Forthcoming in : American Economic Journal: Applied Economics

The Price of Prejudice

By MORTEN STØRLING HEDEGAARD AND JEAN-ROBERT TYRAN *

We present a new type of field experiment to investigate ethnic prejudice in the workplace. Our design allows us to study how potential discriminators respond to changes in the cost of discrimination. We find that ethnic discrimination is common but highly responsive to the "price of prejudice", i.e. to the opportunity cost of choosing a less productive worker on ethnic grounds. Discriminators are on average willing to forego 8 percent of their earnings to avoid a co-worker of the other ethnic type. The evidence suggests that animus rather than statistical discrimination explains observed behavior.

* Hedegaard: University of Copenhagen, Department of Economics, Øster Farimagsgade 5, building 26, Copenhagen, DK-1353 (Morten.Hedegaard@econ.ku.dk). Tyran (corresponding author): University of Vienna, Department of Economics and University of Copenhagen, Department of Economics, and CEPR (London), Oskar-Morgenstern-Platz 1, Vienna, A-1090 (Jean-Robert.Tyran@univie.ac.at). This paper is a thoroughly revised and extended version of Hedegaard's PhD thesis "Essays in experimental economics: preferences and information", chapter 1. We gratefully acknowledge generous financial support by the Rockwool Foundation's Research Unit (RFF). We are particularly grateful to RFF's former Research Director Torben Tranæs for his encouragement and support. We also thank Dirk Engelmann, Simon Gächter and Ernesto Reuben as well seminar participants at Columbia U, NYU, Oxford, WZB Berlin and U Zurich for helpful comments and Nete Daly, Pelle Flachs, Anne Schjellerup Olsen and Pablo Torija for assistance in conducting the experiment.

Field experiments have been used for more than 40 years to investigate the causes of ethnic discrimination in the workplace (**s**ee Riach and Rich 2002 for a survey).[1]

[1] Field experiments have also been used to measure gender discrimination (e.g. Goldin and Rouse 2000) and other types of discrimination in the labor market (e.g. Neumark, Bank and van Nort 1996, Behaghel, Crépon, Le Barbonchon 2015), and discrimination in other markets (e.g. Ayres and Siegelman 1995, Edelman, Luca and Svirsky 2017, Gneezy, List and Price 2012, Levitt 2004, List 2004, Yinger 1998).

In so-called correspondence tests (e.g. Bertrand and Mullainathan 2004), pairs of fictitious resumes are submitted to employers by mail. Discrimination is inferred from differential callback or job-offer rates across pairs of workers which are similar in all respects except for ethnicity. This approach has many advantages but a limitation is that since applicants are equally productive by design, *discrimination is free* for the discriminator. Correspondence tests may therefore exaggerate the true extent of discrimination (e.g. Heckman and Siegelman 1993). In addition, correspondence tests are silent on how discrimination responds to *changes* in the price of discrimination because they usually do not vary the cost of choosing one candidate over the other (see Neumark 2012 for a discussion).

This paper presents a new type of field experiment to investigate ethnic prejudice in the workplace. Our experiment combines tight control and accurate observation of the earnings consequences of discrimination with the desirable property that subjects do not know that they are participating in an experiment. We achieve this combination by hiring juveniles for a job that is perfectly natural to them – preparing letters for a mailing – and have them do the job in office facilities, under conditions we tightly control.[2] The main innovation of our design is that it allows us to vary the price a discriminator has to pay, and to control the information potential discriminators have about that price. Our approach thereby allows us to observe discrimination when discrimination is costly to the discriminator.[3] Hence, we are able to put a price tag on discrimination choices or, borrowing Gary Becker's

[2] The combination of tight control and subject unawareness has been implemented previously, e.g. in experimental studies of labor markets. It is common practice to call such experiments "natural field experiments" even when the job is performed on the premises of a university, see e.g. Gneezy and List (2006) or Kube, Maréchal, and Puppe (2013) who recruit students to catalogue the holdings of a small library at a university.

[3] A few studies have been able to relate variations in price to discrimination choices in a context not related to work. For example, Baccara et al. (2014) use variation in the cost of adopting children in the US to estimate the willingness to pay for babies with particular (ethnic, among others) characteristics. Pope and Sydnor (2011) use variation in interest rates in online peer-to-peer lending to show that statistical discrimination of black borrowers absent animus cannot explain net returns observed in loan-performance data. Levitt (2004) uses data from a TV show to test how statistical discrimination of candidates reacts to changes in cost.

(1957) expression[4], to estimate how discrimination responds to the "price of prejudice", rather than just observing that discrimination occurs when it is costless.[5]

In our main treatment (called Info), we randomly assign a price of discrimination by giving decision makers the choice between two co-workers of different ethnicity and of *known* productivity. The decision maker and the chosen co-worker then form a team and are paid according to team output. This procedure allows us to estimate the causal effect of changes in the price of discrimination (the amount of money lost from not choosing the more productive worker) on the "demand for discrimination" (the propensity to choose the less productive worker). We find that ethnic discrimination is common but remarkably responsive to the price of discrimination.

Our experiment proceeds as follows. We hire 162 juveniles from secondary schools in Copenhagen, Denmark, with Danish-sounding and Muslim-sounding names to prepare letters for a large mailing and pay a piece rate. Workers are requested to show up for work twice in two consecutive weeks. In the first round, they work by themselves and we measure their individual productivity on the job. Before they come back for the second round, we call randomly selected workers on the phone and inform them that they will again do the same job but now have to work in teams of two. They are informed that they are paid the same piece rate as in round one and share earnings from team output in round two with the co-worker. These randomly selected workers can choose whom to work with. The choice is between a candidate from the ethnic majority group and a candidate from an ethnic minority group. In treatment Info, we provide the decision maker with information

[4] "Price and Prejudice" is the title of part 2 in Becker (1976, The economic approach to human behavior) which is a revised version of his PhD thesis, published in 1957.

[5] The reason why most tests involve a zero cost of discrimination to the discriminator is probably that they take the standard definition of discrimination as a point of departure. Altonji and Blank (1999: 3168) define discrimination in the labor market as "a situation in which persons who provide labor market services and who are equally productive in a physical or material sense are treated unequally in a way that is related to an observable characteristic such as race, ethnicity, or gender".

about the individual productivity of the two candidates, i.e. the number of letters they prepared in round one, and their first names as a marker of ethnicity. Because the candidates are randomly selected from the pool of workers, the productivity difference between the two, and therefore also the price to be paid when choosing a less productive worker on ethnic grounds, is random. Rational decision makers who choose the less productive worker of the same ethnic type thus discriminate knowingly and deliberately.

We find that discrimination is common even at a substantial cost and that the tendency to discriminate is not different across ethnic types. We estimate that discriminators on average are willing to forego 8 percent of their earnings in round two to avoid a co-worker of the other ethnic type. Our main result from treatment Info is that discrimination is highly responsive to the price of prejudice. Our best estimate is an elasticity of -.9, i.e. we find that the probability to discriminate falls by about 9 percent if the price of discrimination goes up by 10 percent. We interpret discrimination in treatment Info as being “animus-driven” (or “taste-based” in the words of Becker 1957).

We demonstrate the power of our estimate of taste-based discrimination out of sample, in a treatment called NoInfo. Treatment NoInfo is the same as Info except that now decision makers do *not know* candidates’ productivities. Decision makers thus have to form beliefs about the average productivity of ethnic groups and both taste-based and statistical discrimination can therefore shape hiring choices in NoInfo. Decision makers who engage in statistical discrimination are assumed to hold unbiased beliefs about the relative average productivity of ethnic types and to choose the worker of the more productive type (Arrow 1973, Phelps 1972).

Our analysis shows that statistical discrimination does not explain observed outcomes in NoInfo well. We find a large gap between observed earnings and earnings predicted by statistical discrimination (about 4 percent of total output). To account for taste-based discrimination, we use our estimate from treatment Info and find that it predicts well out of sample; about 40 percent of that gap is explained by animus-driven prejudice. Thus, our results suggest that prejudice is an important cause of ethnic discrimination in the workplace, and that it needs to be taken into account above and beyond the theory of statistical discrimination.

We proceed as follows. Section I presents the experimental design, section II describes our framework for empirical analysis, and section III presents results. Section IV shows that our interpretation in terms of taste-based discrimination is robust to alternative accounts and that our measure of animus predicts well out of sample. Section V concludes.

I. Experimental Design

A general description of the experiment is as follows. We recruit an approximately balanced sample of juveniles with Danish-sounding and Muslim-sounding names from secondary schools in central Copenhagen for a job involving stuffing envelopes for a large-scale mailing. Workers commit to show up twice and indicate their availability for work. In the first round, they work at a piece rate in isolation. This round serves to measure individual productivity on the job. In the second round, workers are required to work in teams of two, and some randomly selected workers (the "decision makers") can choose their partner. We construct triples of workers by randomly drawing one decision maker and two "candidates", one with a Danish-sounding name and one with a Muslim-sounding name.

