The Price of Prejudice

By MORTEN STØRLING HEDEGAARD AND JEAN-ROBERT TYRAN

Appendix A: Flyer Used for Recruiting

Figure A1: Flyer Used for Recruiting

Appendix B: Location and Participants

Figure B1: Location of Schools from Which Participants Were Recruited

Figure B2: Control Room

Figure B3: Floor plan

Appendix C: Description of the Work Task

Figure C1: Work Station

Appendix D: Validation of Classification of First Names

Table D1: Effectiveness of First Names as Marker of Ethnic Type

Appendix E: Using Productivity Differences as Proxy for the Price of Discrimination

Table E1: The Demand for Discrimination Using Output Differences as a Proxy for Price

Appendix F: Robustness of Price Effect With Respect to the Decision Maker's Productivity

Table F1: Discrimination and the Decision Maker's Productivity

Appendix G: Testing for Random Assignment of Price (Simulation)

Appendix H: Eliciting Productivity Beliefs

Table H1: Average Output Guesses by Participants in Complementary Study

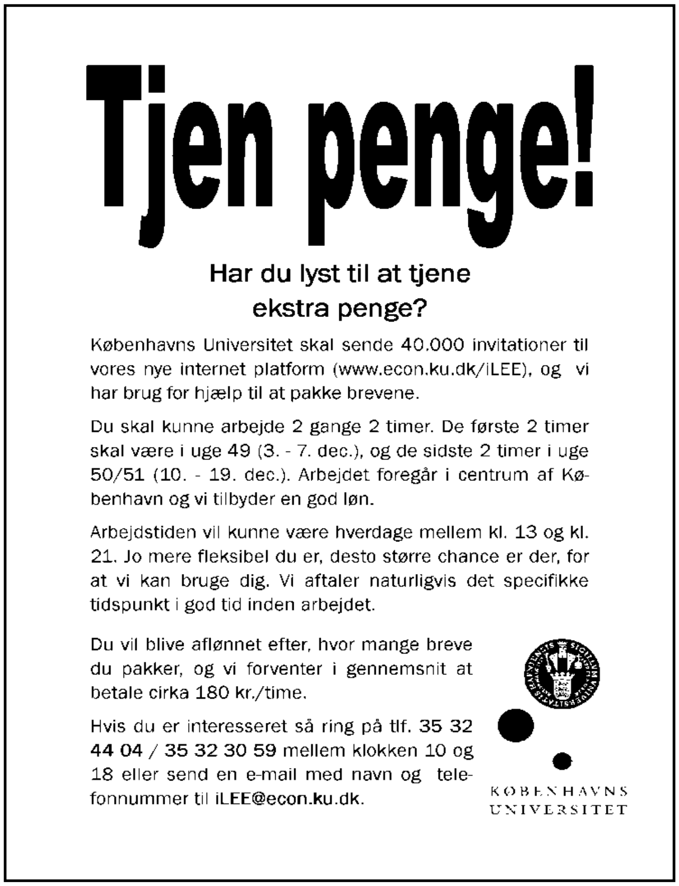
Appendix I: Decomposition of the Earnings Gap

Appendix J: Treatment NoName

Figure J1: Work Station

Figure J2: Control Room

Appendix A. Flyer Used for Recruiting



Tjen penge!

Har du lyst til at tjene ekstra penge?

Københavns Universitet skal sende 40.000 invitationer til vores nye internet platform (www.econ.ku.dk/iLEE), og vi har brug for hjælp til at pakke brevene.

Du skal kunne arbejde 2 gange 2 timer. De første 2 timer skal være i uge 49 (3. - 7. dec.), og de sidste 2 timer i uge 50/51 (10. - 19. dec.). Arbejdet foregår i centrum af København og vi tilbyder en god løn.

Arbejdstiden vil kunne være hverdage mellem kl. 13 og kl. 21. Jo mere fleksibel du er, desto større chance er der, for at vi kan bruge dig. Vi aftaler naturligvis det specifikke tidspunkt i god tid inden arbejdet.

Du vil blive aflønnet efter, hvor mange breve du pakker, og vi forventer i gennemsnit at betale cirka 180 kr./time.

Hvis du er interesseret så ring på tlf. 35 32 44 04 / 35 32 30 59 mellem klokken 10 og 18 eller send en e-mail med navn og telefonnummer til iLEE@econ.ku.dk.

KØBENHAVNS UNIVERSITET

FIGURE A1: FLYER USED FOR RECRUITING

Translation.— Earn money! Would you like to earn some extra money?

The University of Copenhagen has to mail 40’000 invitation letters for a new internet platform, and we are looking for help to stuff these envelopes. You are supposed to work twice for 2 hours. The first 2 hours are in week 49 … the second in week 50/51. Work is to be done in the city center and we pay a good salary.

Work times are between 1 p.m. and 9 p.m. You are more likely to be hired if you are more flexible with respect to work times. We will of course make a specific agreement with sufficient notice. You will be paid according to how many envelopes you stuff and we expect to pay about 180 kr. (about €24) per hour. Call us on …between .. and .. or send an e-mail with your name on phone number to … if you are interested.

Appendix B. Location and Participants

Figure B1 shows the secondary schools from which participants were recruited for the experiment (red symbols), for the belief elicitation and name validation studies (blue symbols) and the pre-test (purple marker in the lower left corner). The flag indicates the location of the University premises where work was carried out.

