FORTHCOMING, *THE ECONOMIC JOURNAL*
FIRST VERSION: NOVEMBER 2012

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 favour their own kind—endophilia? Using a field experiment that assigned graders randomly to students' exams that did/ did not contain names, we find favouritism but no discrimination by nationality, but neither by gender. We are able to identify these preferences under a wide range of behavioural scenarios regarding the graders. That endophilia dominates exophobia alters how we should measure discriminatory wage differentials and should inform the formulation of anti-discrimination policy.

JEL classification: J71, I24, B40

Keywords: favouritism, discrimination, field experiment, economics of education

*Feld: Gothenburg University; Salamanca: Melbourne Institute of Applied Economic and Social Research (MIAESR); Hamermesh: Professor of 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, Adam Booij, 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, several referees, 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]

1. Introduction

Economists have studied labour-market discrimination at least since Becker (1957). Differences in labour-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 favoured while the minority group's outcome is the norm.

The possibility that we are measuring the extent of favouritism 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 by economists in the past quarter century, including by the more recent *Handbook* surveys of the literature on discrimination (Altonji and Blank, 1999; Fryer, 2011). Once we recognise that favouritism need not be the obverse of discrimination, the importance of studying preferences for favouritism/discrimination increases. Although the distribution of discriminating agents' tastes underlay Becker's theory, in most empirical research the demand side—the behaviour of discriminatory agents—has not been studied explicitly. Only recently has there been a small upwelling of interest in examining their behaviour and its impacts on outcomes.[1] These studies typically consider how agents' behaviour toward those who match them along some dimension differs from their behaviour toward those who do not match them, again only estimating relative differences.

[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 favouritism and discrimination separately under reasonable assumptions about agents' rationality. The key to doing this is 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 favouritism/discrimination.[3] Hence, we have evidence for favouritism 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 favouritism/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 favouritism to discrimination. Hence, *endophilia* denotes preferences for member of one's own group, while *exophobia*

[2]In the immense literature on discrimination the majority of studies focus on discriminatory penalties to out-groups. More recently some others have looked at in-group favoritism in various markets (e.g., Laband and Piette, 1994; Garicano e*t al*., 2005; Bernhard *et al*, 2006). Unless these differences are assessed versus a neutral group (i.e., a group that receives neither a penalty nor a premium), any distinction made between favoritism and discrimination is purely semantic. 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.

denotes preferences against members of other groups. One can also imagine, however, that some agents prefer members of other groups—are *exophilic*, while other agents are *endophobic*—discriminate against people like themselves.

2. Motivation

The importance of the distinction between favouritism and discrimination can be seen both theoretically and empirically. Goldberg (1982) adapted Becker's model to show that if favouritism 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 favouring *instead of* the discriminatory preferences as in Becker (1957). Employers can, however, have both discriminatory *and* favouring preferences; and a merging of both models would show how variations in the importance of these different preferences can generate variations in observed "discriminatory" wage differentials.

That the concepts of endophilia and exophobia are different is reflected by survey evidence. 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 favour members of their own group more disfavour 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 that between 1996-2000 and 2004-2006 Whites' endophilia remained constant, while Whites' exophobia (the negative of the measure in the Table) decreased. Becker's model (where only exophobia matters) would predict that the decline in exophobia shown in the Table decreased the wage gap. Goldberg's model (where only endophilia matters) would predict that the wage gap remained constant. That the black-white earnings ratio in the U.S. remained constant over this period hints that favouritism in the labour market may be more important than discrimination, although with so many other shocks over this short period attributing causation is difficult. The main point here is that favouritism and discrimination evolve over time as distinct constructs, and thus to understand the true nature of discriminatory outcomes it is important to acknowledge and carefully consider them as such.

3. Constructing the Experiment

3.1 The Environment

To make the distinction between favouritism 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 favouritism 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.[4] German students have a reputation for being more hard-working than Dutch and other

[4]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/favouritism, although it is unclear *a priori* in which direction these will be.[5]

The grading of final exams, which we examine here, is a good setting for identifying discrimination/favouritism, because graders have no incentives to favour or disfavour 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.[6] 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.[7] To prevent cheating, the location of each student's desk is predetermined by the Exams Office (the organisation 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.[8] Figure 1 illustrates the arrangement of desks in each block.

---

[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 (e.g., Donald and Hamermesh, 2006, although that study cannot distinguish between these two types of preferences).

[6]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*.

[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 centralised 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.

3.2 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 balanced 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).[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, so that the grader would only observe their ID numbers when grading their exams. 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.[10]

---

[9]We placed the sheets on entire rows instead of scattered seats within each block for simplicity. The Exams Office informed the course coordinators—who were in charge of organising 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.

[10]This was done immediately after the exam, before the course coordinators received the exams and started the grading process. The blind treatment group had a little over 80-percent compliance, 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.

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 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.[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 behaviour 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 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.[13]

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.[14] 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

[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. Grading behavior includes whether graders looked up any names while grading.

[13]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.

[14]Because actual treatment can be endogenous, due to students mistakenly writing their names when instructed not to, all our analyses are made using ITT as the actual treatment. Our results change very little if we use the actual treatment instrumenting it by the ITT.

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 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 any of these characteristics. 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, there are no differences between the Blind and the Visible groups in this respect either.

4. Empirical Approach and Basic Results

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).[15] 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 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.[16]

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

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

[16]Presumably all particular (s, g) combinations are either blind or visible (although we investigate the extent of blindness in the blind grading in Section 5).

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 $C(s)\neq C(g)$ and $v=1$, 0 if not;

and

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

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:

(2a) $e^* = [S^*(I1) - S^*(I2)]$;

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

(2b) $x^* = [S^*(I4) - S^*(I3)]$,

where $S^*(Ij)$ is the average score for the group $Ij=1$. 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 4 we present estimates of each of the effects as discussed here. We discuss the implications of alternative behavioural assumptions in Section 5.

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:

(3) $S = \beta_1 MATCH^*VISIBLE + \beta_2 MATCH^*BLIND + \beta_3 NON\text{-}MATCH^*VISIBLE$

$+ \beta_4 NON\text{-}MATCH^*BLIND + \gamma' Z + \varepsilon$,

where here S is a unit normal deviate calculated for each exam question, and the other variable names are self-explanatory where *BLIND* and *VISIBLE* are based on ITT.[17] 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

[17]The distribution of the standardised 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*.

without a constant. From this equation the estimates of the average extent of endophilia and exophobia are:

(4a) $e^* = S^*(I1) - S^*(I2) = \beta_1 - \beta_2$,

and:

(4b) $x^* = S^*(I4) - S^*(I3) = \beta_4 - \beta_3$.

Thus the estimates of (3) provide direct analogues 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 out in this comparison. This framework also makes it easy to expand Equation (3) 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 6.

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 randomised by blocks of students within each course, we cluster the standard errors at the ITT and 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 0.165 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-normalised, 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 favouritism by nationality exists in grading, there is no apparent exophobia by nationality: The estimated impact of being visible when not matching by nationality is very small, but positive—if anything there is evidence for exophilia.