The discrimination choice is made between rounds one and two. We call the decision makers on the phone and explain that they will do the same job at the same piece rate in round two, but will have to work in teams of two. In treatment Info, they learn the first names and the productivity (i.e. number of envelopes stuffed in round one) of the two candidates. Decision makers know that all candidates are equally experienced and have similar characteristics. In particular, they know that all candidates have worked on the same job under identical conditions and that they are recruited from secondary schools. When the decision maker has made a choice, we call the chosen candidate requesting him or her to show up at a particular time. In round two, teams are formed according to the choices of the decision makers whenever possible, and workers are paid out for both rounds.

We took great care to implement a proper natural field experiment – in which participants are not aware that they are part of an experiment. In particular, we have been careful at all stages of the experiment to assure that the job itself and the work conditions appear natural to participants, that the experiment (in particular the information provided to decision makers) is tightly controlled, and that all aspects of the experiment are consequential and do not involve deception.

A. Detailed Description of Procedures

1. Recruiting.—We distributed hundreds of flyers in eleven upper secondary schools in central Copenhagen.[6] The flyer explains that the University of Copenhagen is looking for part-time workers to prepare a major mailing. The flyer also explains that applicants are expected to show up for two hours in each of two consecutive weeks. Applicants are requested to call us on a phone number indicated on the flyer.

We recruited in upper secondary schools because these juveniles have relatively low outside options, are similar with respect to age (16-20 years old) and education, are legally allowed to work for money, and because there is considerable naturally occurring ethnic heterogeneity in this group (23% of juveniles in these schools are immigrants). Using a homogenous subject pool has the advantage of minimizing unobserved heterogeneity across ethnic types, for example with respect to language

[6] The flyer is reprinted in appendix A. Appendix B shows the location of the schools.

skills. In addition, it is feasible to recruit an approximately balanced sample by gender from this pool. The reason for wanting a balanced sample is that we keep the triples (see below) separate by gender to avoid confound of ethnicity and gender.

2. Names as Markers of Ethnicity.—Upon calling us, we record the applicants' names, phone numbers, and where they saw the flyer. Applicants indicate when they are available for work in both rounds and are requested to make a commitment to show up at any of these slots. We classify the applicants according to their first names as Danish-sounding or Muslim-sounding. We call 169 persons with high availability[7] in approximately balanced proportions (see table 1).[8]

We focus on applicants with typical Danish-sounding and Muslim-sounding names because these ethnic groups are by far the largest in Denmark.[9] We use first names as markers of ethnicity since it is natural to refer to a person in Denmark by first name across all social strata. Using first names to evoke stereotypes is common practice in correspondence tests. These tests use fictitious first names which can be chosen to be particularly strong markers of ethnicity (e.g. Lakisha vs. Emily in Bertrand and Mullainathan 2004). In contrast, we use participants' actual first names to mark ethnicity. In a follow-up study with 144 subjects, we find that our ethnic markers are highly effective and confound rarely occurs. For example, names we classify as Muslim-sounding are thought to be Danish-sounding only in about 1 percent of the cases (see appendix D for details).

[7] 95 percent (*n* = 169) of participants were available on 3 or more days, 55 percent on 6 or more days in round 2.

[8] Table 1 shows that the names of 7 workers did not fit either ethnic type. These workers (and the teams they worked in) are excluded from our analysis below. Table 1 does not list 27 workers who participated in a pre-test. These workers were recruited from a school where we did not recruit for the main experiment.

[9] According to official statistics (2009, www.statistikbanken.dk), 69 percent of immigrants in Denmark are from non-Western countries, and most of these originate from countries with high proportions of Muslims such as Turkey, Iraq and Pakistan.

Note that the first names of the ethnic minority group are Muslim-sounding but may also be foreign-sounding to native Danes. Thus, our study cannot not provide a definitive answer on whether the animus we measure among Danes is directed at Muslims or foreigners living in Denmark more generally. However, a correspondence test designed to investigate this issue (Adida, Laitin and Valfort 2010) for France suggests that animus against Muslims is more pronounced than animus against foreigners in general.[10]

3. Measuring Individual Productivity.—A total of 169 persons work in round 1 of our experiment. Workers are requested to show up at particular times and are led to separate rooms to minimize interaction between them. The job is explained and demonstrated to each worker individually. The job involves stuffing letters marked with an ID-number into envelopes. These numbers have to be looked up in a binder and are associated with different letter types. Depending on the type, letters have to be complemented with a gift and sorted into specific bins (see appendix C for details). When participants indicate that they understand the task, the payment scheme (the piece rate is DKK 4, approx. €0.5 per letter), and that they are ready to start, an alarm clock is set in the control room (see Figure B2 in appendix B). After exactly 90 minutes a staff person returns to the worker and counts the number of envelopes stuffed. Each worker got a receipt confirming their entitlement and was paid at the end of round 2 to provide them with incentives to return.

The job is ideal for our purposes for several reasons. First, the job is easy to explain and easy to learn for juveniles within the given time frame. Second, the job can meaningfully be done both in isolation and in a team of two workers. Third, teamwork on the job requires minimal spoken interaction which minimizes the

[10] The study combines a foreign-sounding last name (Diouf, a typical name in Senegal) either with a Christian (Marie) or Muslim (Khadija) first name. Response rates for Marie Diouf and a reference candidate with a typical French name (Aurélie Ménard) were not different. However, response rates for Khadija Diouf were significantly lower than for Marie Diouf.

motive to discriminate against members from a different "speech community" (e.g. Grogger 2011). Fourth, the task produces sufficient variation in individual output which is essential to make discrimination costly. Fifth, the job is not artificial. It is not unusual for juveniles to work in a temporary job like stuffing envelopes and the job is real in the sense that we effectively used the letters for a large-scale mailing.[11]

TABLE 1—NUMBER OF WORKERS IN ROUND 1 BY GENDER AND ETHNICITY

	Ethnicity			
Gender	Danish-sounding name	Muslim-sounding name	Other name	*Total*
Female	40	46	5	*91*
Male	40	36	2	*78*
Total	*80*	*82*	*7*	*169*

4. Matching Procedure.—Upon completion of round 1, we match workers into triples as follows. We randomly select a person to be the decision maker. Thus, the decision maker may have a Danish-sounding or a Muslim-sounding name. We then determine the set of all suitable candidates for this decision maker. This is the set of participants who are of the same gender as the decision maker, are from a different school, and are available for work on at least one of the time slots indicated by the decision maker. We randomly draw two candidates from this set. One candidate is of the same ethnic type as the decision maker (*same* for short), one is of the other type (*other* for short). In treatment Info, the draws are repeated until *same* is *less* productive than *other* and the two candidates are available on different weekdays. If no such pair exists, we randomly draw a new decision maker from the pool.

[11] We used the letters for a mailing to recruit participants for a large-scale internet study. This study used different letter types necessitating sorting the letters. We randomly checked 5 letters for each participant in round 1. The error rate was low (0.05) and did not differ by ethnic type (p = 0.270, χ^2-test). Error rates also do not differ by team composition in round 2 (p = 0.688, χ^2-test).

We randomly draw decision makers to observe discrimination choices by both ethnic types. The ability to observe discrimination choices by minority decision makers is, to the best of our knowledge, a unique feature of this study. For example, correspondence tests usually do not observe the ethnicity of the employer and simply assume that he or she belongs to the ethnic majority. We match candidates and decision makers from different schools to exclude that they personally know each other, thus avoiding confound of ethnic discrimination with a preference for a personal acquaintance. We are able to match teams from different schools because we gather information about school affiliation from participants when they apply for the job over the phone. Randomly drawing two candidates serves to generate a random price of discrimination (i.e. the earnings foregone by choosing *same* over *other*). Thus, the price is independent of any animus that may be present. Random assignment of price to decision makers is a precondition for identifying taste-based discrimination. The restriction imposed in Info that *same* has lower productivity than *other* serves to maximize the number of informative choices. Choices are informative in the sense that decision makers with strong animus are likely to be detected. The reason why the candidates must be available on different weekdays is that we frame the discrimination choice as a choice between two weekdays rather than between two persons, as is explained next.