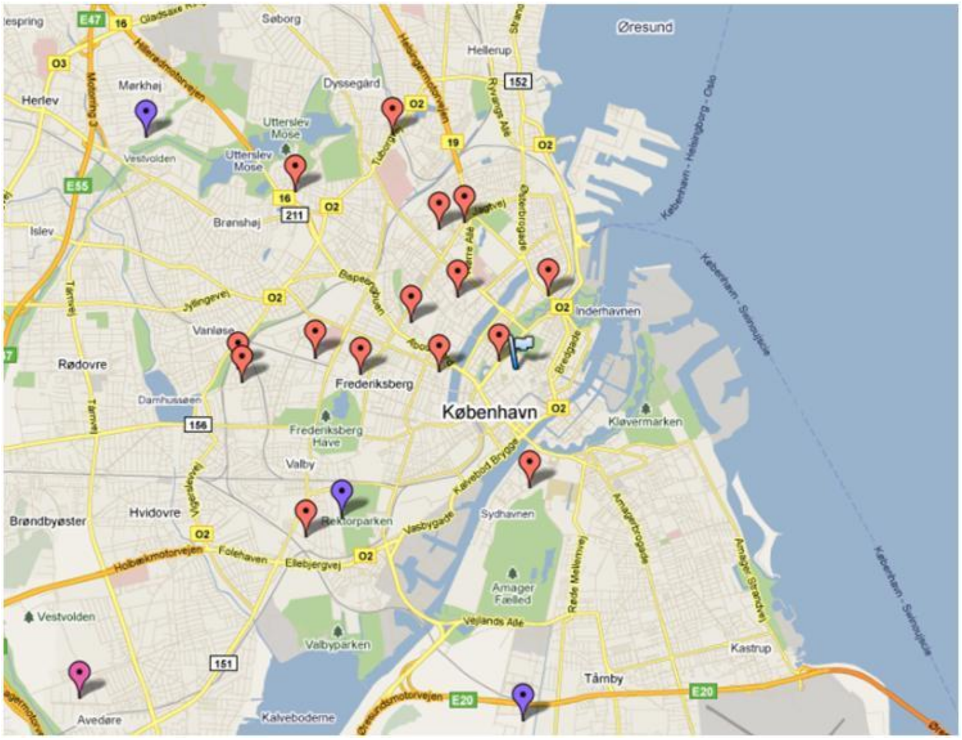


FIGURE B1: LOCATION OF SCHOOLS FROM WHICH PARTICIPANTS WERE RECRUITED[1]

[1] The figure has more than eleven red markers as some of the schools where we recruited for the experiment have several campuses in Copenhagen.

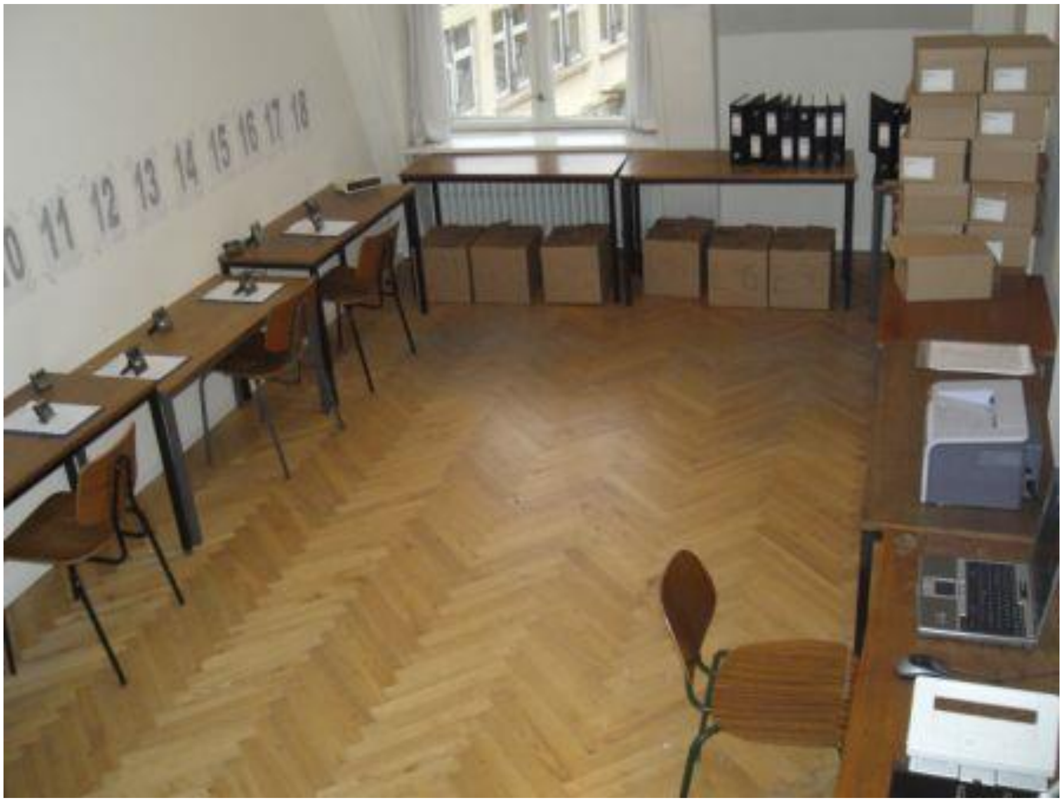


Figure B2: Control Room

The University of Copenhagen generously provided us with an entire floor (app. 320 m^2) of 11 offices which were furnished with tables and chairs. Two offices were used for storage of materials, one office was used as control room (see figure B2) and work was carried out in the remaining eight offices.

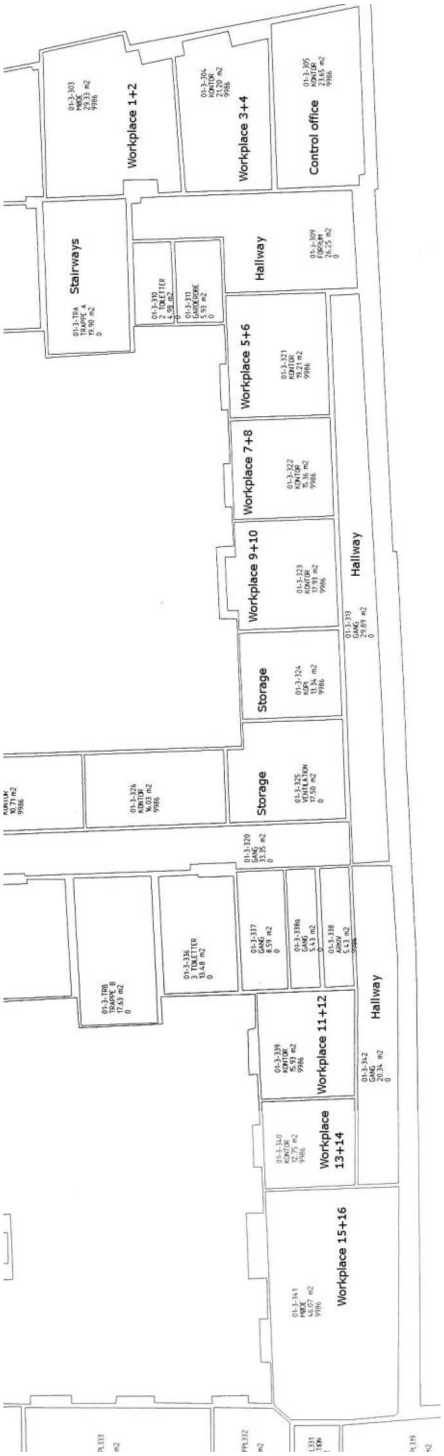


FIGURE B3: FLOOR PLAN

Appendix C. Description of the Work Task

The participants were seated in a two-person office at a workstation facing the wall. Figure C1 shows a photograph of the workstation.