The results of estimating the regression examining gender matching are shown in Column (2) of Table 3. The point estimate suggests the existence of a small degree of endophilia. For non-matches there is exophilia, but the impact is statistically insignificant and also minute. On average gender matching seems to be of little importance for grading.[18]

Going behind the information in Columns (1) and (2), we can ask whether, for example, endophilia by nationality is the same for Dutch and German graders, and whether endophilia and exophobia are absent among both male and female graders. We do this by expanding Equation (3) to include interactions of student's 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 behaviour of Dutch than of German graders; it also shows significant exophilia by German graders, although the differences by nationality between the estimates are not statistically significant. The larger point estimate for endophilia by Dutch compared to German graders (and to some extent of exophilia by German graders) is interesting for answering the question whether our results are driven by statistical discrimination. At the SBE Dutch students have a reputation of being less hard-working, and they also receive significantly lower grades than German students. The high estimate of Dutch endophilia is therefore inconsistent with a form of statistical discrimination in which the grader uses the nationality as a signal of student ability.

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

Columns (4) of Table 3 show estimates of expanding Equation (3) by gender. The results reveal that the small endophilia by gender in Column (2) is mostly driven by the behaviour of male graders, although here too the differences are not statistically significant. Neither male nor female graders exhibit significant exophobia, and for both men and women the absolute impacts are smaller compared to the impacts based on nationality.

5. Identification and the Treatment of the Blind Group

The interpretation of our results above as direct estimates of graders' endophilic and exophobic preferences hinges on the assumption that graders are neutral toward the exams in the blind group. In this section we first argue that this is the case, since we are likely observing a form of *implicit* discrimination—discrimination driven by attitudes outside of the awareness of the discriminator. We then consider two alternatives of how the blind group might be treated by graders: 1) What if graders recognise the characteristics of some of the students in the blind group? And what if graders penalise students in the blind group for not writing their names? We show the direction in which these two different ways of treating the blind group could alter our results. 2) We then consider the case in which graders have rational expectations regarding the composition of the blind group. We show that where graders are perfectly rational about the blind group, we can correctly identify only the sum of endophilia and exophobia, i.e., what has traditionally been called "discrimination," although we can no longer disentangle the two. In cases where graders are less than perfectly Bayesian about their expectations, however, we demonstrate how to recover the separate estimates of endophilia and exophobia.

5.1 Implicit Discrimination

Becker's model of discrimination and the models that follow it implicitly assume that agents consciously act on their discriminatory preferences. Bertrand *et al*. (2005) argue that discriminatory outcomes can also be driven by what they call implicit discrimination: behaviour driven by "*unconscious* mental associations between a target […] and a given attribute." (p. 94.) Their claim is based on an extensive body of experimental evidence in social psychology which, when incorporated into the discrimination literature, suggests that discrimination and favouritism may be unintentional and outside of the agent's awareness. The authors also argue that implicit discrimination matters particularly in situations where people are inattentive to the task, under time pressure, when their cognitive capacity is overloaded, and when the task at hand is ambiguous. Implicit attitudes—often measured with implicit association

tests—have been shown to predict a number of outcomes over and above explicit attitudes (e.g., Greenwald *et al*., 2009).[19]

In our setting, graders are arguably motivated to grade objectively and have no incentives to treat matching and non-matching students differently. While there might be a small rivalry between Dutch and Germans when it comes to football, all the graders we talked to about our results agreed that students' nationality and gender should not be considered in the grading process. If explicit taste-based or statistical discrimination is unlikely to play a role in the grading process of exams with their names visibly written on them, it is even less likely that explicit motives play a role in the treatment of the blind group. Of course, implicit differences in treatment can exist, and, indeed, are what we are trying to isolate.

Grading exams may be susceptible to influence by implicit attitudes. Grading requires high levels of concentration; it is time-sensitive in nature, since at the SBE the grading has to be finished within 15 working days after the exam takes place; and students' answers in the exams are often ambiguous. In this context the graders who face exams where the name is clearly visible can thus be influenced by implicit attitudes, whereas when the exams only have a student ID number these attitudes do not affect their judgment. Without further reasons for treating the blind exams in a particular way (other than those discussed below) and without further information about these blind exams, a neutral treatment of the blind exams seems to be the most plausible assumption.

5.2 Recognising Student Characteristics and Being Annoyed with Students in the Blind Group

Even though the exams in the blind group only had student ID numbers, graders might have identified the gender or nationality of some of the students. They could have done so by looking up the student ID numbers to find out the students' names, or by making inferences (consciously or subconsciously) based on other information on the exams such as handwriting or writing style. Both of

[19]Note that implicit discrimination is not inconsistent with modelling discrimination through (dis)utility, as Becker and Goldberg do. In a decision-theoretic framework, utility functions are used to map out choices regardless of whether they are driven by implicit or explicit attitudes.

these cases would result in the blind group being "less blind" and thus in an attenuation of our estimates of endophilia and exophobia.

We can infer the extent of the attenuation due to graders looking up the names of the students by re-estimating (3) including only those graders who explicitly stated in the grader survey that they did not look up the names of the students in the blind group. In Columns (1) and (2) of Table 4 we show that this exercise results in very similar estimates of endophilia by nationality and, interestingly, in a substantially larger estimate of exophilia by gender. Both results are consistent with the intuition outlined above.[20]

It is also possible that graders have negative attitudes toward students who did not write their names, perhaps because the graders take this as a signal that these students are sloppier, or perhaps because they become annoyed at the students' inability to follow instructions. This negative attitude could result in a systematic grade penalty toward the blind group and would, in turn, mean that we overestimate endophilia and underestimate exophobia. To see to what extent this is the case, in the survey of graders we asked what they thought when they saw that students did not write their name on the answer sheet. Out of 33 graders who answered the survey, only four indicated that they "felt annoyed with the students" for not writing their name. In Columns (3) and (4) of Table 4 we show that re-estimating (3) excluding those graders who felt annoyed results in a larger estimate of exophilia by gender—which is consistent with this hypothesis—but also a larger estimate of endophilia by nationality—which is not. In any event these modifications do not alter the qualitative conclusion about the importance of endophilia by nationality. We conclude that even though some graders felt annoyed, this did not translate into systematic downgrading of students in the blind group.

5.3 Rational and Quasi-rational Graders

[20]The attenuation in our estimates from graders recognising the characteristics of the students in the blind group will be proportional to the share of the students that could be correctly identified. Since it is arguably easier to recognise students' gender from their handwriting than students' nationality, we expect our estimates by gender to be more attenuated due to this reason. This is consistent with our findings of larger estimates of endophilia and exophobia by nationality than by gender.

We argue above that endophilia and exophobia as identified in this paper are reflecting implicit attitudes rather than explicit preferences. Let us, however, consider the case in which graders not only have explicit endophilic and exophobic preferences but also are fully rational in a Bayesian sense in their treatment of exams in the blind group. A rational agent eager to exert her preferences would, when confronted with an exam that has no name, form expectations about the student's characteristics and treat the exam accordingly. From the grader's perspective the only information available to her is the overall share of students in the pool of exams whose characteristics do or do not match hers.[21] She will therefore reward or penalise the blind exam based on a weighted average of her endophilic and exophobic preferences, where the weights will be the share of students who do or do not match her characteristics. Under these assumptions our estimates of endophilia and exophobia from (3) can be expressed as:

(4a') $\beta_1 - \beta_2 = e^* - (pe^* - (1-p)x^*) = (1-p)(e^* + x^*)$,

and:

(4b') $\beta_4 - \beta_3 = (pe^* - (1-p)x^*) - (-x^*) = p(e^* + x^*)$,

where the β coefficients are the OLS estimates from (3), now expressed in terms of the latent endophilic and exophobic preferences, e^*, and x^*, and the share of matching students, p.

