered by student-course. Columns (1) and (2) are based on regressions of S on MATCH, VISIBLE, and their interaction. (3) and (4) are based on regressions of S on MATCH, VISIBLE, CHARACTERISTIC dummies, and their interactions, where CHARACTERISTIC are dummies for German and Other in (3) and for Female in (4). MATCH x Other interactions in (3) are empty because we define MATCH = 1 only for German and Dutch students. Other nationalities almost never matched.

Table 4. Effects of Grader Experience and Grader Teaching Quality on Outcomes*

At the m^{th} percentile of:	Percentile:	(1) Nationality	(2) Gender	(3) Nationality	(4) Gender
		Experience		*Teacher Quality*	
Endophilia	5^{th}	0.152	0.077	0.378	0.021
	p =	[0.185]	[0.274]	[0.075]	[0.896]
	50^{th}	0.163	0.074	0.168	0.072
	p =	[0.109]	[0.234]	[0.045]	[0.185]
	95^{th}	0.247	0.048	0.056	0.100
	p =	[0.018]	[0.629]	[0.708]	[0.247]
Exophobia	5^{th}	-0.022	0.005	-0.018	-0.253
	p =	[0.721]	[0.947]	[0.891]	[0.124]
	50^{th}	-0.014	-0.009	-0.005	-0.013
	p =	[0.791]	[0.896]	[0.910]	[0.838]
	95^{th}	0.043	-0.117	0.001	0.114
	p =	[0.679]	[0.303]	[0.987]	[0.310]

*p-values clustered at the student-course level in square brackets. We report linear combinations based on regressing S on MATCH, VISIBLE, EXPERIENCE or TEACHERQUALITY, and their interactions, and we evaluate the linear combinations at different values of EXPERIENCE and TEACHERQUALITY.

Table 5. The Average Treatment Effect (ATE) of the Visibility of Student Characteristics*

	Total ATE	*p-value*	(1) Share matched $(\eta_1+\eta_2)$	(2) Endophilia	(3) Share not matched $(1-\eta_1-\eta_2)$	(4) Exophobia
German	0.103	[0.109]	0.29	0.140	0.71	-0.087
Dutch	0.107	[0.155]	0.41	0.193	0.59	-0.047
Other	-0.088	[0.268]	-	-	-	-
Female	0.078	[0.246]	0.45	0.055	0.55	-0.097
Male	0.028	[0.618]	0.62	0.062	0.38	0.030