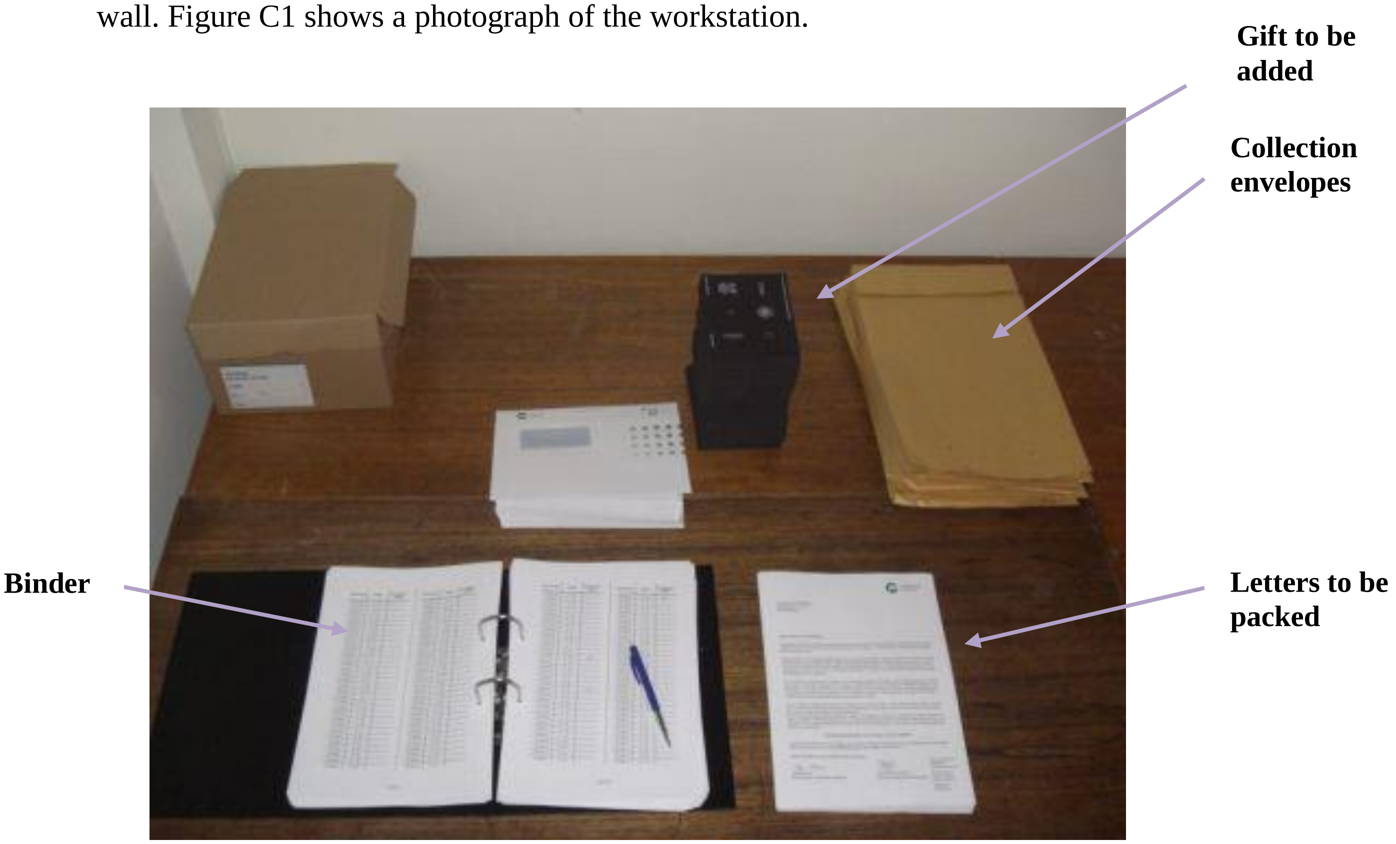


Figure C1: Work Station

Each letter had an ID number (ranging from 12,000 to 51,999). The order of the letters was randomized such that each participant was given letters from the entire interval.

The 40,000 letters had to be sorted into 5 main categories (A to E). These were then split further into a total of 96 subcategories (A-1 to E-96). The sub-categories were assigned randomly and were not printed on the letters. Each participant would get letters belonging to six subcategories and would have to sort the letters accordingly.

For each letter, the task was to: Look up the letter’s ID number in a binder with 600 pages and see which category (A-1 to E-96) the letter belongs to; Look up the category type (A to E) in a separate list and see whether the letter should include a gift (letters in categories B and D should include a small foam puzzle); Fold the letter and stuff it into an envelope. If category B or D, then also include a gift; Close the envelope; Sort the envelope into the collection envelope marked with the corresponding subcategory label.

The participants received both oral and written instructions on how to do the task. These instructions were given individually and we demonstrated how to prepare an envelope. The participant then stuffed an envelope under supervision to verify understanding of the procedure. If successful, the participants worked alone for 90 minutes. An alarm clock was set in the control room to enforce to time limit. After the 90 minutes, we stopped the participants and counted the number of envelopes stuffed. In total, the participants spent less than two hours at the University in each round.

Appendix D. Validation of Classification of First Names

As in correspondence tests, we use names as a marker of ethnicity. However, we do not use fictitious and highly stereotypical names but the actual names of workers. We categorize these names into ethnic types using our judgment complemented by lists of “typical” Danish and Muslim names we found on the web (such as www.muslimbabynames.net).

To test if actual names are effective markers of ethnicity, we run a complementary study with n = 144 juveniles in a secondary school on the outskirts of Copenhagen where we do not recruit for the experiment. The questionnaire (available from the authors on request) presents respondents with 4 randomly drawn pairs of candidates (i.e. using the actual names and actual pairs decision makers faced) and asks them classify the names as either Danish or Muslim. More specifically, respondents have the option to classify either, both or none of the two names as ‘Danish’ or ‘Arab/Muslim’. We randomize the order of names for a given choice in any given pair. This task is presented to respondents as part of a “classification study” which also contains 9 other, unrelated, tasks (e.g. classify cities as German or French). Participants are paid a flat fee of DKK 100 (€13.3) for completing the survey.