There are two issues to note in these equations. First, since we observe p we can identify the sum of the two estimated preferences, $e^* + x^*$, using Equations (4a') and (4b'), but we cannot identify endophilia and exophobia separately. The second issue to note is that, if graders do have perfectly rational expectations, there will be a systematic relation between the differences of these coefficients. Specifically, we will observe that $(\beta_4 - \beta_3)/(\beta_1 - \beta_2) = p/(1 - p)$. In other words, with perfect Bayesian updating, the ratio of exophobia to endophilia will equal the ratio of matched to unmatched exams.

[21] A fully Bayesian agent would gradually incorporate the information about her previous guesses of blind exams, and more information that could be of value for guessing the student's characteristics, such as the handwriting and writing style. Such a degree of rationality would imply that the treatment of the blind group would be continuously changing within grader. We abstract from this degree of analysis for two reasons. First, we do not have the information to test most of its implications, e.g., no information on the order in which each student's exam was graded. Second, and more important, our discussion gains little from increasing the complexity to this level given the strong evidence against full rationality.

We can test this very tight prediction about rational behaviour using a non-linear Wald test. That test rejects the null that the two ratios are equal for nationality [p = 0.053], and does so even more strongly for gender [p = 0.027]. We take this as very strong evidence against the existence of rational expectations in grading. This finding aligns with the literature in behavioural economics showing that people's judgement of probabilities is unresponsive to changes in the base rates (Tversky and Kahneman, 1972).

The scenario where all graders are fully rational is an extreme case. The more likely scenario is that graders are partially rational. This scenario can easily be incorporated into our empirical framework by assuming that some fraction of graders (α) hold rational expectations about the blind exams, whereas other graders, a fraction (1 - α), treat those exams neutrally.[22] Under this model our estimates of endophilia and exophobia can be written as weighted averages of neutral and fully rational behaviour:

(4a'') $\beta_1 - \beta_2 = (1 - \alpha)e^* + \alpha(1\text{-}p)(e^*+x^*) = (1- \alpha p)e^* + \alpha(1 - p)x^*$,

and:

(4b'') $\beta_4 - \beta_3 = (1 - \alpha)x^* + \alpha p(e^*+x^*) = \alpha p e^* +[1-\alpha(1 - p)]x^*$.

We can thus express endophilia and exophobia in terms of the OLS estimates, the share of matching and non-matching students, and the share of fully rational graders as:

(5a) $e^* = [(\beta_1 - \beta_2)(\alpha(1 - p) - 1) + (\beta_4 - \beta_3)\alpha(p - 1)]/(\alpha - 1)$,

and:

(5b) $x^* = [(\beta_4 - \beta_3)(\alpha p - 1) + (\beta_1 - \beta_2)\alpha p]/(\alpha - 1)$.

Expressing endophilia and exophobia in this way clarifies the meaning of the estimates in Table 3 under various behavioural scenarios. When all agents grade blind exams neutrally (α = 0), the differences in the estimates correspond to the structural parameters of endophilia and exophobia. When all agents are

[22]It might be more intuitive to think that all agents are partially rational, in the sense that they hold rational expectations about the base group but only incorporate them to a certain degree. The treatment of the blind group *for each grader* can then be modelled as the weighted average of the neutral treatment and the rational expectations treatment, weighted with α and 1- α respectively. This model yields observationally equivalent results to that discussed in the text.

fully rational ($\alpha = 1$) the structural parameters cannot be identified from our estimates. But, more importantly, for every $\alpha < 1$ we can recover the corresponding endophilia and exophobia conditional on α.

Figures 2 and 3 show estimates of endophilia and exophobia by nationality for different levels of graders' rationality. They illustrate three important findings. First, they demonstrate that our reduced-form parameters underestimate endophilia and exophilia by nationality in the presence of rational graders. Second, they show that the estimates of the structural parameters of endophilia by nationality remain statistically significantly different from zero until α rises above 0.5. Based on the results from the literature on base-rate neglect and imperfect Bayesian updating (e.g., Kahneman and Tversky, 1972), we believe, although we obviously cannot prove it, that 0.5 is an upper bound on α. Third, they show that even for large values of α the estimates of endophilia are substantially larger than exophilia, demonstrating once again that meaningful group differences can arise—and in our case do arise—from in-group favouritism rather than from out-group discrimination or favouritism.[23]

In sum, the assumption that graders treat the blind group neutrally is reasonable, because graders' behaviour is likely driven by implicit attitudes. Our results could be attenuated if a share of graders holds rational expectations toward the blind group, or if a few graders recognise some of the students in that group. If graders penalise students in the blind group systematically we would overestimate endophilia and underestimate exophobia, but the data do not support this possibility. Our initial estimates are therefore likely to provide good measures of endophilic and exophobic preferences, and our results remain well identified under the most plausible scenarios about graders' behaviour.

6. Extensions

The graders and exams differ along several dimensions that might affect the extent to which they favour/discriminate for/against students. We first look at whether the graders knew the students they

[23]Figures A2 and A3 in the Appendix show estimates of endophilia and exophobia by gender for different levels of graders' rationality. Incorporating partial recognition and systematic downgrading of the exams with no name (as discussed in Section 5) would not substantially shift the relation between the structural parameters and the level of rationality, α. These results are available upon request.

graded, and thus whether endophilia/exophobia is present towards anonymous and familiar students alike.[24] 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 (3), expanded to include interactions of the *GRADER_KNOWS_STUDENT* 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 almost twice as large as the mean effect in the baseline model, reflecting the combination of no effect when the grader knew the student and a large effect when she did not. There is also evidence of exophilia by nationality only when the grader did not know the student. There is evidence of endophilia and exophilia by gender, but again only when the grader did not know the student.

The exams at the SBE differ in the extent to which they contain mathematical questions, depending mostly on the nature of the courses. Answers on the more mathematical exams, especially answers that can be easily checked against an answer key, are arguably less ambiguous and therefore less likely to be influenced by implicit attitudes. To separate the more from the less mathematical exams we

[24]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.

asked three raters (from the SBE's pool of potential graders) to rate the exams as mathematical or not. We 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 (3), 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 mathematical exams to be less susceptible to implicit discrimination since their answers are arguably less ambiguous. None of the other results in the two columns is statistically significant.

We also examine whether discrimination or favouritism varies with grader experience or 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 more experienced graders may be more used to grading and do so in an efficient and cognitively less demanding manner, resulting in a less pronounced bias. On the other hand, more experienced graders may be more strained for time and their cognitive capacity more demanded due to other obligations, resulting in a more pronounced bias.

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 results, are 1, 8 and 59 courses.[25] Figure 4 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.

[25] 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.