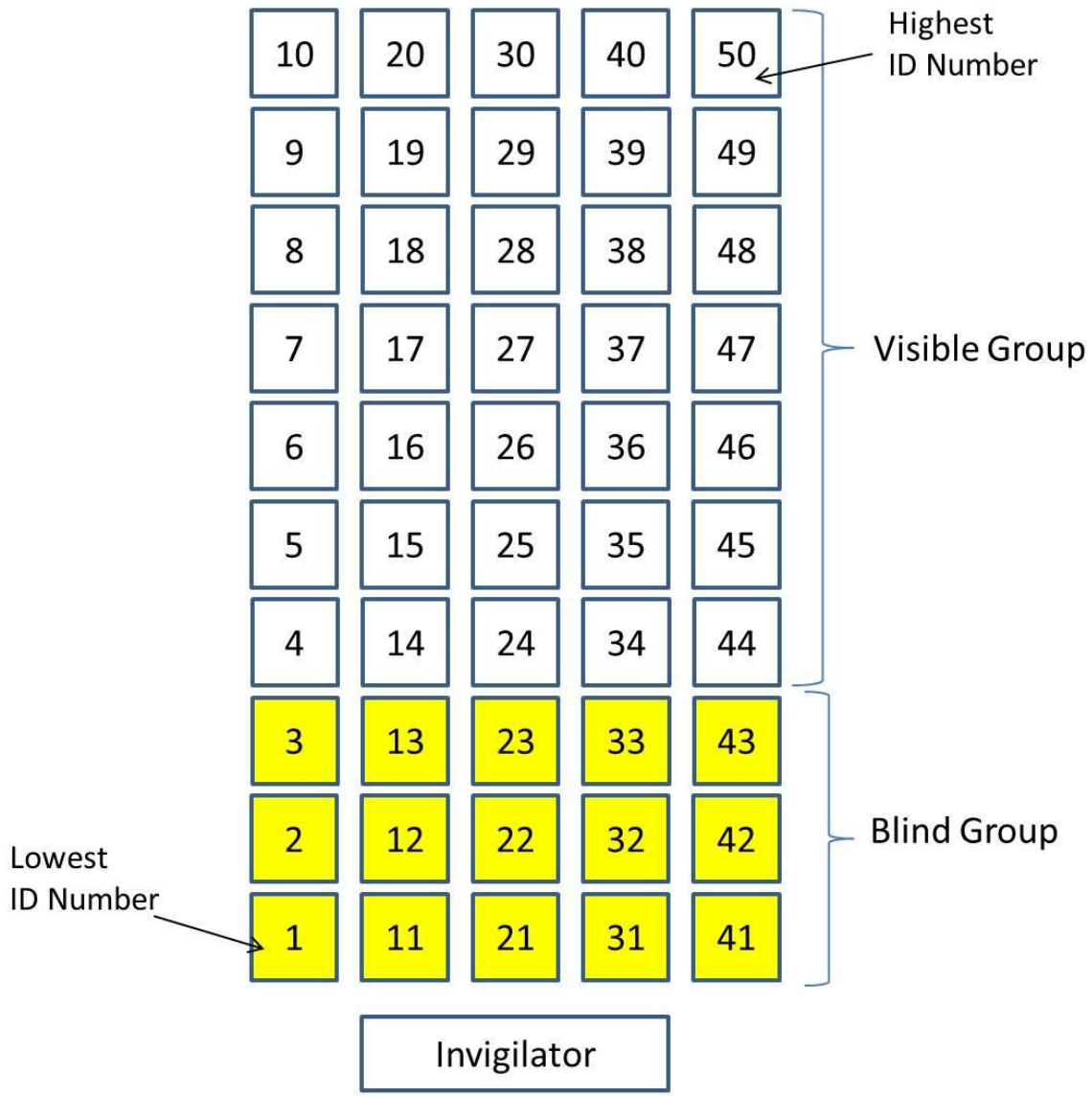
*The ATE is calculated as shown in Equation (5). The p-values are based on standard errors clustered at the student-course level. Columns (1) and (3) show the share of questions, for a given characteristic, which were graded by graders with matching and non-matching characteristics. Columns (2) and (4)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

Table 6. Robustness Checks on Grader Information and Behavior*

		(1) Nationality	(2) Gender		(3) Nationality	(4) Gender
	Difficulty?	*Looked up names*			*Grader suspected*	
Endophilia	No	0.194	0.004		0.182	0.039
	p =	[0.029]	[0.951]		[0.050]	[0.505]
	Yes	-0.064	0.187		-0.004	0.082
	p =	[0.770]	[0.290]		[0.986]	[0.640]
Exophobia	No	-0.013	-0.122		-0.048	-0.131
	p =	[0.828]	[0.092]		[0.369]	[0.063]
	Yes	0.101	0.138		0.194	0.189
	p =	[0.402]	[0.362]		[0.121]	[0.207]
N		7,431	7,431		8,138	8,138

*p-values clustered at the student-course level in square brackets. We report linear combinations based on regressing S on MATCH, VISIBLE, LOOKUP or SUSPECTED, and their interactions. We control for grader experience, grader instructor evaluation, and grader role in the course (course coordinator, course planner, tutor, lecturer, or other) in all regressions. Fewer observations than in Table 3 are used because of survey non-response.

Figure 1: Seating Arrangement for the Experiment[a]



[a]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.