5. Discrimination Choice.—The discrimination choice is made on the phone prior to round 2. Upon answering the phone, the decision maker is asked to confirm availability on the two time slots determined by the matching procedure (Tuesday and Wednesday 2 p.m. - 4 p.m., say). If the decision maker cannot reconfirm availability or says that he has a preference for work on a particular day, we say we have to make new arrangements and call back later. In this case, the triple pertaining to this decision maker is reinserted into the pool and a new triple is drawn according to the matching procedure described above. If the answer is affirmative, decision makers are informed that the job in round 2 is the same and is paid according to the same piece rate as in round 1. They are told that, unlike in round 1, they have to work in teams of two and that they have to share the revenue from teamwork.[12] Decision makers are told that which person they are going to work with depends on which day they choose. In treatment Info, the decision makers are told the first names and the number of envelopes stuffed in round 1 for both candidates and asked to make a choice. For example, "If you choose Tuesday, you will work with Ahmed who stuffed 150 envelopes last week. If you choose Wednesday, you will work with Christian who stuffed 110 envelopes last week. So, when would you like to work, Tuesday or Wednesday at 2 p.m.?" In treatment NoInfo, the procedure is the same except that we do not mention the individual output of candidates in round 1.[13]

An important advantage of this procedure is the high degree of control it provides over the information available to the decision makers. In both treatments, decision makers know that candidates are similar (they are recruited from the same set of schools) and have the same experience on the job (they all worked in round 1 under the exact same conditions). Beyond that, in treatment Info, the decision makers know *only* the names and productivities of the candidates. Since they cannot personally identify or see the candidates, factors such as attractiveness or personal appearance cannot affect decisions in our design (see e.g. Möbius and Rosenblat 2006 for experimental and Hamermesh and Biddle 1994 for field evidence on personal attractiveness and discrimination).

We frame the discrimination choice as a choice of workdays rather than persons to minimize so-called Hawthorne or experimenter demand effects (see Zizzo 2010

[12] If asked, we justified that the job has to be done in teams of two by explaining that "we found out that working in teams of two is more effective and therefore workers on average earn more than last week". We knew from a pretest with 27 participants that this claim is true.

[13] Non-chosen candidates were reinserted into the pool of participants and were matched into another triple, either as decision maker or candidate. Thus, our design does not necessarily imply a cost to the discriminated.

for a general discussion). Such effects are particularly relevant in experiments on discrimination because of the illicit nature of discrimination and of participants' concerns to conform to notions of political correctness (see e.g. Kawakami et al. 2009). Framing the choice in terms of days is thus a precondition for measuring discrimination preferences when decision makers think they should comply with a non-discrimination norm. The drawback of this framing is a potential confound of a taste for discrimination with a preference for a work day (section 5.1 shows that this concern is unwarranted).

6. Credibility and Consequentiality.—We take great care to create a natural setting, to measure output and to provide information with tight control, to insure that all information provided to decision makers is truthful, and that choices are consequential. For example, decision makers were presented with a choice between two real people, we indicate their actual first names, and their actual productivity in round 1. Decision makers are matched to work with the partner of their choice in round 2 whenever possible (i.e. when both show up on time) which implies that the chosen candidate cannot make a discrimination choice.

We believe that our choice of the location and work task was highly credible in the sense that workers did not know that they were participating in an experiment. We made choices consequential to avoid deception and disappointment. For example, decision makers who opt for a highly productive co-worker would be antagonized if forced to work with a low-productive partner in round 2.

II. Conceptual Framework for Econometric Analysis

Section II.A describes a discrete choice framework for treatment Info which serves as a basis for estimating the price of discrimination, the demand for discrimination, and the willingness to pay for discrimination. Section II.B.1 explains how we estimate $Price_i$, i.e. the cost of discrimination each individual decision maker faced, from the productivities of the two workers in round one. Section II.B.2 explains how we use variation in $Price_i$ and the variation in observed choices across decision makers to estimate the demand for discrimination, i.e. how decision makers on average respond to changes in *Price*, and the distribution of willingness to pay for discrimination.[14]

A. Discrete Choice Framework

Assume that decision maker *i*'s utility for choosing *same* can be written as:

$$U_i\ (same) = a_i - b\ Price_i + \varphi_i \tag{1}$$

where a_i is the decision maker's animus towards working with a person of the other ethnic type, b is the response to the $Price_i$ an individual faces for choosing the less productive worker of the same type, and φ_i is a mean-zero error term.

The error term captures effects that may influence choices on top of the decision maker's animus. In choice situations in "the wild", i.e. in non-experimental situations, many factors may play a role. An important advantage of studying discrimination in our controlled experimental setting is that we can exclude most of these factors by design. However, choices in our experiment may still be driven by factors we cannot fully control or neutralize by randomization, particularly a preference to work on a particular day. As explained in section I, the choice between the two candidates is presented to decision makers as a choice between two weekdays. This design choice has the advantage of minimizing Hawthorne effects but it potentially introduces a bias if decision makers have preferences over days on which to work. In section IV.A, we present the procedural aspects of our design

[14] In terms of methodology, our approach is similar to (Mas and Pallais 2016) who experimentally vary wages to elicit preferences for alternative work arrangements.

to address day preferences and results from a follow-up treatment called NoName showing that day preferences do not drive our results. Given these results, we can safely ignore φ_i in eq. 1 and decision maker *i*'s utility for choosing *same* is then reduced to:

$$U_i\ (same) = a_i - b\ Price_i \qquad (2)$$

In eq. (2), a_i reflects animus, i.e. the utility the decision maker gets from working with *same* rather than *other*. This utility can be translated into a monetary equivalent which corresponds to a maximum willingness to pay (WTP_i) to work with *same* (i.e. to avoid *other*).

If choices are utility-maximizing and the decision maker knows $Price_i$, decision maker *i* reveals to have a (sufficiently strong) "taste" for working with *same* $a_i \geq bPrice_i$ if he chooses *same*.[15] In this case, we say the decision maker engages in taste-based discrimination (*Discr* = 1). Conversely, the decision maker reveals to have $a_i < b\ Price_i$ if he chooses *other*, and we say the decision maker does not discriminate (*Discr* = 0). Analogously, the decision maker reveals to have $WTP_i \geq Price_i$ if he chooses *same* and reveals to have $WTP_i < Price_i$ if he chooses *other*. Given a distribution of animus *a* in the sample, utility maximization implies that fewer decision makers prefer to discriminate as its price increases. In other words, the demand for discrimination is downward-sloping and our main interest is in estimating the slope *b*.

We estimate the demand for discrimination from the variation in $Price_i$ (which is assigned exogenously to each decision maker) and the variation of choices across decision makers. Our estimate shows for each *Price* the probability that a particular

[15] Below, we use the estimation results from the team production function to calculate *Price*. This procedure assumes that decision makers know the team production function. Appendix E shows that our results are robust to this assumption. In particular, Appendix E shows that using raw productivity differences between candidates in round 1 as a proxy for *Price* yields the same qualitative results as those reported in table 3.

decision maker discriminates (or alternatively, the share of discriminators who have $WTP_i \geq Price_i$). Given parametric assumptions about the CDF of *WTP*, we can back out the distribution of *WTP* from observed choices by use of maximum-likelihood estimation.

B. Econometric Analysis

1. Estimating the Price of Discrimination.—We define the price of discrimination to the discriminator in Info as earnings foregone by choosing a less productive co-worker of the same ethnic type rather than a more productive worker of the other ethnic type. To measure this price, we estimate a *team production function* showing how productivities in round 1 map into output of ethnically homogeneous and heterogeneous teams in round 2. We then estimate for each decision maker the marginal product of labor for the two candidates. This analysis yields the important result that there is a positive price to pay for choosing *same* (see section 4.1).

We estimate the team production function using all observations of workers who completed both rounds[16] as

$$\ln\left(Y_{i,j}^{i}\right) = \beta_0 + \beta_1 \cdot \ln x_i + \beta_2 \cdot \ln x_{j,j\neq i} + \beta_3 \cdot \ln x_i \cdot Alone + \gamma \cdot \mathbf{X} + \varepsilon_i , \tag{3}$$

where $Y_{i,j}^{i}$ is worker *i*'s share of the team output in round 2 when working with co-worker *j*. We estimate team production as a function of worker *i*'s own production in round 1, x_i, the production of the co-worker *j* in round 1 (x_j), an interaction term to capture different learning effects when working alone (*Alone* = 1 and $x_j = 1$ if *i* is working alone in phase 2), and a vector of variables characterizing the team composition (e.g. by ethnic type).

[16] In total, 140 workers completed both rounds according to the description in section 3. Observations from teams with workers having names which do not fit either ethnic type are excluded from our regression.