The point estimate of endophilia by nationality is very similar to the estimate in Table 3 at all levels of grader experience. The significant average endophilia shown in Table 3 results disproportionately from the behaviour of the more experienced graders, but the difference by experience is not very large. Inexperienced graders show less endophilia, although the point estimate of their behaviour is still 90 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 graders' 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 equalling 8.8. As Figure 5 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 (3) 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 behaviour 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—substantially favouring those who match her nationality, unsurprisingly given the symmetry in the distribution of teaching evaluations. The worst-rated instructors, however, favour 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 67th 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, favouring students of their own nationality and, to a lesser extent, the other gender.

7. 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 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 favouring the opposite gender on average, though it is less definitive.

The demonstrated importance of graders' expectations about the demographic mix of the "control" groups in our experiment has important general implications for any social experiment in which agents are deciding between suppliers whose characteristics are or are not visible (e.g., so-called audit, or correspondence studies). So long as the agents can draw some inferences about the nature of the suppliers whose individual characteristics are not visible, the simple differential between the treatments of different groups does not measure discriminatory preferences (see, e.g., Heckman, 1998). With appropriate assumptions about the behaviour/knowledge of the agents, however, it can, as we have shown, be the basis for correctly inferring the extent of discrimination.

Assuming that the dominance of endophilia over exophobia that we have demonstrated for nationality is ubiquitous in labour markets, the fact has important implications for the measurement of "discrimination" in labour markets. Decompositions that adjust a gross wage differential into parts due to different characteristics or different treatments in the labour market can be made using either the majority or the minority wage as the base case. In the literature (e.g., Neumark, 1988; Arulampalam *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 behaviour, 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 labour 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 labour-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

Abrevaya, J., and Hamermesh, D. (2012). 'Charity and favoritism in the field: Are female economists nicer (to each other)?' *Review of Economics and Statistics*, vol. 94(1), pp. 202-7.

Ahmed, A. (2007). 'Group identity, social distance and intergroup bias', *Journal of Economic Psychology*, vol. 28(3), pp. 324-37.

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

Altonji, J., and Blank, R. (1999). 'Race and gender in the labor market', in Ashenfelter O., and Card, D., eds. *Handbook of Labor Economics, Vol 3C*. Amsterdam: North-Holland, pp. 3143-3259.

Arulampalam, W., Booth, A., and Bryan, M. (2007). 'Is there a glass ceiling over Europe? Exploring the gender pay gap across the wage distribution', *Industrial and Labor Relations Review*, vol. 60(2), pp. 163-86.

Baguës, M., and Esteve-Volart, B. (2010). 'Can gender parity break the glass ceiling? Evidence from a repeated randomised experiment', *Review of Economic Studies*, vol. 77 (4),pp. 1301-28.

Becker, G. (1957). *The Economics of Discrimination*. Chicago: University of Chicago Press.

Bernhard, H., Fehr, E., and Fischbacher, U. (2006). 'Group affiliation and altruistic norm enforcement', *American Economic Review,* vol. 96(2), pp. 217-21.

Bertrand, M., Chugh, D., and Mullainathan, S. (2005). 'Implicit discrimination', *American Economic Review,* vol. 95(2), pp. 94-98.

Burgess, S., and Greaves, E. (2013). 'Test scores, subjective assessment, and stereotyping of ethnic minorities', *Journal of Labor Economics*, vol. 31(3), pp. 535-76.

Cain, G., (1986). 'The economic analysis of labor market discrimination: A survey', in Ashenfelter, O., and Layard, R., eds., *Handbook of Labor Economics, Vol. 2*. Amsterdam: North-Holland, pp. 693-785.

Cardoso, A.R., and Winter-Ebmer, R. (2010). 'Female-led firms and gender wage policies', *Industrial and Labor Relations Review*, vol. 64(1), pp. 143-63.

Dee, T. (2005). 'A teacher like me: Does race, ethnicity or gender matter?' *American Economic Review*, vol. 95(2), pp. 158-65.

Dillingham, A., Ferber, M., and Hamermesh, D., (1994). 'Gender discrimination by gender: Voting in a professional society', *Industrial and Labor Relations Review*, vol. 47(4), pp. 622-33.

Donald, S., and Hamermesh, D. (2006). 'What is discrimination? Gender in the American Economic Association, 1935-2004', *American Economic Review*, vol. 96(4), pp. 1283-92.

Elder, T., Goddeeris, J., Haider, S. (2010). 'Unexplained gaps and Oaxaca-Blinder decompositions', *Labour Economics*, vol. 17(1), pp.: 284-90.

Feld, J. and Zölitz, U. (2014). 'Understanding peer effects: on the nature, estimation and channels of peer effects', Scandinavian Working Papers in Economics (S-WoPEc), No. 596.

Fershtman, C., Gneezy, U., and Verboven, F. (2005). 'Discrimination and nepotism: The efficiency of the anonymity rule', *Journal of Legal Studies*, vol. 34(2), pp. 371-96.

Fong, C., and Luttmer, E. (2009). 'What determines giving to Hurricane Katrina victims? Experimental evidence on racial group loyalty', *American Economic Journal: Applied Economics*, vol. 1(22), pp. 64-87.

Fryer, R. (2011). 'Racial inequality in the 21st century: The declining significance of discrimination', in Ashenfelter, O., and Card, D., eds., *Handbook of Labor Economics, Vol. 4B,* 2011, Amsterdam: Elsevier, pp. 855-971.

Garicano, L., Palacios-Huerta, I., and Prendergast, C. (2005). 'Favoritism under social pressure', *Review of Economics and Statistics*, vol. 87(2), pp. 208-16.

Giuliano, L., Levine, D., and Leonard, J. (2011). 'Racial bias in the manager-employee relationship: An analysis of quits, dismissals and promotions at a large retail firm', *Journal of Human Resources*, vol. 46(1), pp. 26-52.

Goldberg, M. (1982). 'Discrimination, nepotism and long-run wage differentials', *Quarterly Journal of Economics*, vol. 97(2), pp. 307-19.

Goldin, C., and Rouse, C. (2000). 'Orchestrating impartiality: The impact of "blind" auditions on female musicians', *American Economic Review*, vol. 90(4), pp. 715-41.

Greenwald, A., Poehlman, T.A., Uhlmann, E., and Banaji, M. (2009) 'Understanding and using the implicit association test: III. Meta-analysis of predictive validity', *Journal of Personality and Social Psychology,* vol. 97(1), pp. 17-41.

Hanna, R., and Linden, L. (2012). 'Discrimination in grading', *American Economic Journal: Economic Policy*, vol. 4(4), pp. 146-68.

Heckman, J. (1998). 'Detecting discrimination', *Journal of Economic Perspectives*, vol. 12(2), pp. 101-16.

Hinnerich, B.T., Höglin, E., and Johannesson, M. (2011). "Are boys discriminated in Swedish high schools?' *Economics of Education Review*, vol. 30(4), pp. 682-90.

Kahneman, D., and Tversky, A. (1972). 'Subjective probability: A judgment of representativeness', *Cognitive Psychology*, vol. 3(2), pp. 430-54.

Laband, D., and Piette, M. (1994). 'Favoritism versus search for good papers: Empirical evidence regarding the behavior of journal editors', *Journal of Political Economy,* vol. 102(1), pp. 194-203.

Lavy, V. (2008). 'Do gender stereotypes reduce girls' or boys' human capital outcomes? Evidence from a natural experiment',"*Journal of Public Economics*, vol. 92(10-11), pp. 2083-105.