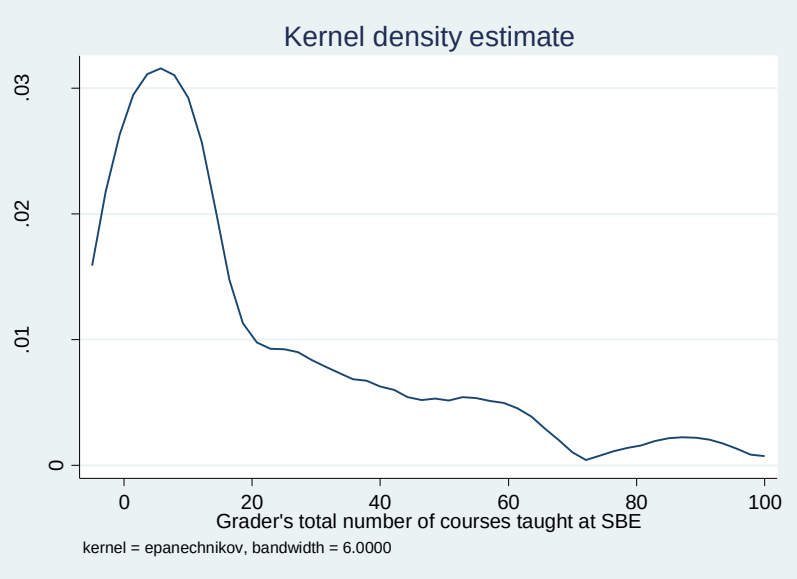


Figure 2a. Kernel Density of the Distribution of Grader Experience

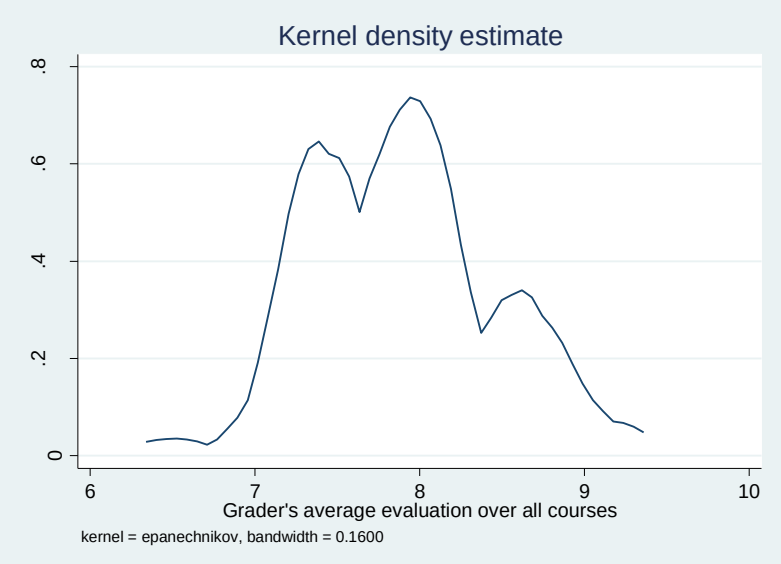


Figure 2b. Kernel Density of the Distribution of Student Evaluations of Graders