Table D1 shows that concordance rates are very high and confound is rare. In particular, the last column shows that 83 percent of the names we classify as Danish-sounding and 92 percent of those we classify as Muslim-sounding are categorized by respondents in concordance with our classification. Importantly, it very rarely happens (1 percent of the cases) that names we classify as belonging to one ethnic type are classified as belonging to the other category by respondents. Concordance and confound rates are similar for respondents with Danish-sounding and Muslim-sounding names.

TABLE D1: EFFECTIVENESS OF FIRST NAMES AS MARKER OF ETHNIC TYPE

	Boys		Girls		
Concordance	Danish-sounding	Muslim-sounding	Danish-sounding	Muslim-sounding	*Overall*
Danish names	80%	87%	84%	94%	*83%*
Muslim names	97%	94%	86%	92%	*92%*
Overall	*89%*	*90%*	*85%*	*93%*	*88%*
Confound					
Danish names	2%	3%	3%	5%	*2%*
Muslim names	0%	0%	1%	0%	*1%*
Overall	*1%*	*2%*	*2%*	*3%*	*1%*

Notes: The table shows the percentage (over of all names and respondents) of classifications in the survey study that are in line (“concordance”) or conflict (“confound”) with the classification into ethnic types in the experiment. Concordance occurs, for example, if a name we classify as Danish-sounding in the experiment is classified by respondents as Danish-sounding. Confound occurs, for example, if a name we classify as Danish-sounding is classified by respondents as Muslim-sounding. The number of respondents is *n* = 144.

Appendix E. Using Productivity Differences as Proxy for the Price of Discrimination

This appendix shows that our main result in Info (that an increase in price causally reduces taste-based discrimination) is robust to using a different type of team production function to estimate prices.

In section 4.1, we estimate the price from the marginal productivity of labor obtained from a particular type team production function (model A in table 2). We then use these (randomly assigned) prices to estimate the demand for discrimination (and the willingness to pay). By doing so, we assume that the price, and implicitly also the team production function, is known to decision makers. To demonstrate robustness, we use "raw" round 1 output differences as a proxy for the price in the estimation of the demand for discrimination and therefore tie the price of prejudice directly to observables. We find very similar results either way.

Table E1 replicates the analysis in table 3 using (half of) the difference of round 1 output between the candidates as a proxy for the price of discrimination. The significant ($p < 0.05$) coefficient of $\Delta Prod_{jk}$ in model (8) shows that if price goes up by €1, decision makers are about 3 percent less likely to discriminate. This estimate is similar to our result for *Price* in table 3 (3.0 vs. 3.6 percent). Also note that models (9) to (11) yield very similar results as models (2) to (4) in table 3.

Table E1: The Demand for Discrimination Using Output Differences as a Proxy for *Price*

Dependent variable: Discr	(8)	(9)	(10)	(11)
$\Delta Prod$	-0.030	-0.029	-0.028	-0.029
	(0.013)	(0.014)	(0.014)	(0.016)
Danish-sounding		0.014		0.088
		(0.160)		(0.273)
Male		-0.063		-0.138
		(0.152)		(0.266)
Danish-sounding * $\Delta Prod$			-0.001	-0.010
			(0.020)	(0.035)
Male * $\Delta Prod$			-0.005	0.010
			(0.017)	(0.029)
N	37	37	37	37
Adj. R^2	0.073	0.076	0.074	0.079

Notes: The table shows average marginal effects for probit regressions. Numbers in parentheses are robust standard errors. The dependent variable *Discr* = 1 for a discriminator and 0 otherwise. The variable $\Delta Prod_{jk}$ is the difference in output in round 1 by *other* minus output by *same*. To make the numbers comparable, we multiply the difference by 0.5 as the joint output was split among the two team members and express values in Euro, i.e. multiply with €0.5 per envelope stuffed. *Danish-sounding* and *Male* are dummy variables characterizing decision maker *i*.

Appendix F. Robustness of Price Effect with Respect to the Decision Maker's Productivity

Our discussion of the response of discrimination to the price of prejudice in Info in section 4.2 is entirely cast in terms of earnings foregone by choosing one candidate over the other, i.e. is based on opportunity cost. Below, we address issues relating to the absolute and relative productivity of the decision maker.

Table F1 investigates if decision makers with high productivity in round 1 tend to be less likely to discriminate. Such an effect is plausible if those with a strong preference for money work hard and also tend to choose a co-worker primarily on the basis of monetary concerns. But we find that the effect is weak is best ($Prod_1$ is insignificant in models 5 and 6). The table also serves to investigate whether the decision maker's productivity in round 1 relative to the productivities of the two candidates biases our estimates of the demand for discrimination. Our conclusion from the discussion below is that it does not.

TABLE F1: DISCRIMINATION AND THE DECISION MAKER'S PRODUCTIVITY

Dependent variable: Discr	(5)	(6)	(7)
Price	-0.030	-0.030	-0.017
	(0.016)	(0.015)	(0.018)
Prod_1	-0.046	-0.044	-0.043
	(0.026)	(0.033)	(0.032)
Prod_1^2	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)
Abs. distance to *same*		0.001	-0.003
		(0.008)	(0.007)
Same candidate below			-0.153
			(0.199)
Both candidates below			0.101
			(0.268)
N	37	37	37
R^2	0.147	0.147	0.177

Notes: The table shows average marginal effects estimated from Probit regressions. Numbers in parentheses are robust standard errors. The dependent variable *Discr* = 1 for a discriminator and 0 otherwise. The variable *Price* is expressed in Euro. $Prod_1$ and $Prod_1^2$ are decision maker *i*'s productivity and its square in round 1. *Abs. distance to "same"* is the absolute difference in round 1 productivity between decision maker *i* and the candidate of the same ethnic type as *i*. *"Same" candidate below* is a dummy variable taking the value 1 if the productivity of the decision maker in round 1 is between the two candidates. *Both candidates below* is a dummy variable taking the value 1 if the productivity of the decision maker in round 1 is higher than that of both candidates.