Levitt, S., and List, J. (2007). 'What do laboratory experiments measuring social preferences reveal about the real world?' *Journal of Economic Perspectives*, vol. 21(2), pp. 153-74.

Neumark, D. (1988). 'Employers' discriminatory behavior and the estimation of wage discrimination', *Journal of Human Resources*, vol. 23 (3), pp. 279-95.

Parsons, C., Sulaeman, J., Yates, M., and Hamermesh, D. (2011). 'Strike three: Discrimination, incentives and evaluation', *American Economic Review*, vol. 101(4), pp. 1410-35.

Price, J., and Wolfers, J. (2010). 'Racial discrimination among NBA referees', *Quarterly Journal of Economics*, vol. 125(4), pp. 1859-87.

Rivkin, S., Hanushek, E., and Kain, J. (2005). 'Teachers, schools and academic achievement', *Econometrica*, vol. 73(2), pp. 417-58.

Tversky, A., and Kahneman, D. (1974). 'Judgment under uncertainty: Heuristics and biases', *Science*, vol. 185(4157), pp. 1124-31.

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		7.071	6.992
		(0.030)	(0.038)
Feel Close to Blacks		5.138	5.525
		(0.032)	(0.038)
N		3,550	2,174
ρ		0.145	0.230
BLACKS			
Feel Close to Whites		5.810	5.907
		(0.082)	(0.108)
Feel Close to Blacks		7.588	7.655
		(0.078)	(0.096)
N		651	387
ρ		0.241	0.285

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

Table 2. Student Characteristics by Intended Treatment Status*

Internal validity: Pre-experiment							
	(1) Blind			(2) Visible			*p-value of difference*
	Mean	*SD*	*Obs.*	*Mean*	*SD*	*Obs.*	*Blind-Visible*
Female	0.369	0.483	452	0.352	0.478	1,043	[0.502]
German	0.374	0.484	452	0.353	0.478	1,043	[0.420]
Dutch	0.363	0.481	452	0.343	0.475	1,043	[0.452]
GPA	7.197	0.628	443	7.215	0.665	1,021	[0.607]
Participation	7.690	0.986	306	7.633	1.031	706	[0.386]
Presentation	7.795	1.164	191	7.930	1.059	436	[0.179]
Term paper	7.870	0.665	109	7.743	0.898	281	[0.126]

Internal validity: Within-experiment							
	(1) Blind			(2) Visible			*p-value of difference*
	Mean	*SD*	*Obs.*	*Mean*	*SD*	*Obs.*	*Blind-Visible*
Multiple Choice exams	5.829	1.972	277	6.043	1.942	661	[0.128]
Fill-In-The-Blank exams	5.325	2.208	152	5.555	1.996	382	[0.264]

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

Table 3. Basic Estimates of the Extent of Graders' Endophilia (Favouritism) and Exophobia (Discrimination) by Nationality and Gender with the Blind Group Viewed Neutrally (N = 9,330)*

	(1)	(2)	(3)			(4)	
	Nationality	Gender	Nationality			Gender	
Interaction with:	-	-	*German*	*Dutch*	*Other*	*Female*	*Male*
*(1) MATCH*VISIBLE*	0.289	-0.032	0.307	-0.010	-	0.143	-0.024
	(0.038)	(0.026)	(0.023)	(0.102)	-	(0.029)	(0.026)
*(2) MATCH*BLIND*	0.124	-0.108	0.174	-0.195	-	0.138	-0.128
	(0.079)	(0.038)	(0.095)	(0.107)	-	(0.070)	(0.039)
*(3) NON-MATCH*VISIBLE*	0.183	-0.073	0.161	-0.048	-0.116	0.150	-0.093
	(0.050)	(0.041)	(0.069)	(0.053)	(0.067)	(0.048)	(0.050)
*(4) NON-MATCH*BLIND*	0.155	-0.111	0.036	-0.078	-0.066	0.063	-0.093
	(0.061)	(0.044)	(0.083)	(0.076)	(0.078)	(0.036)	(0.083)
Endophilia [(1)-(2)]	0.165	0.076	0.133	0.185	-	0.005	0.104
p =	[0.033]	[0.087]	[0.180]	[0.071]	-	[0.944]	[0.032]
Exophobia [(4)-(3)]	-0.028	-0.038	-0.125	-0.024	0.050	-0.087	-0.000
p =	[0.546]	[0.556]	[0.027]	[0.750]	[0.489]	[0.150]	[0.998]
Adj. R^2	0.016	0.010	0.016			0.010	

*Standard errors in parentheses and p-values in square brackets. Both are clustered by ITT and Course. *VISIBLE* and *BLIND* are defined based on the intention to treat (ITT). Columns (1) and (2) present the estimates of Equation (3) without a constant. Columns (3) and (4) are based on Equation (3), with the main variables interacted with *CHARACTERISTIC*, where *CHARACTERISTIC* represents indicators for student nationality in Column (3) and for student gender in Column (4). *MATCH*Other* interactions in Column (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 other variables.

Table 4. Endophilia and Exophobia by Other Characteristics Based on Graders' Survey Responses*

	(1) Nationality	(2) Gender	(3) Nationality	(4) Gender
Regression:	*Did Not Look up Names*		*Were Not Annoyed*	
Endophilia	0.151	0.008	0.255	0.061
p =	[0.030]	[0.851]	[0.006]	[0.298]
Exophobia	-0.035	-0.145	-0.050	-0.152
p =	[0.490]	[0.004]	[0.314]	[0.005]
N	5,108		5,526	
Adj. R^2	0.015	0.008	0.014	0.010

*p-values in square brackets, clustered by ITT and Course. We report linear combinations based on extensions of Equation (3) with all results based on the ITT. Columns (1) and (2) are based on the sample of graders who did not look up any of the names in the Blind group of exams. Columns (3) and (4) are based on the sample of graders who did not report feeling annoyed with the exams in the Blind group. Main effects are included throughout.

Table 5. Endophilia and Exophobia by Other Characteristics of Graders and Exams (N = 9,330)*

		(1) Nationality	(2) Gender			(3) Nationality	(4) Gender
Grader knows the student?:				*Exam was mathematical?:*			
Endophilia	*No*	0.280	0.151	Endophilia	*No*	0.255	0.032
	p =	[0.005]	[0.007]		p =	[0.012]	[0.632]
	Yes	0.060	-0.002		*Yes*	-0.030	0.136
	p =	[0.583]	[0.982]		p =	[0.671]	[0.001]
Exophobia	*No*	-0.114	-0.144	Exophobia	*No*	0.046	-0.014
	p =	[0.021]	[0.020]		p =	[0.479]	[0.875]
	Yes	0.072	0.090		*Yes*	-0.142	-0.089
	p =	[0.449]	[0.355]		p =	[0.000]	[0.171]
F-test differences:		[0.122]	[0.100]			[0.000]	[0.349]

*p-values in square brackets, clustered by ITT and Course. We report linear combinations based on extensions of Equation (3) with all results based on the ITT. Columns (1) and (2) report interactions of the main variables with *GRADER_KNOWS_STUDENT*, Columns (3) and (4) show interactions of the main variables with *MATHEMATICAL_EXAM*. The F-test differences report the p-values from testing the null hypothesis that Endophilia and Exophobia are equal for the groups defined by *GRADER_KNOWS_STUDENT* and *MATHEMATICAL_EXAM*, respectively. Main effects are included throughout.