The price of discrimination is defined as earnings foregone by choosing to work with *same* rather than *other* in treatment Info. This price is not directly observed in our experiment because the decision maker only works with the chosen candidate but not with the non-chosen candidate. We thus have to estimate the counterfactual. The price of discrimination is then the difference between decision maker *i*'s earnings with *other* minus the earnings with *same*[17]

(4) $$\text{Price}_i = p(\hat{Y}^i_{i,other} - \hat{Y}^i_{i,same}) > 0 \ \forall \ i \,.$$

2. Estimating the Demand and Willingness to Pay for Discrimination.—To estimate the demand for discrimination, we model the probability that a decision maker chooses *same* as:

(5) $$\Pr\left(Discr_i = 1 \mid \mathbf{X}\right) = F\left(\mathbf{X}'\beta + \varepsilon_i\right),$$

where F is a CDF and $\mathbf{X}$ is a vector of variables consisting of the price of discrimination (see eq. 4) and gender and ethnicity as additional controls. We use maximum likelihood and assume a probit model to estimate equation (5).[18] The demand for discrimination then shows how likely decision makers on average are to discriminate given that a particular price is assigned to them.

Our estimation of demand for discrimination and the WTP to discriminate are related as follows (where we include only $Price_i$ as an explanatory variable):

[17] The price of discrimination expressed in Euro is obtained by multiplying the difference in output with *p* which is the product of the piece rate (DKK 4 per letter) and the exchange rate (0.13 Euro per DKK).

[18] We report probit estimates throughout the paper. Logit regressions yield qualitatively similar results. We also ran linear regressions with the appropriate finite-sample correction (see Imbens and Kolesar 2016), and probit regression with inference based on permutation. We obtain qualitatively similar results in all cases.

$$
\begin{aligned}
\Pr(Discr = 1 \mid \text{Price}_i) &= \Pr(WTP_i \geq \text{Price}_i) \\
&= 1 - \Pr(WTP_i < \text{Price}_i) \\
&= 1 - \text{F}(\text{Price}_i)
\end{aligned} \tag{6}
$$

We use the estimates $(\hat{a}, \hat{b})$ from eq. (5) to obtain the distribution of the willingness to pay:

$$
\text{F}_{WTP}(x) = 1 - \Phi(\hat{a} + \hat{b} \cdot x), x \in \Re \tag{7}
$$

Note that by experimental design $Price_i > 0$. As a consequence, predictions for $Price_i \leq 0$ are out of sample and should therefore be interpreted with caution. This is particularly true if there is a mass point at $Price_i = 0$ which would be the case if some participants are indifferent in terms of ethnic preference for co-workers. Availability of price data at positive, zero and negative values would have allowed us to estimate a breakpoint model (as in Mas and Pallais 2016) and thereby identify any discontinuities in the distribution of *WTP*.

III. Results

A. The price of discrimination

The price of discrimination is estimated from a team production function which maps individual outputs in round 1 into team outputs in round 2 (see section 3 for explanations). Before discussing the estimation of the team production function, we provide some descriptives.

Figure 1 shows a scatterplot of worker *i*'s share of production in round 2 (i.e. half of the envelopes jointly stuffed) by production in round 1 (i.e. envelopes stuffed in isolation). Black diamonds represent individuals in heterogeneous teams (52 individuals) and white diamonds represent individuals in homogenous teams (36 both Danish-sounding, 32 both Muslim-sounding). Crosses represent individuals working alone in round 2 (20 individuals) because they or their partner did not show up on time.

The figure shows that there is considerable variation in both round 1 production (average 107, sdv = 24) and in round 2 (average 115, sdv = 24). As expected by virtue of random treatment allocation, decision makers' distributions of round 1 production are not different across treatments (*p* = 0.528, Kolmogorov-Smironov test). Workers with Danish-sounding names tended to be more productive in round 1 than those with Muslim-sounding names (116 vs. 100, *p* = 0.000, Mann-Whitney test).[19] This finding has important implications for our analysis in both treatments and is discussed in detail below.

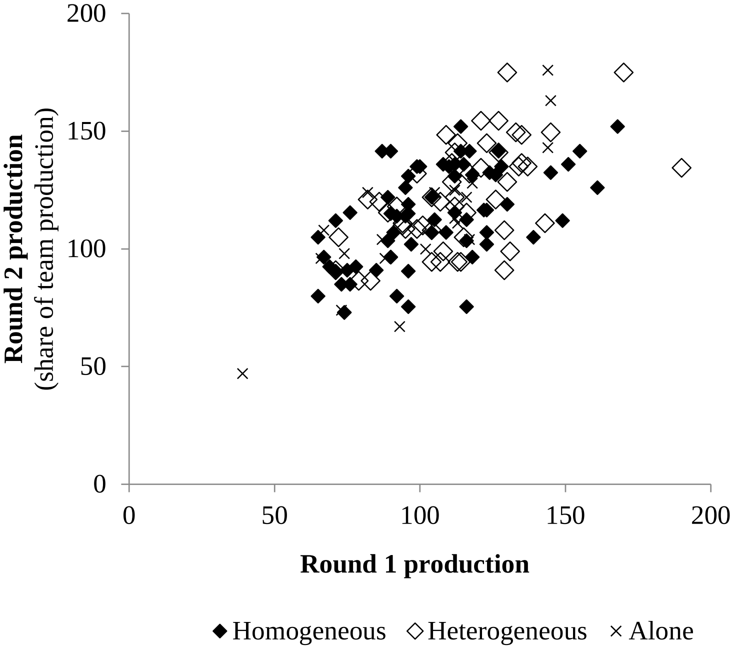

FIGURE 1. PRODUCTION IN ROUND 1 AND ROUND 2

[19] While there are clear productivity differences across ethnic types, the two types were similarly "risky" in terms of variation of output (see Aigner and Cain 1977). The standard deviation for workers with Danish-sounding and Muslim-sounding names is 22.6 and 24.5, respectively.

Notes: The figure shows the number of envelopes stuffed in isolation in round 1 and the share of envelopes stuffed in round 2 for individuals who worked in round 2 in homogeneous teams (black diamonds, n = 68), in heterogeneous teams (white diamonds, n = 52), or alone (crosses, n = 20). The share is 50% of team output for those working in teams, and 100% of individual output for those working alone.

1. Team Production Function.—Table 2 shows various estimates for the team production function. The positive coefficients in the first two lines are all significant ($p < 0.01$) and show that teams tend to be more productive if their members have high productivity in round 1. The coefficients in the third line (all $p < 0.01$) reflect learning by those working alone in round 2. These coefficients are very similar in size to the previous ones suggesting that there is no gain from specialization in our task since those who happened to work alone are on average equally productive as those working in teams.[20] The significant coefficient for *Male* ($p < 0.01$) shows that males are about 6 percent more productive than females in round 2. Taken together, round 1 output explains a considerable share of variation in team output (adjusted R^2 is about .61 in all specifications) which implies that the information available to decision makers is an excellent predictor for the price of discrimination.

[20] Average earnings are the same whether working alone or in a team in round 2, holding everything else constant (p = 0.573, t-test). The coefficients in the first three lines of specification A are very similar because the share of team output for worker i and j is the same (one half) by definition and the share for a worker i working alone is estimated assuming a team mate j with the same round 1 production as worker i.

TABLE 2—TEAM PRODUCTION FUNCTION

Dependent variable: ln(prod$_{2i}$)	(A)	(B)	(C)	(D)
ln(prod$_{1i}$)	0.416	0.408	0.421	0.419
	(0.044)	(0.044)	(0.050)	(0.051)
ln(prod$_{1j}$)	0.416	0.426	0.421	0.428
	(0.044)	(0.045)	(0.050)	(0.050)
ln(prod$_{1i}$) · Alone	0.416	0.424	0.324	0.327
	(0.044)	(0.044)	(0.107)	(0.109)
Male	0.064	0.063	0.064	0.064
	(0.022)	(0.022)	(0.023)	(0.023)
Decision maker		-0.018		-0.017
		(0.024)		(0.030)
Alone			0.452	0.468
			(0.545)	(0.549)
Danish-sounding team			0.037	0.041
			(0.025)	(0.033)
Muslim-sounding team			-0.019	-0.010
			(0.035)	(0.039)
Decision maker · Heterogeneous				0.012
				(0.045)
Constant	0.841	0.843	0.785	0.768
	(0.219)	(0.220)	(0.315)	(0.317)
Adj. R^2	0.611	0.610	0.615	0.610
N	140	140	140	140

Notes: Dependent variable is (the logarithm of) the number of envelopes stuffed in round 2 by worker *i* if working alone (n = 20) or, if working in a team (n = 120), half of the number of envelopes stuffed by *i*'s team. *prod*$_{1i}$ is the number of envelopes stuffed in round 1 by worker *i*, *prod*$_{1j}$ the number of envelopes stuffed by *i*'s co-worker in round 2. *Alone* is a dummy set to 1 if worker *i* works alone in round 2, *Male* is worker *i's* gender, *Decision maker* indicates if worker *i* makes a choice of co-worker. The remaining dummies characterize team composition in round 2. Numbers in parentheses are robust standard errors.