A potential concern with using an opportunity cost concept is that it does not take relative standing into account. Due to random matching of decision makers into triples, decision makers have a choice between candidates who can be more or less similar to the decision maker in terms of productivity in round 1. A particular concern is that choosing *same* may not reflect a preference for an ethnic type, but a preference for a co-worker with similar productivity. For example, a decision maker may choose *same* to avoid peer pressure and feeling uncomfortable when working with a much more productive co-worker. Model (6) in table F1 includes a variable *Abs. distance to "same"* which measures the absolute productivity difference between the decision maker and the candidate of the same ethnic type. The insignificant coefficient suggests that this concern does not affect the choice of co-worker.[2]

Model (7) in table F1 investigates a potential confound of loss aversion and taste-based discrimination. Due to the randomness of our matching procedure, decision makers have a choice between a) two candidates which are both less productive, b) both more productive, or c) a more and a less productive candidate. Compared to the case of being in a team with a co-worker with the same productivity, discrimination in case a) means incurring an additional loss, in b) foregoing an additional gain, and in c) incurring a loss rather than making a gain. Thus, loss aversion predicts that choosing *same* is less likely in case c) than in a) or b), and less likely in a) than b) for a given price of discrimination. To test, we add *"Same" candidate below* (equal to 1 in case c) and *Both candidates below* (a dummy equal to 1 in case a). The insignificant estimates suggest that loss aversion does not seem to have affected the choice of a co-worker. However, this result should be taken with a grain of salt due to multi-collinearity and the large number of explanatory variables compared to the number of observations.

Appendix G. Testing for Random Assignment of Price (Simulation)

A precondition for identifying the causal effect of prices on discrimination choices in treatment Info is that the price of discrimination (i.e. the opportunity cost choosing *same* over *other*) is randomly assigned to decision makers. In particular, the distribution of animus and the distribution of the prices must be independent.

Our matching procedure (see section 2) is sequential and matches (randomly drawn) decision makers with candidates from a pool of suitable candidates. That is, once a decision maker is determined, the candidates are drawn from a constrained

[2] We also find that decision makers do not have a bias in favor of the candidate with more "similar" productivity in a simple non-parametric test. Out of 37 decision makers, 21 choose the "closer", 16 the "further" candidate. This split is not statistically different from a 50:50 split (p = .560, χ^2 test).

set (e.g. the candidates and the decision maker have to be available on the same days). A possible concern is that our matching procedure caused selection in the sense that characteristics of the decision maker constrain the set of set of suitable candidates in such a way that the resulting distribution of prices is not random and independent of decision makers' animus.

Below, we provide three tests for random assignment of prices to decision makers. The tests do not reject the hypothesis of random assignment.

First, we test if the distribution of prices observed in our experiment is normal. Unconstrained random drawing of pairs of candidates implies that the distribution of $Price_i$ follows (half a) normal distribution. Because $Price_i$ is positive by design in Info, we mirror the experimental distribution on 0, and test this distribution for normality using standard tests. We cannot reject the normality assumption (p = 0.818, Shapiro-Wilk; p = 0.721, Shapiro-Francia; p = 0.901, Skewness/Kurtosis test for normality).

Second, we test if the sequentiality of our matching procedure caused a bias in the distribution of productivity differences between candidates. We test for productivity differences because these are directly observable and are a good proxy to $Price_i$ (see appendix E for a discussion). In particular, we test if the observed distribution of productivity differences is different from a simulated distribution which is obtained from random draws without (unintended) constraints. The simulated productivity differences are obtained by sampling from all participants who complete round 1 (n = 162) with two constraints which are intended consequences of our design (rather than unintended consequences of sequential sampling). Our simulation imposes that a decision maker is always matched with candidates of the same gender (to avoid confound of gender and ethnicity) and that *same* is by design less productive than *other* (to make choices informative). We sample 1'000 productivity differences for each type of decision maker. From this pool, we randomly draw 37 productivity differences and test the resulting distribution against the experimentally observed distribution using Mann-Whitney (MW) and Kolmogorov-Smirnov (KS) tests. We repeat the draw and run the tests 1’000 times. At a level of significance α, we expect fewer than α percent of these tests to reject (i.e. to have a p-value < α) if the null is true. At $\alpha = 0.05$, we find that these tests reject in less than 1 percent of the cases (MW: 0.009, KS: 0.005). At $\alpha = 0.1$, we find that the tests reject in less than 3 percent of the cases (MW: 0.029, KS: 0.009). In summary, our sequential matching procedure yields productivity differences which are indistinguishable from purely random draws of candidates and the sequential matching we use does therefore not seem to bias prices.

Third, we test for the independence of the distribution of animus and the distribution of prices by means of a simulation. This is a joint test for independence and other assumptions which are simultaneously imposed in the simulation. In particular, the simulation imposes a normal distribution of prices, a normal distribution of animus (an assumption we make in using probit regressions), and independence of the two distributions. We also impose utility maximization in that the decision maker discriminates if and only if $a_i \geq b\ Price_i$, just as we do in our estimations (see section 4.2). We compare the simulated distributions to the observed distribution in the experiment using non-parametric tests. We find that our experimental observation is likely to come from a population in which the assumptions above, including independence, jointly hold.

We proceed as follows. We randomly sample $n = 37$ pairs of $Price_i$ and a_i. $Price_i$ is drawn from the best fit of a normal distribution to estimated prices and a_i is drawn from the estimated distribution as explained in section 3.2. If $a_i \geq b\ Price_i$, we assign a value of $Discr_i = 1$, and = 0 otherwise. We calculate the conditional distribution of price for discriminators ($Discr_i = 1$) and non-discriminators, and the share of discriminators. We test these 3 distributions against the respective distributions as observed in the experiment using non-parametric tests. We repeat 1000 times for each distribution and expect a share of less than α (the significance level) of these tests to have *p*-values < α if the null hypothesis is true.

For the conditional distribution of the price of discriminators we find no significant difference between simulated and observed data. At $\alpha = 0.05$, we find that non-parametric tests reject in less than 3 percent of the cases (Mann-Whitney (MW): 0.024, Kolmogorov-Smirnov (KS): 0.020). At $\alpha = 0.1$, we find that the tests reject in 5 percent or less of the cases (MW: 0.050, KS: 0.040).

For the conditional distribution of the price of non-discriminators we find no significant difference between simulated and observed data. At $\alpha = 0.05$, we find that non-parametric tests reject in less than 2 percent of the cases (MW: 0.011, KS: 0.010). At $\alpha = 0.1$, the tests reject in 3 percent of the cases (MW: 0.030, KS: 0.030).