Table 6. Endophilia and Exophobia by Teachers' Experience and Quality (N = 9,197)*

		(1)	(2)	(3)	(4)
	Percentile:	Nationality	Gender	Nationality	Gender
At the m^{th} percentile of:		*Experience*		*Teacher Quality*	
Endophilia	5th	0.180	0.075	0.334	0.076
	p =	[0.098]	[0.149]	[0.012]	[0.313]
	50th	0.182	0.076	0.154	0.080
	p =	[0.064]	[0.109]	[0.037]	[0.068]
	95th	0.200	0.080	-0.006	0.083
	p =	[0.031]	[0.274]	[0.952]	[0.126]
Exophobia	5th	-0.020	-0.023	-0.051	-0.152
	p =	[0.740]	[0.790]	[0.491]	[0.034]
	50th	-0.023	-0.031	-0.026	-0.026
	p =	[0.660]	[0.681]	[0.569]	[0.686]
	95th	-0.045	-0.093	-0.004	0.085
	p =	[0.600]	[0.081]	[0.941]	[0.353]
F-test interactions:		[0.967]	[0.721]	[0.164]	[0.061]

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

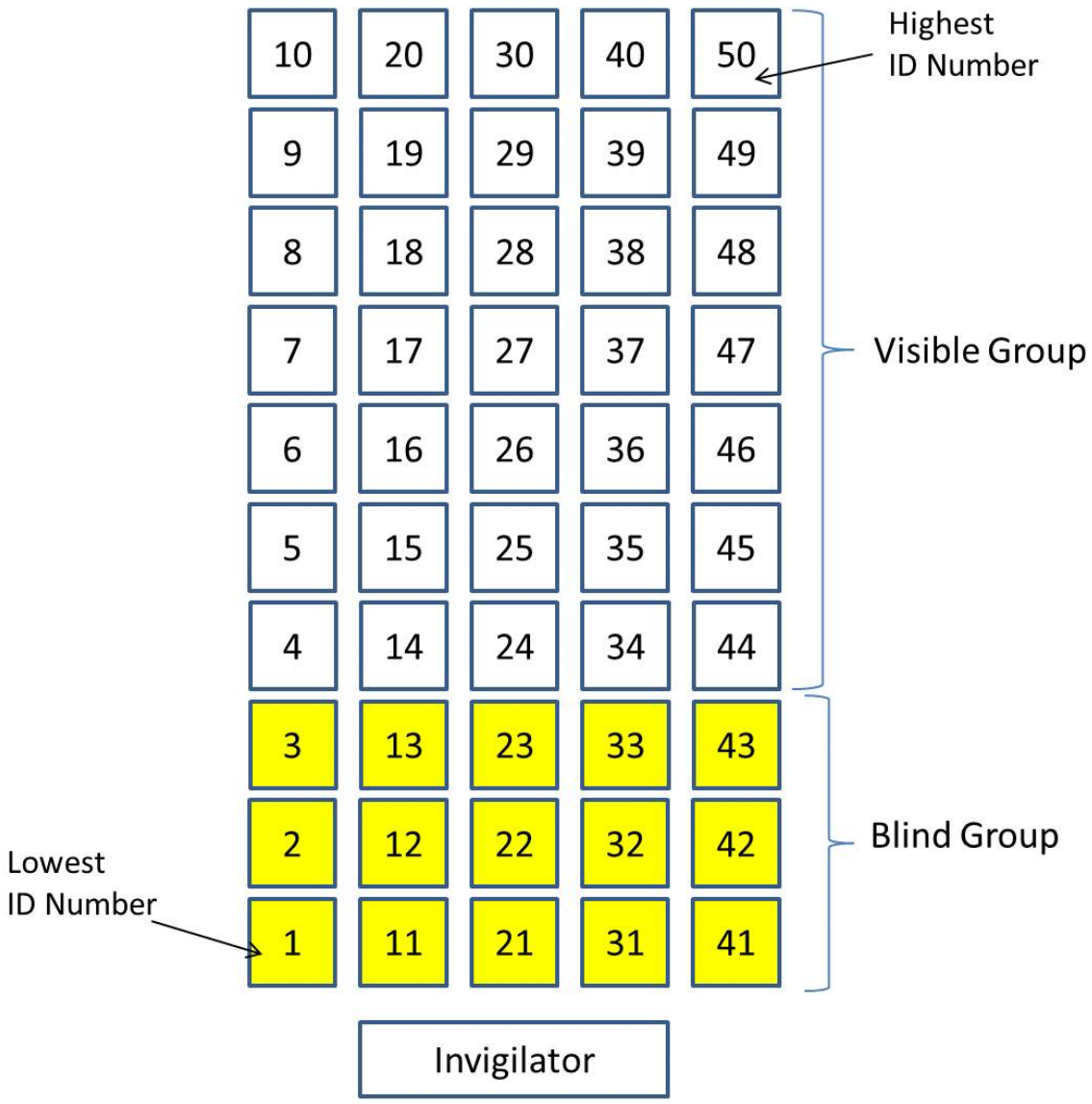

Figure 1: Seating Arrangement for the Experiment*

*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, as usual, thus creating the Visible group.

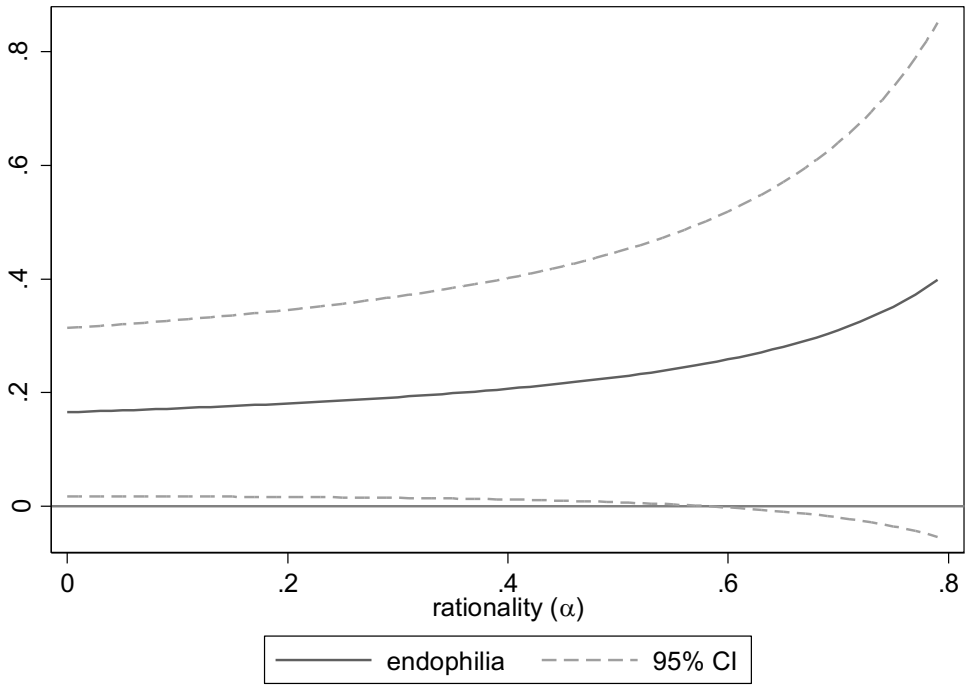

Figure 2. Estimates of Endophilia by Nationality for Different Shares of Fully Rational Graders (α)