Model B adds the dummy variable *Decision maker* to address the concern that selection effects may have driven the results. For example, those who "cannot work well" with people of the other type may systematically select into homogeneous couples and those who can may select into heterogeneous couples. Our experimental design allows us to address this concern because one half of the subjects had no choice to make and were, from their perspective, forced into teams. We find that decision makers (after controlling for individual productivities) do not have significantly different productivity from those who are forced into the team. This is true for decision makers in general as well as for decision makers selecting into heterogeneous teams (see insignificant interaction term *Decision maker · Heterogeneous* in model D). Models C and D add dummies for team composition to test if ethnically homogenous teams are more productive than heterogeneous teams (which is the reference category in the regression). The insignificant estimates show that the team production function is not type-specific. That is, given individual productivities, heterogeneous teams are equally productive as homogeneous teams.[21]

Taken together, the estimates on the production function show that much of the variation in team production is explained by one's own productivity and the productivity of the co-worker (which are both known to the decision maker in Info when making the choice), but essentially nothing is explained by the ethnic type of the co-worker. This finding is important because it implies a monetary incentive to choose *other* in treatment Info. In other words, there is a price to pay for discrimination, and decision makers had all the required information to know the price.

2. The Price of Discrimination.—To estimate $Price_i$, i.e. the cost of discrimination, we use specification A in table 2 because all variables included in the other models are insignificant. $Price_i$ ranges from 0.65€ to 21.39€ with a mean of 6.70€ and a standard deviation of 4.72€. The 25th and 75th percentile are 2.98€ and 9.24€, respectively. We find that the distribution of $Price_i$ (mirrored on 0) is normal (p = 0.818, Shapiro-Wilk; p = 0.721, Shapiro-Francia; p = 0.901, Skewness/Kurtosis test for normality), as is expected by virtue of random sampling of candidates.

[21] This is a surprising finding in the light of the literature. For example, (Hjort 2014) studies groups of three people who package flowers in a Kenyan plant. One upstream ''supplier'' prepares flowers that are passed on to two downstream ''processors'' who assemble the flowers into bunches. Hjort finds that suppliers (who are paid according to total team output) were "willing to accept lower own pay to lower the pay of non-coethnic co-workers" (p. 1902). Hjort finds that vertically mixed teams are 8 percent less productive than homogenous teams. (Lyons 2017) finds in an online experiment that teams of programmers composed of people with different nationalities are less productive than homogeneous teams.

B. Discrimination in Treatment Info

We observe that 38 percent of decision makers in treatment Info choose to discriminate, i.e. choose *same*. This result is novel since we show that taste-based discrimination is common even when decision makers face a positive and known price of discrimination.

A first finding supporting our claim that higher (randomly assigned) prices causally reduce discrimination is that discriminators face lower prices on average than non-discriminators (€4.9 vs. €7.8). Both a Kolmogorov-Smirnov test (KS, p = 0.091) and Wilcoxon rank-sum test (p = 0.052) show that prices are different for the two groups (see appendix G for tests showing that prices are randomly assigned). The average expected price of €4.9 for discriminators may seem low in absolute terms but is strikingly high in relative terms. For example, the average discriminator gives up 8 percent of round 2 earnings to work with *same* for 90 minutes.

1. The Demand for Discrimination.—Table 3 presents the main results for treatment Info. Model (1) provides the most parsimonious specification showing that the law of demand holds for taste-based discrimination. The table shows marginal effects from probit regression and the significant ($p < 0.05$) coefficient on *Price* therefore shows that discrimination falls by 3.6 percent if the price of discrimination goes up by €1. Note that this number is our best estimate for the average marginal change. Due to the non-linearity of the demand relation, this marginal effect is not informative for larger changes in cost. We provide estimates for such changes in the discussion of figure 2.

Model (2) adds dummy variables for gender (*Male*) and ethnic type (*Danish-sounding*) of the decision maker. The insignificant estimate on *Danish-sounding* indicates that the tendency to discriminate is not different across ethnic types, after controlling for differences in prices. We think that this is a remarkable result for two reasons. First, attention both in the literature and policy debates usually focuses on discrimination of the minority group by the majority group because members of the majority group are more often in the position to discriminate, and workers from the minority group tend to be disadvantaged. However, our results suggest that observing more frequent discrimination of minorities may simply be due to the fact that majority decision makers have more opportunities to discriminate rather than a stronger ethnic animus.

Second, this result highlights the importance of controlling for prices when measuring discrimination. From simply looking at discrimination percentages, a layperson may be misled to conclude that decision makers with Danish-sounding names are more likely to discriminate. In fact, decision makers with Danish-sounding names discriminate in 44 percent of the cases, while those with Muslim-sounding names do so in only 33 percent of the cases (however, $p = 0.517$, χ^2 test). Yet, these differences do not reflect differences in animus because decision makers with Danish-sounding names face a lower price on average than decision makers with Muslim-sounding names (€5.2 vs. €7.8, $p = 0.078$, KS). The reason is that workers with Danish-sounding names are systematically more productive (116 letters) in round 1 than participants with Muslim-sounding names (100 letters). According to regressions (2) and (4) in table 3, these price differences explain the observed differences in taste-based discrimination across ethnic types (*Danish-sounding* is insignificant, but *Price* is significant).

TABLE 3. THE DEMAND FOR DISCRIMINATION

Dependent variable: Discr	(1)	(2)	(3)	(4)
Price	-0.036	-0.035	-0.034	-0.038
	(0.016)	(0.017)	(0.016)	(0.020)
Danish-sounding		0.020		-0.045
		(0.160)		(0.286)
Male		-0.056		-0.022
		(0.152)		(0.284)
Danish-sounding · Price			0.005	0.011
			(0.022)	(0.040)
Male · Price			-0.007	-0.004
			(0.018)	(0.036)
R^2	0.082	0.085	0.086	0.087
N	37	37	37	37

Notes: The table shows average marginal effects estimated from Probit regressions. Numbers in parentheses are robust standard errors. *Discr* = 1 for a decision maker choosing *same* and 0 otherwise. *Male* and *Danish-sounding* are dummies characterizing the decision maker.

Model (3) adds the interaction terms *Danish-sounding · Price*, and *Male · Price*. The respective estimates are insignificant, suggesting that responses to changes in price are not different across ethnic types and gender. Model (4) combines (2) and (3) and yields the same results. Appendix E shows that our results are robust to using alternative types of team production functions to estimate prices, and appendix F shows robustness to issues relating to the absolute and relative productivity of the decision maker.

Figure 2 summarizes our main finding. The solid line shows that decision makers respond strongly to changes in prices in Info. For example, increasing the price of discrimination by one standard deviation from the average (i.e. from €6.7 to €11.4) reduces the probability of discrimination by 45 percent (from .36 to .20). Conversely, decreasing the price by one standard deviation from the average (i.e. from €6.7 to €2.0) increases the probability by 54 percent (from .36 to .55). Another way to describe the price-responsiveness is to estimate an elasticity which indicates the percentage decrease in the probability to discriminate in response to a 1% increase in price. Our best estimate is -0.9. This elasticity is an average of all elasticities, evaluated at each observation. In conclusion, we find that the demand for taste-based discrimination is downward-sloping and is highly elastic.

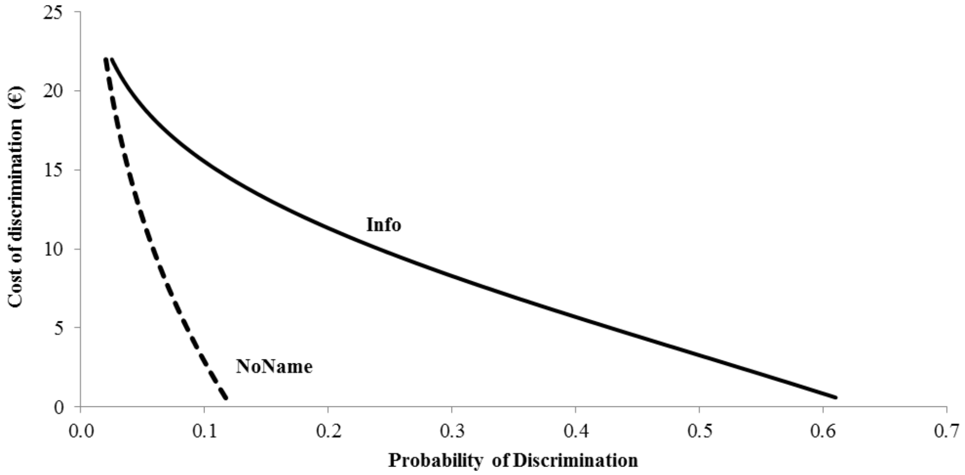

FIGURE 2. THE DEMAND FOR DISCRIMINATION

Notes: The figure shows the relation between the probability of discrimination (choosing *same*) and the price of discrimination in Euro in treatment Info (solid line) and NoName (dotted line), calculated using specification (1) in table 3.