We find a mean simulated discrimination rate of 38.4 percent (observed is 37.8 percent, $n = 37$). We run 1'000 χ^2 test to test for differences in the simulated and observed discrimination rate. At $\alpha = 0.05$, we find that the tests reject in less than 1 percent of the cases (χ^2: 0.007), at $\alpha = 0.1$, the tests reject in less than 2 percent of the cases (χ^2: 0.013).

In conclusion, the tests for the conditional prices of discriminators, of non-discriminators and the discrimination rates reveal that the observed data in our experiment does not look different from simulated data imposing random allocation of prices to decision makers.

Appendix H. Eliciting Productivity Beliefs

We recruit $n = 353$ juveniles to elicit beliefs about individual and team output across ethnic types in the envelope stuffing task from two secondary schools where we do not recruit for the experiment. We carefully explain the work task to these participants and ask them to guess the productivity of actual workers in our experiment. We provide incentives for guessing correctly (the full questionnaire is available from the authors on request).

In particular, we present participants with a table of 7 randomly selected workers of the same gender and ask them to guess how many envelopes each worker stuffed when working in isolation in round 1. We also ask them to guess round 2 output for 6 randomly selected teams (2 homogeneous Danish-sounding, 2 homogeneous Muslim-sounding and 2 heterogeneous teams). As a point of reference, we provide participants with the observed median production in rounds 1 and 2. In total, 204 juveniles with Danish-sounding and 149 with Muslim-sounding names participate (42 have names that are classified as "other" and are omitted from the study). Beliefs are incentivized using a quadratic scoring rule. Participants receive max(0; 50 - 0.03d^2) where d is the difference between the true productivity and the guess. Average earnings are €13.6.

Table H1 shows that both types of participants tend to believe that workers with Danish-sounding names are more productive than workers with Muslim-sounding names when working alone (109 vs. 106 and 101 vs. 98, respectively). Remarkably, these beliefs about individual productivity differences across ethnic types are qualitatively in line with our results for round 1 production (116 vs. 100). However, both types of participants underestimate the true difference across ethnic types (3 vs. 16 letters).

Concerning team output, table H1 shows that both groups expect homogeneous Danish-sounding teams to be more productive than productive than heterogeneous teams which, in turn, are believed to be more productive homogeneous Muslim-sounding teams. The differences in beliefs about team production almost perfectly reflect the differences in beliefs about individual production. In particular, expected output increases by 3 letters by replacing a team worker with a Muslim-sounding name by one with a Danish-sounding name. Note that this almost perfect correspondence holds for participants of both ethnic types.

TABLE H1: AVERAGE OUTPUT GUESSES BY PARTICIPANTS IN COMPLEMENTARY STUDY

Participant	Individual workers		Teams		
	Danish-sounding	Muslim-sounding	Danish-sounding	Muslim-sounding	Hetero-geneous
Danish-sounding	109	106	225	220	223
Muslim-sounding	101	98	215	207	211

Notes: The table shows the average guesses for output of individuals and teams by participants in the belief elicitation study with Danish-sounding (n = 204) and Muslim-sounding (n = 149) names.

The analysis below shows that participants do not think that workers earn more in a homogeneous team than a heterogeneous team, for given round 1 output. Put differently, neither do the 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.

To test, we regress

$$(1) \qquad \Delta_i = \beta_0 + \beta_1 \delta_i + \beta_2 \mathit{Danish} + \varepsilon_i,$$

where Δ_i and δ_i capture the participants' beliefs about output of teams and individuals of different ethnic types. More specifically, Δ_i is participant *i*'s belief about output in a homogenous team of the same type as *i* minus *i*'s belief about output in a heterogeneous team. Thus, Δ_i captures how much participants with Danish-sounding names thought that all-Danish teams outperform heterogeneous teams, and vice versa for participants with Muslim-sounding names. The variable δ_i is *i*'s belief about output of individual workers of the same type as *i* minus *i*'s belief about output of workers of the other type. Thus, δ_i captures how much participants with Danish-sounding names thought that Danish workers outperform Muslim workers, and vice versa for participants with Muslim-sounding names. The dummy variable *Danish* equals 1 if the participant has a Danish-sounding name and is used to check whether the two groups differ in their beliefs about the production function.

The regression yields an insignificant coefficient β_0 which suggests that participants do not expect homogeneous and heterogeneous teams to be different, after controlling for beliefs about differences in individual productivity. We find $\beta_1 > 0$ which suggests that differences in beliefs about individual productivity translate into differences in beliefs about team productivity. The estimate for β_2 is not significant, indicating that the two groups do not have different beliefs about the type-specificity of the production function after controlling for beliefs about individual productivity differences. In summary, beliefs about individual productivity differences across types explain differences across homogenous and heterogeneous teams. In addition, homogenous teams are not generally believed to outperform heterogeneous teams, and these beliefs are not different across ethnic type of participant.

Appendix I. Decomposition of the Earnings Gap

The earnings gap discussed in section 5.3 is the difference in decision makers' total earnings between the benchmark case of statistical discrimination and observed earnings. A gap results if decision makers choose a worker of the on average less productive type. Such a choice can in principle result from holding a biased belief about the average price by type, from animus against a type of worker, or from other sources (unexplained part) In the analysis below, we abstract from biased beliefs and consider only rational expectations (see Hedegaard and Tyran 2014 for a discussion of biased beliefs). Statistical discrimination is profit-maximizing given available information and assumes that decision makers have rational beliefs on the price of discrimination and no animus.

Rational expectations ($Price_{\mathrm{i}}^{\mathrm{RE}}$) are determined for each i of the n = 37 decision makers as follows. We draw two co-workers (of the same gender as i) from the population of workers in our experiment (161 other workers, see table 1). We estimate team output with each drawn co-worker using i's production in round 1 and model A in table 2. The price of discrimination is then the difference in i's estimated earnings with either type. We repeat this procedure 1'000 times to obtain a distribution of $Price_{\mathrm{i}}^{\mathrm{RE}}$.

We use the mean of the distribution $\mu_{\mathrm{i}}^{\mathrm{RE}}$ to predict behavior for i in 2 scenarios which differ by whether we allow for animus (no vs. as estimated from treatment Info).