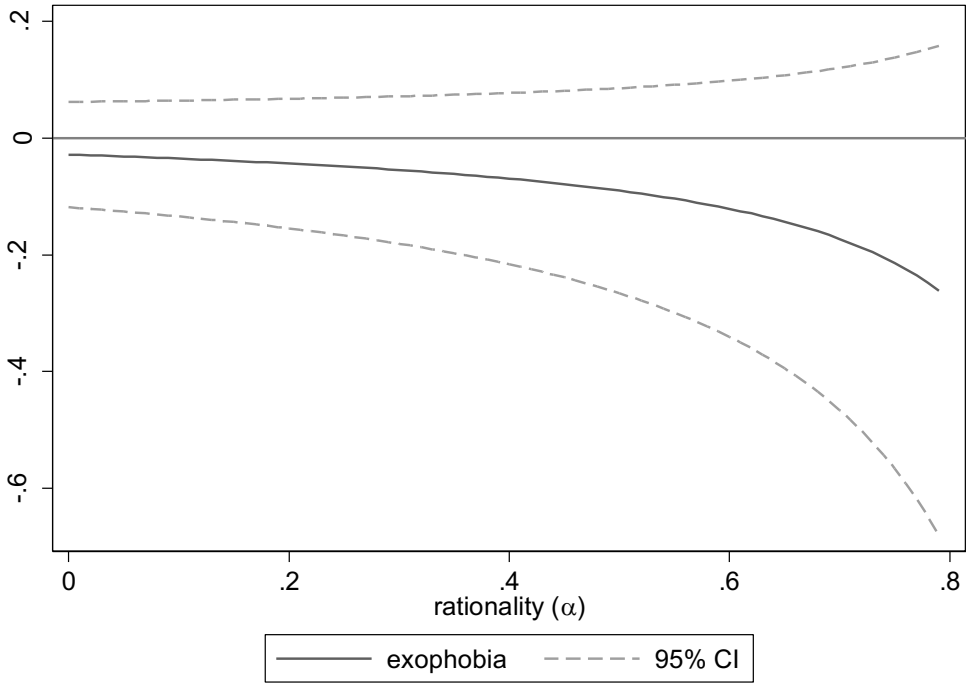

Figure 3. Estimates of Exophobia by Nationality for Different Shares of Fully Rational Graders (α)