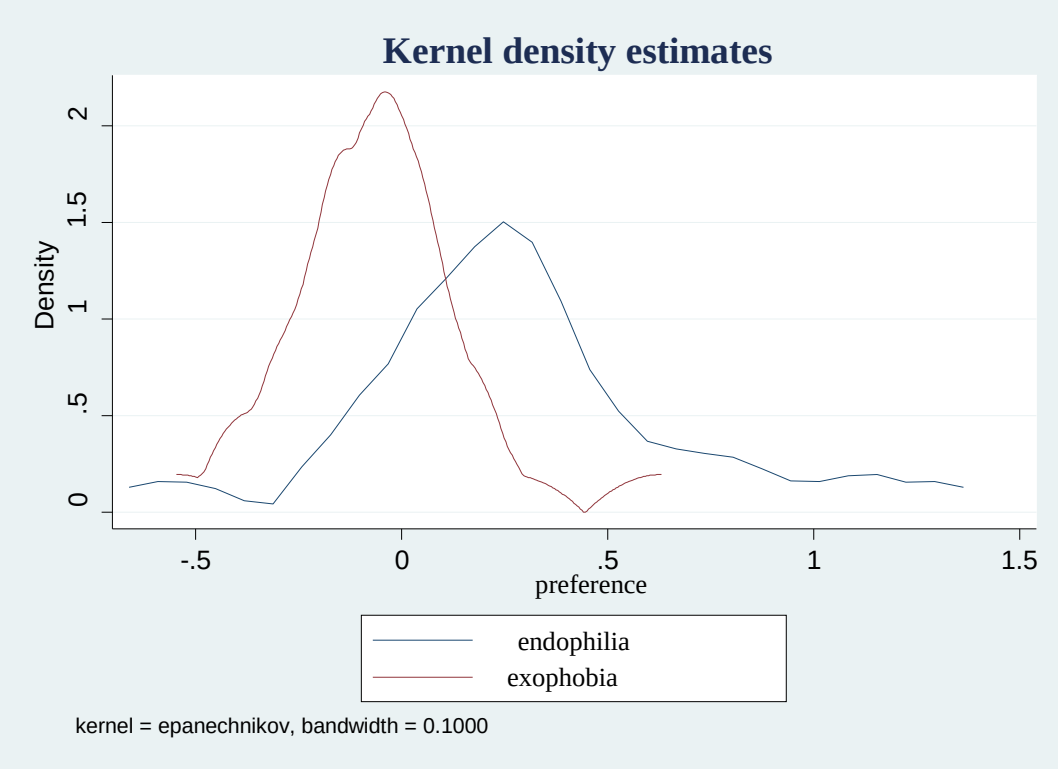


Figure 3. Kernel Density Estimates of Graders' Preferences by Nationality

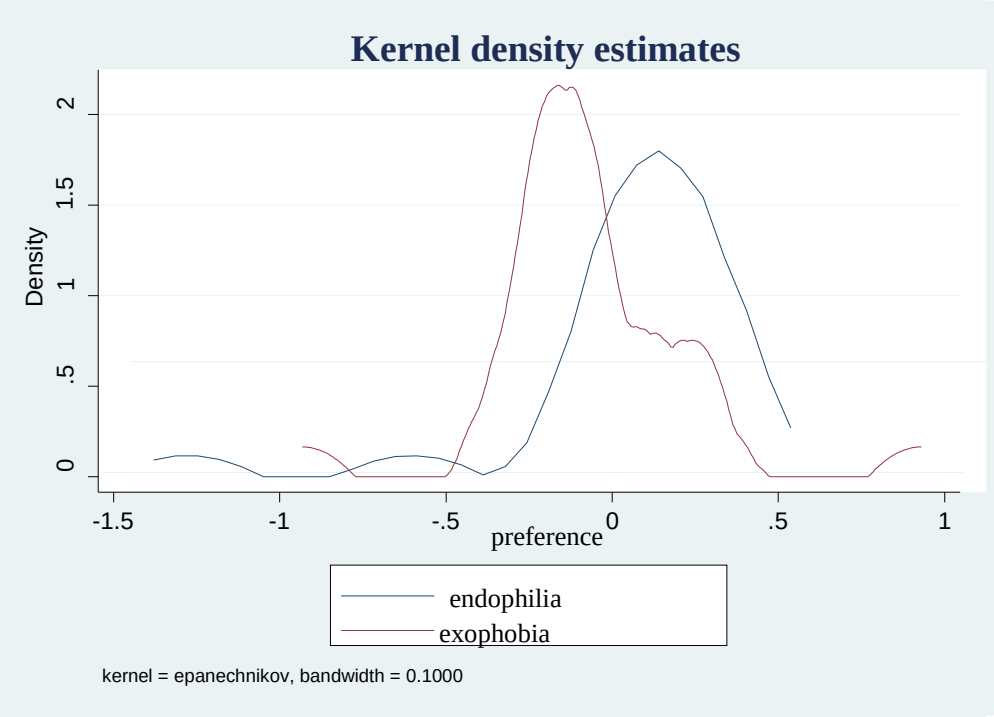


Figure 4. Kernel Density Estimates of Graders' Preferences by Gender