2. Willingness to Pay for Discrimination.—Using the estimation strategy described in section II.B.2, we find that the average decision maker in our sample is willing to pay at least μ_{WTP} = €3.3 to work with *same* rather than *other* (σ_{WTP} = €9.6). The 25th and 75th percentile of the WTP distribution is -€3.2 and €9.7, respectively. To the best of our knowledge, we are the first to back out the mean and variance of a (lower bound of) the willingness to pay for animus-based discrimination from field experimental data.[22]

[22] Related studies include (List 2004) who shows final offers made at a sportscard market are 18-20% lower to blacks than to whites, (Doleac and Stein 2013) who find that offers to black ipod sellers on craigslist are 11% lower than those to white sellers, and (Zussman 2013) who finds that Arab buyers of used cars need to quote prices that are 5-10% higher to obtain the same probability of response from the seller as an Israeli buyer. However, we are not aware of any study providing credible estimates of the distribution of the willingness to pay.

IV. Interpretation of Results

The results above clearly show that decision makers tend to choose co-workers of the same ethnic type even when this is costly to the decision maker, but that this tendency is greatly reduced as the price of doing so goes up. Our estimate of price responsiveness results from heterogeneous animus in our sample as we observe only one choice per decision maker. Hence, as the price of choosing *same* goes up, fewer decision makers will have a sufficiently strong animus great enough to warrant choosing *same* and the probability of discrimination decreases. We interpret our findings as an indication that taste-based discrimination is highly responsive to the price of discrimination.

This section addresses three concerns with this interpretation. First, because we frame the choice between candidates as a choice of days, the result can be interpreted as an indication that decision makers have a preference to work on a particular day of the week but increasingly choose the other day as doing so becomes more costly. In terms of eq. (3), this means that ε_i can either be interpreted as pure noise or a day preference plus noise. Section IV.A addresses this concern with a follow-up treatment called NoName in which taste-based discrimination is ruled out because the ethnicity of the candidates is unknown but in which day preference could drive choices. Second, rather than measuring animus, the estimate on *WTP* may reflect prejudiced beliefs on productivity. Section IV.B addresses this concern with a follow-up study in which we elicit beliefs on a sample of similar juveniles. Third, we show that our estimate of willingness to pay from section III.B has good predictive power out of sample and performs better in explaining observed behavior than statistical discrimination. Section IV.C addresses this concern by means of treatment NoInfo.

A. Preference for Work Days?

The downward-sloping curve labeled Info in figure 2 can be interpreted in terms of a day preference as follows. Suppose decision makers have no ethnic preference whatsoever but they have an unobserved preference to work on a particular day (the earlier date, say). Because of random allocation of types to work days such decision makers will choose *same* in 50 percent of the cases when the price is zero (i.e. whenever *same* happens to be on the first day), but choose *same* at lower rates as the price goes up (because they trade off their preference for the earlier date against the price). If decision makers had a day preference but no ethnic preference, we would thus see a falling demand for *same* with increasing prices. The rest of this section argues that this interpretation is not consistent with the data.

Three procedural aspects of our experiment make it implausible that day preferences are a driving force behind the patterns of choice we observe in Info. First, at the end of round 1 all participants indicated on which weekdays they will be available for work in phase 2. At that time, they did not know that they will later have a choice of which day of the week to work on. Therefore, a decision maker with a preference against working on a particular weekday would rationally not indicate that day. In addition, when decision makers were called on the phone before round 2, they had to confirm their availability on the days they had indicated earlier. Again, confirmation came before knowing that they will have a choice on which day work. Second, Friday was not in the choice set. Subjects with Muslim-sounding names may have a preference against working on Fridays relative to other weekdays for religious reasons. Third, our subjects were highly available and flexible with respect to working days. For example, 95 percent (n = 169) of participants indicated that they were available on at least 3 (out of 4) days in round 2.

To provide a direct test for the relevance of a day preference in worker choice, we ran a treatment called NoName with n = 51 decision makers. The structure of NoName is the same as treatment Info insofar as subjects are recruited to stuff envelopes for a large mailing and are paid at a piece rate. The main difference to Info is that the decision maker *does not know the name* (but does know the productivity) of the candidates when making the choice in NoName.[23] The treatment rules out (taste or statistical) discrimination as a motivation for choice. But choosing the less productive candidate is consistent with a day preference.

We find that 92 percent (= 47/51) of decision makers choose the more productive candidate. This share is different from both the distribution of choices in Info and from 50/50 choices (both tests: $p < 0.001$, χ^2 test). An alternative interpretation is that 8% of decision makers made "wrong" choices because they were inattentive. This share is indeed similar to the 13% of people who make "wrong" (i.e. dominated) choices in the labor market experiment by Mas and Pallais (2016). The finding from NoName indicates that choices were clearly guided by the relative profitability of the candidates and not by a day preference.

A probit regression of choice of the less productive worker on *Price* and a constant as in model (1) in table 3 reveals an insignificant coefficient on *Price* of -0.040 (sdv = 0.056, p = 0.474) and a significant constant of -1.165 (sdv = 0.425, p = 0.006). The result of that regression is shown in figure 2, see dotted line. The significant negative constant shows that decision makers were systematically more likely to choose the more productive over the less productive candidate, and the insignificant coefficient on *Price* shows that these choices did not depend on *Price*. This non-reaction to *Price* together with the fact that almost all decision makers

[23] NoName also differs in a number of other ways from Info. For example, the work task was shorter and more complex in NoName and average output in phase 1 was therefore lower in NoName (45.6 in 60') than in Info (106.8 in 90'). The subject pool is different since NoName was conducted at the University of Vienna with a total of 51 students from all fields, only 4 of which happened to have Muslim-sounding names. See Appendix J for a detailed description.

choose the more productive candidate means that φ_i in eq. (1) reflects noise and not a day preference. In other words, decision makers' choices in Info were not driven by a preference for days.[24]

B. Prejudiced Beliefs on Productivity?

This section addresses the concern that our estimates in section III.B do not reflect animus but a belief that the production function is type-specific. The concern here is as follows: decision makers may choose *same* over *other* not because they dislike *other* but because they think that working with *other* is less productive than working with *same* for given individual productivities. Section III.A has shown that the production function is in fact not type-specific. That is, decision makers had no material reason to choose *same* over *other*. Now, we argue that on average, they did not think that there was a material reason to choose *same* over *other*.

Decision makers can be said to have prejudiced beliefs if they falsely believe that collaborating with a worker of the other type is less productive than collaborating with a worker of the same ethnic type. Unfortunately, simply asking decision makers whether they are prejudiced in this way is unlikely to yield meaningful results.[25] As a consequence, we elicit beliefs about the average price of discrimination indirectly, from a sample of similar juveniles.

[24] Additional evidence for this conclusion comes from treatment NoInfo which is identical to Info, except that the price of choosing *same* is known in Info but is not known in NoInfo. The overall rate of choosing *same* is about twice as high in NoInfo than in Info (78 vs. 38 percent, $p = 0.001$, χ^2 test). This difference is an impressive demonstration that knowing about the price of discrimination makes a big difference for discrimination choices. Furthermore, decision makers with Muslim-sounding names almost all choose *same* (92%) in NoInfo, but they all should have chosen *other* if we assume that they have no animus, no day preference, and hold correct beliefs. The reason is that workers with Danish-sounding names are much more productive in round 1 than those with Muslim-sounding names (116 vs. 100 letters stuffed). This extreme choice pattern can easily be explained by an ethnic preference for *same* (i.e. animus against *other*) but not by a day preference.

[25] People tend to be more prejudiced than they admit. For example, Kawakami et al. (2009: 277) show "that people's predictions regarding their emotional distress and behavior in response to a racial slur differ drastically from their actual reactions". Studies using survey-based data on animus may therefore yield lower-bound estimates for animus. For example, Charles and Guryan (2008)) use 21 survey questions to argue that animus-based prejudice explains up to 25 percent of the racial wage gap in the US. Another reason for not asking about beliefs in NoInfo is to implement a ceteris paribus variation compared to Info where asking about beliefs would have undermined the important property that participants do not know that they are in an experiment.

We recruit n = 353 participants with Danish-sounding and Muslim-sounding names from secondary schools on the outskirts of Copenhagen where we did not recruit for treatments Info and NoInfo. We carefully describe the work task to participants and elicit beliefs about the productivity of individuals and teams. In particular, each participant is presented with the names of 7 randomly selected workers and 6 randomly selected teams, all of the same gender as the participant. Participants are asked to guess how many envelopes each worker stuffed in round 1 and each team stuffed in round 2 (see appendix H for details). Participants are rewarded for guessing correctly (using a quadratic scoring rule).