Absent any animus and assuming rational expectations, i chooses *same* if $\mu_{\mathrm{i}}^{\mathrm{RE}} < 0$ and *other* otherwise. In particular, we find that $\mu_{\mathrm{i}}^{\mathrm{RE}} < 0$ for all decision makers with Danish-sounding names, and $\mu_{\mathrm{i}}^{\mathrm{RE}} > 0$ for all decision makers with Muslim-sounding names.

To predict behavior in the case with animus, we feed $\mu_{\mathrm{i}}^{\mathrm{RE}}$ into model 1 from table 3 to calculate the probability that i chooses the co-worker of the same ethnic type ($Prob_i^{\mathrm{RE}}$).

Appendix J. Treatment NoName

This appendix describes a treatment designed to test for the effect a preference for working on a particular weekday on choices. In the main treatments (Info and NoInfo), the choice of candidates is framed in terms of workdays. The advantage of this frame is that it minimizes experimenter demand effects which might undermine our ability to measure taste-based discrimination. The limitation is that our observation that the candidate of the same ethnic type is chosen less often as the price (the productivity differential between the two candidates) of doing so goes up might not entirely be due to a taste for collaborating with a particular ethnic type but due to a taste for working on a particular day. In treatment Info, the decision maker knows the productivity and the names of the two candidates. In NoName, the decision maker only knows the productivity but not the names. Therefore, animus can play no role in NoName. The results of NoName show that day preferences do not significantly affect choices between candidates.

J1. General Description

The structure of NoName is the same as in the main treatments insofar as subjects are recruited to stuff envelopes for a large mailing and are paid at a piece rate. Workers are requested to show up for work twice in two consecutive weeks. In round 1, they all work by themselves and we measure their individual productivity on the job. At the end of round 1, we ask them to indicate time slots on which they are available for work in the coming week. We then call them on the phone and inform them that they will again do the same job but now have to work in teams of two, that they are paid the same piece rate but share the total revenue. We ask whether they are still available on two time slots on different weekdays they indicated at the end of round 1. If yes, they can choose whom to work with. So far, the control treatment is the same as the main experiment.

The control treatment differs in a number of ways from the main treatments. Importantly, in NoName, the decision maker only knows the productivity but not the first name of the candidates when making the choice. This contrasts with treatment Info where the decision maker knows the productivity and the first name of the two candidates (and NoInfo where he only knows the name but not the productivity).

The work task was shorter and more complex in NoName than in the other treatments 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. The main treatments were run with juveniles from secondary schools in Copenhagen and that sample was selected to consist of half of the subjects with Muslim-sounding, half with Danish-sounding names (in Vienna only 4 participants happened to have Muslim-sounding names). Those with Muslim-sounding names were less productive in both samples, but the difference was somewhat smaller in NoName (-10.4% vs. -13.8% in Info). However, both samples were gender-balanced. Procedures in phase 2 were also different insofar as round 2 was not actually run in NoName. Instead, subjects were called on the phone after all subjects had made their choice of partner, informing them that round 2 had to be cancelled. Observing actual output in round 2 is not necessary for the purpose of NoName and dropping phase 2 allowed us to have everyone choose between two candidates. A more detailed description is given in section J3 below.

J2. Results

We randomly allocate the candidate with the higher production value on the first (51.0%) and the second date (49.0% of the cases). We find that 51.0% of subjects (= 26/51) choose the first date on offer, 49.0% choose the second date which means that subjects clearly did not just choose the "first" date on offer (p = 1.000, χ^2 test). The vast majority (92.2 percent) of choices are for the day with high productivity candidate, only 4 out of 51 decision makers choose the day with a low-productivity candidate. This distribution clearly is different from equal distribution (p < 0.001, χ^2 test) indicating that choices in NoName were indeed driven by productivity differences. We find 2 choices for low-productivity candidates below the median "price" for going against material incentives (i.e. half the difference between the candidates' productivity), and 2 at or above the median. According to the hypothesis that people have pronounced preferences for which day to work on (assume these preferences are randomly distributed in the population), we should see a high incidence of choices of the low-productivity candidate at low “prices” (in fact, we should see close to 50% for prices close to zero) which then declines with prices. However, we find no evidence that the percentage is higher at lower cost (there is no observation in conflict with money maximization in the bottom decile, and only one in the bottom quintile). Section 4.3C in the main text provides regression analysis showing that choices do not significantly react to the price. We can therefore *safely reject the hypothesis that day preferences were driving choices* in the follow-up experiment.

J3. Detailed Description

The experiment has been conducted at the University of Vienna in two parts. Subjects were recruited to work one hour in the first in the week and one hour in the following week to prepare a mass mailing for the Faculty of Business, Economics and Statistics at its premises. Recruitment took place by sending our e-mail invitations to students of all faculties at U Vienna. We recruited a total of 66 subjects. We did a pretest with 5 subjects to estimate the amount of materials we need to provide. Of the remaining 61 subjects, 10 did not show up or did not respond to our call. A total of 51 subjects completed the entire experiment.

Two seminar rooms were available for letter packing for the duration of the experiment. One room had a total of three workstations as shown in figure J1. The other room served for storage and as a control room (see figure J2).