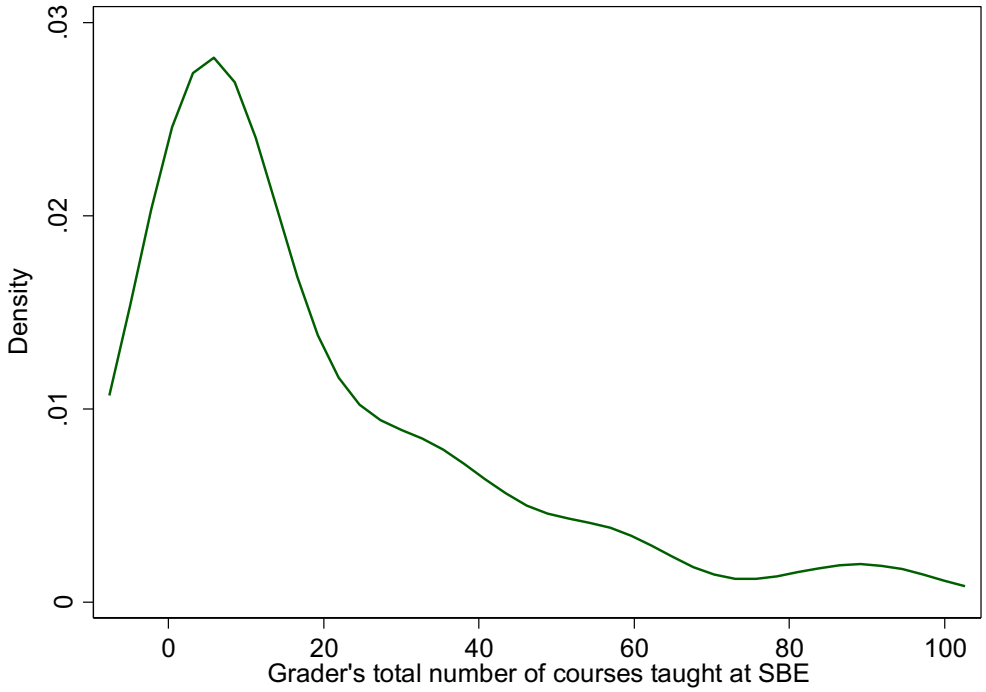

Figure 4. Kernel Density of the Distribution of Grader Experience

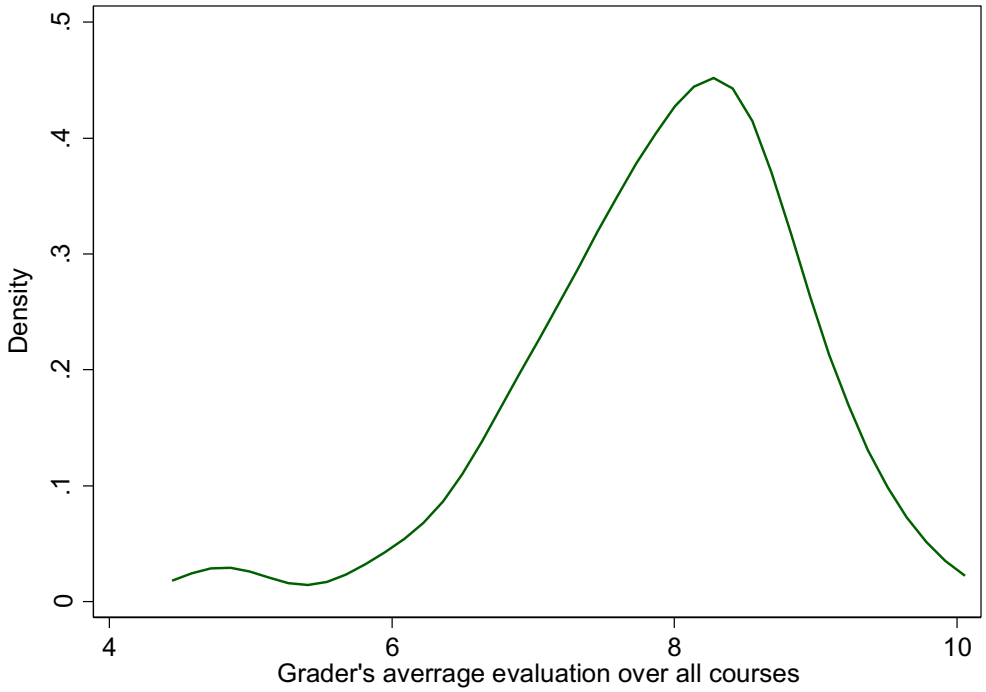

Figure 5. Kernel Density of the Distribution of Student Evaluations of Graders

Appendix

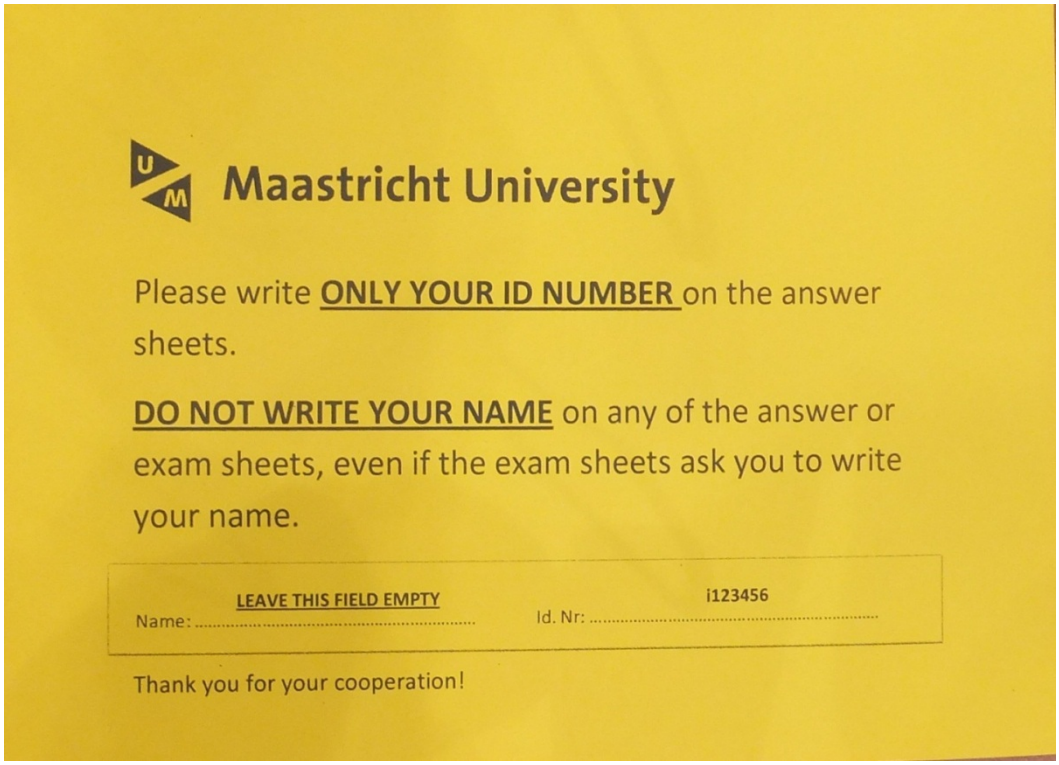

Figure A1. Yellow Sheet Placed on Some Students' Desks Before the Exam.

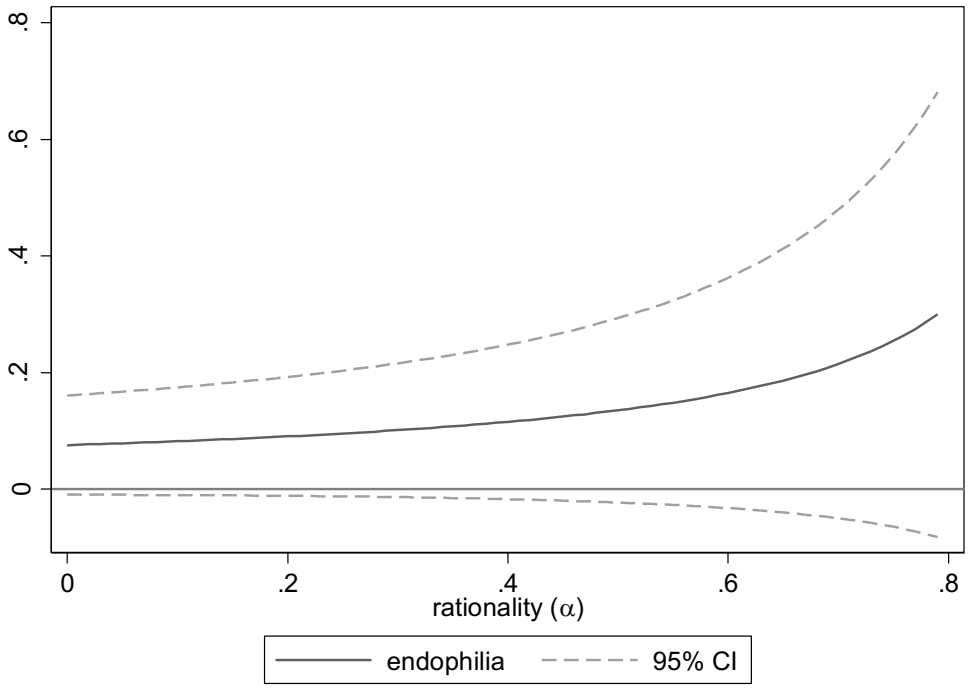

Figure A2. Estimates of Endophilia by Gender for Different Shares of Fully Rational Graders (α)

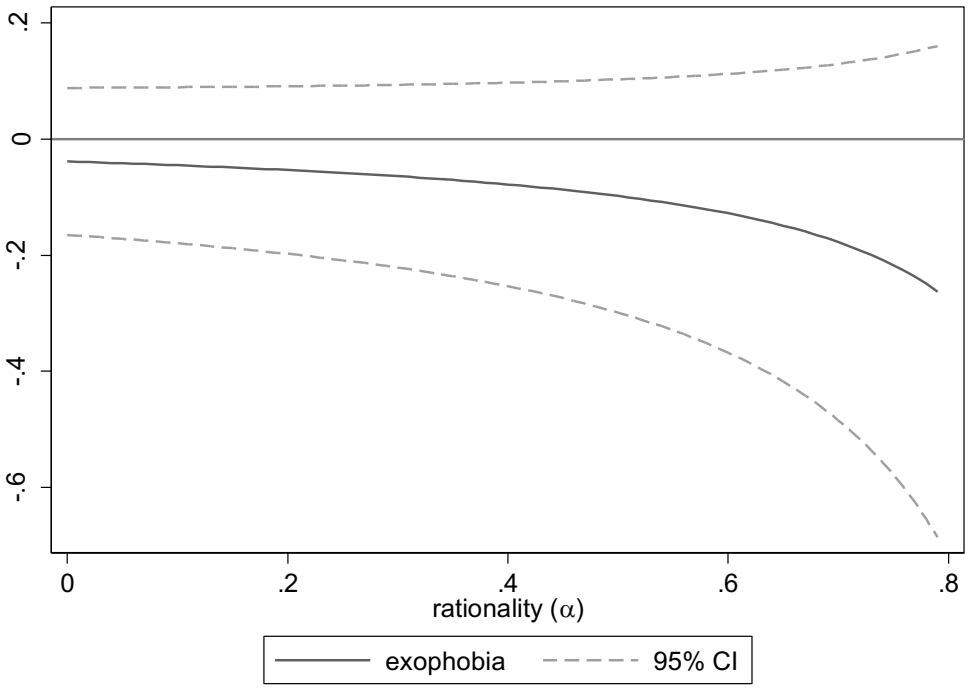

Figure A3. Estimates of Exophobia by Gender for Different Shares of Fully Rational Graders (α)