We find that participants have qualitatively correct beliefs in the sense that they believe that individual workers with Danish-sounding names stuff more envelopes on average than workers with Muslim-sounding names (p = 0.004, Wilcoxon signed-rank test, WSR). However, average beliefs are quantitatively biased since the true difference across types of workers is larger than the expected difference (16 vs. 3 letters).[26] In other words, participants tend to underestimate the true productivity difference across types. Consistent with the belief that workers with Danish-sounding names are individually more productive, we find that teams with more Danish workers are believed to be more productive.

Importantly, we find no evidence for the claim that the team production function is thought to be type-specific. Our analysis in table 2 has shown that, after controlling for individual productivity, heterogeneous teams in fact are as productive as homogeneous teams. Our analysis (see appendix H) shows that participants *do not think* that workers earn more in a homogeneous team than a heterogeneous team, for given round 1 outputs. Put differently, neither do the

[26] We reject the hypothesis that the median person believes the difference to be equal to the true difference (p = 0.000, WSR). This result also holds for each ethnic group separately (p = 0.000, WSR).

juveniles believe nor do they have a reason to believe that selecting a co-worker of the same type is more profitable for given productivities of workers.

C. Explanatory Power of Estimated Animus vs. Statistical Discrimination

This section provides a comparative evaluation of the explanatory power of our estimate of animus from section III.B vs. statistical discrimination in a situation where they both can matter, i.e. in a treatment called NoInfo. This treatment is identical to treatment Info except that decision makers *do not know* how productive the candidates are when choosing between them. Therefore, they must form beliefs about productivity of types to make a rational choice.

The standard theory of statistical discrimination assumes that decision makers form rational (i.e. on average correct) beliefs and that decision makers have no animus.[27] Our experiment provides a rare opportunity to test the predictions of statistical discrimination because we can retrieve *rational beliefs* from the distribution of workers' output in round 1 as follows. For each decision maker *i*, we sample observed round 1 output of two candidates of different types. We estimate the marginal product of labor (MPL) for *i* with either candidate using model A from table 2. Decision maker *i*'s price of choosing one worker over the other is the difference between these MPLs. By repeatedly sampling and averaging, we obtain the expected price for *i* of choosing one type over the other (see Appendix I for details). Because workers with Danish-sounding names are on average more productive than those with Muslim-sounding names in our sample (116 vs. 100 envelopes stuffed in round 1), statistical discrimination predicts that all decision makers choose the worker with the Danish-sounding name. However, only about

[27] Altonji and Pierret (2001: 316) explain that they "are using the term 'statistical discrimination' as synonymous with the term 'rational expectations' in the economics literature. We mean that in the absence of full information, firms distinguish between individuals with different characteristics based on statistical regularities. That is, firms form stereotypes that are rational given their information."

half of decision makers (49%) in fact do so. Statistical discrimination blatantly fails to predict choices of decision makers with Muslim-sounding names (only 10.5 percent choose *other*).

To evaluate the predictive power of animus given rational beliefs, we use rational beliefs as described above and feed those beliefs into our estimate of taste-based discrimination (see model 1 in table 3) and estimate the probability that decision maker *i* chooses *same*. By doing so, we assume that the distribution of animus-driven prejudice is the same in treatment Info and NoInfo. This assumption is warranted since decision makers were randomly allocated to treatments.

Taking animus-driven prejudice into account improves the prediction for the decision makers with Danish-sounding names from 100 to 79.1 percent. This prediction is not statistically different from the observed 66.7 percent (p = 0.711, Fisher exact test).[28] The prediction for the decision makers with Muslim-sounding names is also improved. Now, 57.3 percent (rather than 100 percent) are predicted to choose *other*. Yet, the prediction is still different from the observed 10.5 percent (p = 0.013, Fisher exact test).

We can also evaluate the comparative performance of taste-based discrimination as estimated above vs. statistical discrimination in terms of earnings. Statistical discrimination predicts that decision makers make earnings-maximizing choices given the limited information available.[29] We can now calculate how much money decision makers left on the table by deviating from statistical discrimination as a measure of misprediction. We calculate this gap between actual earnings and earnings under statistical discrimination in percent of decision makers' round 2 earnings with statistical discrimination. The total gap is 3.6 percent (or about €2.3

[28] Tests in this section assume an equal number of observations for predicted and observed discrimination rates.

[29] While statistical discrimination maximizes expected earnings absent precise information about candidates' productivities, it does not yield the first-best outcome. Losses may occur when type-productivity distributions overlap. In our setting, the loss due to limited information is 2.5 percent of round 2 earnings. Yet, there is a clear incentive for statistical discrimination in NoInfo. In fact, earnings are 2.5 percent higher with statistical discrimination than with random choice.

per decision maker). The gap is smaller for decision makers with Danish-sounding names (1.6 vs. 5.8 percent) because they tend to choose the Danish-sounding, i.e. on average more productive, candidate more often. We find that our estimate of animus *cum* rational beliefs predicts much better and therefore implies a smaller earnings gap of 1.7 percent relative to the benchmark. That is, animus *cum* rational beliefs explains observed behavior much better than statistical discrimination. We find that our estimate of animus from treatment Info explains about 47 percent (= 1.7/3.6) of the income gap between earnings with statistical discrimination and observed earnings.[30]

V. Concluding Remarks

This study develops a novel experimental approach to measuring how discriminators respond to changes in the price to be paid for ethnic discrimination in the workplace. As in conventional experimental approaches like audit and correspondence tests, discriminators are not aware that they are participating in an experiment. In contrast to these conventional experimental approaches, ours involves real workers who produce actual work output. We measure their productivities and observe actual outcomes of discrimination choices in a tightly controlled setting. Our design allows us to assign a price to discriminators (earnings forgone from discriminating against a more productive worker) randomly.

Using a sample from Denmark, we find that discrimination is common even at a substantial price, that majority and minority groups are equally likely to discriminate for given prices, and that the demand for discrimination is highly

[30] The other half of the income gap can be partly explained by biased expectations as elicited in the follow-up experiment reported in section IV.B (see Hedegaard and Tyran 2014 for details). We find that animus *cum* biased beliefs explains about 60 percent of the total gap. We can only speculate what may explain the remaining gap. A possibility is “implicit discrimination” (Bertrand, Chugh and Mullainathan 2005). However, it is not entirely clear why implicit discrimination should be more important for minority than for majority decision makers. We can safely exclude “attention discrimination” (Bartoš et al. 2016) by design since decision makers need not make any effort in evaluating candidates in our simple setting.

elastic. Our best estimate is that the probability to discriminate falls by about 9 percent if the price of discrimination goes up by 10 percent.

In two follow-up studies we show that these observations are not consistent with alternative accounts like a preference for work days or false (type-specific) beliefs on productivity, and we conclude that the most plausible interpretation is that decision makers engage in taste-based discrimination which is highly price-elastic. We also show that our estimate of taste-based discrimination predicts behavior well out of sample, and that it is more successful in accounting for observed behavior than statistical discrimination in a treatment (called NoInfo) where both types of discrimination can matter.

Two potential sources of mismeasurement of prejudice due to selection effects may in principle limit the external validity of our results. First, we may underestimate animus in the general population because our sample is not representative of the Danish population. We recruit juveniles from secondary schools in Copenhagen who have very similar age and education and are all fluent in the majority language. Such relatively well-educated and integrated juveniles as a group may have systematically lower animus than the average Dane or Muslim living in Denmark. In fact, available research suggests that (voiced) animus decreases with education and income but increases with age (e.g. Charles and Guryan 2008).

Second, we may over- or underestimate differences in animus across ethnic types. We find that minority and majority groups are equally likely to discriminate for a given price. This result is surprising in the light of evidence suggesting that minorities have more pronounced "homophily" (in the diction of Currarini, Jackson and Pin 2009) than majorities. We may underestimate the difference due to unobserved heterogeneity in income in our sample. While the evidence presented in Charles and Guryan (2008) suggests that animus decreases with income, taste-based discrimination may well also increase with income (if it is a "normal" good).

However, we may overestimate the difference due to a subtle name-related selection effect. A juvenile is classified as having a Muslim-sounding name in our experiment if his parents chose such a name, but is classified as having a Danish-sounding (or other) name if they did not. If the name choice by parents is correlated with animus, we would tend to overestimate differences in animus across ethnic groups. However, this effect seems to be of minor relevance since we find no difference in animus across ethnic types.