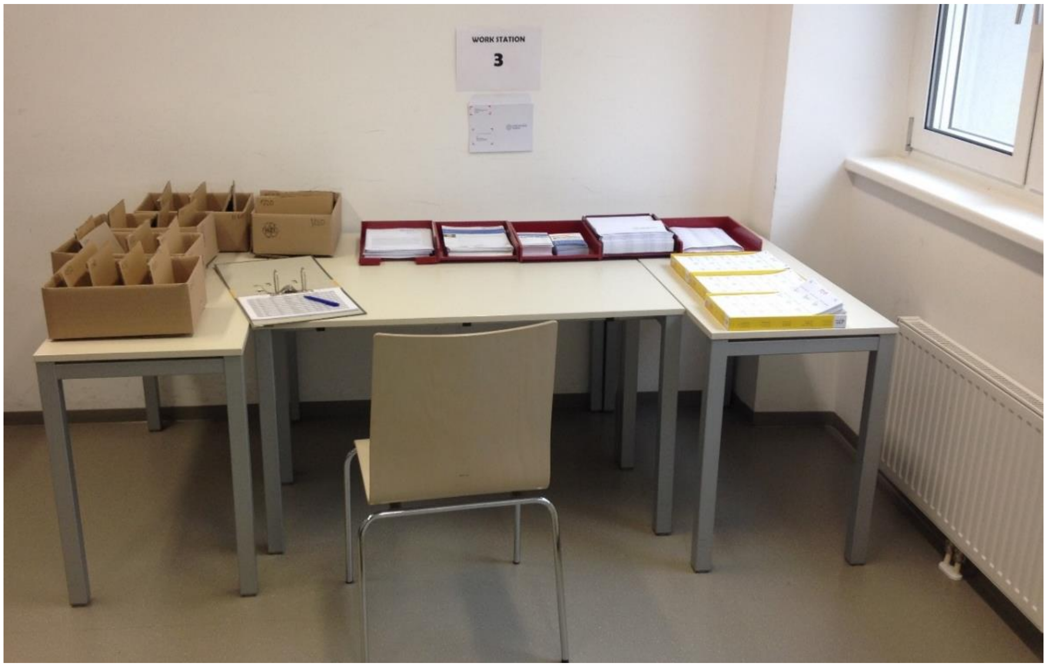


FIGURE J1*:* WORK STATION

The task was to prepare letters for a mass mailing. Names of recipients were listed in binder and depending on whether the recipient was listed as an alumni of U Vienna, the letter would contain announcements for an alumni-related event or only advertisements for other (public) events at the Faculty. The subject had to place the materials inside the envelope, post a sticker for the sender and one for the recipient addresses (to be placed as shown in a specimen fixed on the wall) and tick off the address in the binder. The letters were not to be sealed because we announced that we will check whether the letter was correctly prepared and we would not pay for incorrectly prepared letters. The letters then had to be sorted according to postal codes into cardboard boxes (left on figure J1). The task was carefully explained and demonstrated to the subject twice. An alarm clock was set to go off within 60 minutes in the control room.

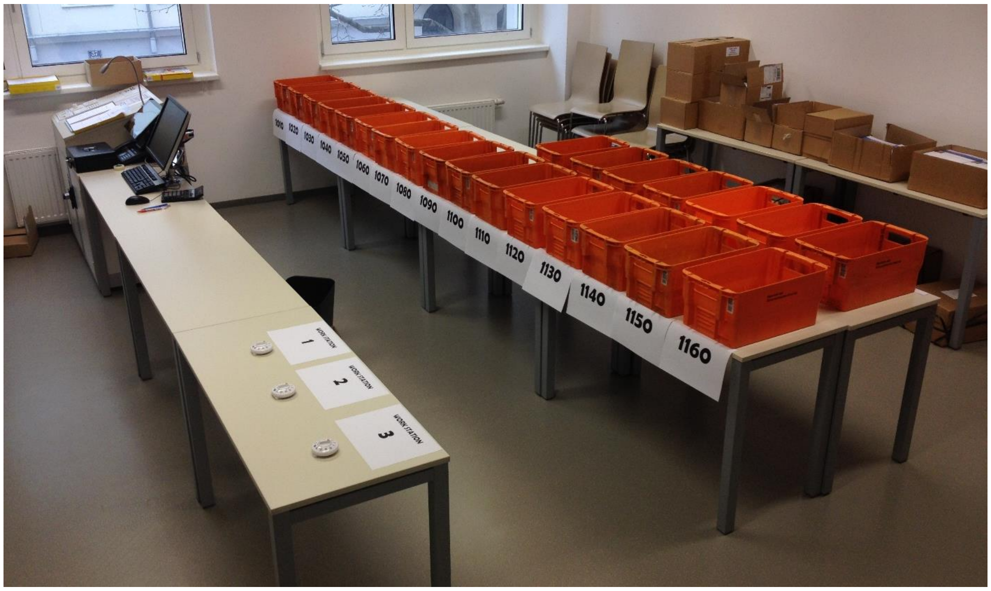


FIGURE J2: CONTROL ROOM

After 60 minutes, an assistant would ask the participant to stop working and to bring the stuffed envelopes to the control room (see figure J2). The assistant would count the number of correctly prepared envelopes and seal them. The letters were then sorted by postal code (see orange boxes in figure J2) to facilitate later postage and billing for the mailing. Subjects were paid on the spot and signed a receipt. They were paid €5 for being on time plus €0.50 per envelope, rounded upwards to the next €0.50. Average earnings were €27.7. Participants were asked to indicate two time slots in the coming week on which they were available to work on the same task.

Work time was between 9am and 8:30pm on three consecutive days. The arrival of subjects was staggered such that there was enough time to carefully instruct the new person arriving while handling the sorting, payment and data recording in the other room with 2 assistants present at any time. All work for round 1 was completed by Thursday evening, and subjects were called on the phone on Friday.

When calling, we say that we are still planning whether we need workers to show up for round 2, but currently it looks like it. Then we ask if they are still available on the days they indicated at the end of round 1. If no, we ask for two new dates and say that we need to reschedule and call them back later. If yes, we say “the task is the same as last time but we now want two people to work together. We have figured out that working in teams of 2 is more effective and the workers therefore also earn more on average. The two workers share the proceeds of work equally. We also pay the same piece rate as last time.” And (using Monday and Tuesday as examples): “I cannot see the names of the workers in my list, but I can see how many envelopes they stuffed. If you come next week on Monday to work, you will be with someone who has stuffed AA envelopes, if you come on Tuesday, you will be with someone who stuffed BB envelopes. When would you like to come, on Monday or Tuesday?”

To make prices in NoName comparable to those in the main treatments, we normalized the productivity differences decision makers faced in Info by the productivity distribution observed in NoName. The average output in Info and NoName was 106.8 and 45.6, respectively, which corresponds to a factor of 2.34. For example, decision-maker #140 in Info choose between candidates with a productivity of 126 and 134, respectively. This set of productivities was transformed to the set 54 (= 126 / 2.34) and 57 (= 134 / 2.34). We transformed all sets in Info and randomly assigned a set to each decision maker in NoName (sampling with replacement). We then randomly assigned productivities to the work days indicated and confirmed by the decision maker.

When the decision maker has made a choice, we record the decision in our sheet and confirm the choice to the subject by emphasizing that the arrangement is provisional: “We are still about to plan whether we need workers next week at all. We will send you a mail confirming that you need to show up on day yy. Please only show up if you got the confirmation mail.” When all decision makers have made their choices, we sent a mail telling them that they do not need to show up and the work for round 2 was cancelled.