The extent to which the quantitative estimates from our experiment extrapolate to hiring choices, in particular in large firms, must remain an open issue. We may over- or underestimate the sensitivity of ethnic discrimination to its cost because decision makers (in Info) faced a clear and known price for discrimination while incentives may be opaque or weak for a personnel officer in a large firm. Decision makers in our experiment are directly affected (monetarily and non-monetarily) by their choices because they make a consequential choice of whom to work with in a team. In contrast, personnel managers do not necessarily physically work with new hires and may also be largely shielded from monetary consequences of their choices. On the other hand, large corporations may have particular policies (like affirmative action programs) on discrimination in place which may provide incentives to clamp down on discrimination.

We close with a reminder of the well-known asymmetry in empirical research that it is easier to refute a clear-cut and simple theory (like statistical discrimination) than to exactly pin down the drivers of behavior, in particular in a complex setting of treatment NoInfo in which taste-based, statistical and potentially other sources of discrimination can play a role. For example, Ayres and Siegelman (1995: 319) note: “It may be that simple theories of discrimination fail to capture the mutually reinforcing nature of multiple causes. In the end, it may prove impossible to parse out the various elements of animus and rational inferences from irrational stereotypes.” Our paper is no exception to this rule. Our main results that discrimination is common when there is a price to pay, and that choices respond to changes in price are, we think, clear and straightforward. The interpretation of what exactly drives behavior is to some extent debatable in NoInfo. Statistical discrimination does not explain behavior well, and our estimate of taste-based discrimination predicts considerably better out of sample. But some unexplained variation remains to be explained, we hope, in future work.

REFERENCES

Adida, Claire L., David D. Laitin, and Marie-Anne A. Valfort. 2010. “Identifying Barriers to Muslim Integration in France.” *Proceedings of the National Academy of Sciences of the United States of America* 107 (52): 22384–90.

Aigner, Dennis J., and Glen G. Cain. 1977. “Statistical Theories of Discrimination in Labor Markets.” *Industrial and Labor Relations Review* 30: 175–87.

Altonji, Joseph G., and Rebecca M. Blank. 1999. “Race and Gender in the Labor Market.” In *Handbook of Labor Economics* Vol. 3, edited by Orley Ashenfelter and David Card, 3143–59. Amsterdam: North Holland.

Altonji, Joseph G., and Charles R. Pierret. 2001. “Employer Learning and Statistical Discrimination.” *Quarterly Journal of Economics* 116 (1): 313–50.

Arrow, Kenneth Joseph. 1973. “The Theory of Discrimination.” In *Discrimination in Labor Markets*, 3–42, edited by Orley Ashenfelter and Albert Rees. Princeton, N. J.: Princeton University Press.

Ayres, Ian, and Peter Siegelman. 1995. “Race and Gender Discrimination in Bargaining.” *American Economic Review* 85 (3): 304–21.

Baccara, Mariagiovanna, Allan Collard-Wexler, Leonardo Felli, and Leeat Yariv. 2014. “Child-Adoption Matching: Preferences for Gender and Race.” *American Economic Journal: Applied Economics* 6 (3): 133–58.

Bartoš, Vojtěch, Michal Bauer, Julie Chytilová, and Filip Matějka. 2016. “Attention Discrimination: Theory and Field Experiments with Monitoring Information Acquisition.” *American Economic Review* 106 (6): 1437–75.

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

Behaghel, Luc, Bruno Crépon, and Thomas Le Barbanchon. 2015.

"Unintended Effects of Anonymous Resumes." *American Economic Journal: Applied Economics* 7 (3): 1–27.

Bertrand, Marianne, Dolly Chugh, and Sendhil Mullainathan. 2005. "Implicit Discrimination." *American Economic Review* 95 (2): 94–8.

Bertrand, Marianne, and Sendhil Mullainathan. 2004. "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." *American Economic Review* 94 (4): 991–1013.

Charles, Kerwin Kofi, and Jonathan Guryan. 2008. "Prejudice and Wages: An Empirical Assessment of Becker's The Economics of Discrimination." *Journal of Political Economy* 116 (5): 773–809.

Currarini, Sergio, Matthew O. Jackson, and Paolo Pin. 2009. "An Economic Model of Friendship: Homophily, Minorities, and Segregation." *Econometrica* 77 (4): 1003–45.

Doleac, Jennifer L., and Luke C. D. Stein. 2013. "The Visible Hand: Race and Online Market Outcomes." *Economic Journal* 123 (572): F469–92.

Edelman, Benjamin G., Michael Luca, and Dan Svirsky. Forthcoming. "Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment." *American Economic Journal: Applied Economics*.

Gneezy, Uri, and John A. List. 2006. "Putting Behavioral Economics to Work: Testing for Gift Exchange in Labor Markets Using Field Experiments." *Econometrica* 74 (5): 1365–84.

Gneezy, Uri, John A. List, and Michael K. Price. 2012. "Toward an Understanding of Why People Discriminate: Evidence from a Series of Natural Field Experiments." *National Bureau of Economic Research* WP 17885.

Goldin, Claudia, and Cecilia Rouse. 2000. "Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians." *American Economic Review* 90 (4): 715–41.

Grogger, Jeffrey. 2011. "Speech Patterns and Racial Wage Inequality." *Journal* *of Human Resources* 46 (1): 1–25.

Guryan, Jonathan, and Kerwin Kofi Charles. 2013. “Taste-Based or Statistical Discrimination: The Economics of Discrimination Returns to Its Roots.” *Economic Journal* 123 (572): F417–32.

Hamermesh, Daniel S., and Jeff E. Biddle. 1994. “Beauty and the Labor-Market.” *American Economic Review* 84 (5): 1174–94.

Heckman, James J., and Peter Siegelman. 1993. “The Urban Institute Audit Studies: Their Methods and Findings.” In *Clear and Convincing Evidence: Measurement of Discrimination in America,* 187–258, edited by Michael Fix and Raymond Struyk. Urban Institute Press.

Hedegaard, Morten, and Jean-Robert Tyran. 2014. “The Price of Prejudice.” *CEPR* Discussion Paper 9953.

Hjort, Jonas. 2014. “Ethnic Divisions and Production in Firms.” *Quarterly Journal of Economics* 129 (4): 1899–946.

Imbens, Guido W., and Michal Kolesar. 2016. “Robust Standard Errors in Small Samples: Some Practical Advice.” *Review of Economics and Statistics* 98 (4): 701–12.

Kawakami, Kerry, Elizabeth Dunn, Francine Karmali, and John F. Dovidio. 2009. “Mispredicting Affective and Behavioral Responses to Racism.” *Science* 323: 276–8.

Kube, Sebastian, Michel André Maréchal, and Clemens Puppe. 2013. “Do Wage Cuts Damage Work Morale? Evidence from a Natural Field Experiment.” *Journal of the European Economic Association* 11 (4): 853–70.

Levitt, Steven D. 2004. “Testing Theories of Discrimination: Evidence from Weakest Link.” *Journal of Law and Economics* 47 (2): 431–52.

List, John A. 2004. “The Nature and Extent of Discrimination in the Marketplace: Evidence from the Field.” *Quarterly Journal of Economics* 119 (1): 49–89.

Lyons, Elizabeth. forthcoming. “Team Production in International Labor Markets:

Experimental Evidence from the Field." *American Economic Journal: Applied Economics*.

Mas, Alexandre, and Amanda Pallais. 2016. "Valuing Alternative Work Arrangements." *National Bureau of Economic Research* WP 22708.

Mobius, Markus M., and Tanya S. Rosenblat. 2006. "Why Beauty Matters." *American Economic Review* 96 (1): 222–35.

Neumark, David. 2012. "Detecting Discrimination in Audit and Correspondence Studies." *Journal of Human Resources* 47 (4): 1128–57.

Neumark, David, Roy J. Bank, and Kyle D. Van Nort. 1996. "Sex Discrimination in Restaurant Hiring: An Audit Study." *Quarterly Journal of Economics* 111 (3): 915–41.

Phelps, Edmund S. 1972. "The Statistical Theory of Racism and Sexism." *American Economic Review* 62 (4): 659–61.

Pope, Devin G., and Justin R. Sydnor. 2011. "What's in a Picture? Evidence of Discrimination from Prosper.com." *Journal of Human Resources* 46 (1): 53–92.

Riach, Peter A., and Judith Rich. 2002. "Field Experiments of Discrimination in the Market Place." *Economic Journal* 112 (483): F480–518.

Yinger, John. 1998. "Evidence on Discrimination in Consumer Markets." *Journal of Economic Perspectives* 12 (2): 23–40.

Zizzo, Daniel J. 2010. "Experimenter Demand Effects in Economic Experiments." *Experimental Economics* 13 (1): 75–98.

Zussman, Asaf. 2013. "Ethnic Discrimination: Lessons from the Israeli Online Market for Used Cars." *Economic Journal* 123 (11): F433–68.