I Z A Institute of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 13225

Collaboration, Alphabetical Order and Gender Discrimination – Evidence from the Lab

Vegard Sjurseike Wiborg
Kjell Arne Brekke
Karine Nyborg

MAY 2020

I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 13225

Collaboration, Alphabetical Order and Gender Discrimination – Evidence from the Lab

Vegard Sjurseike Wiborg
University of Oslo

Kjell Arne Brekke
University of Oslo

Karine Nyborg
University of Oslo and IZA

MAY 2020

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Collaboration, Alphabetical Order and Gender Discrimination – Evidence from the Lab

If individual abilities are imperfectly observable, statistical discrimination may affect hiring decisions. In our lab experiment, pairs of subjects solve simple mathematical problems. Subjects then hire others to perform similar tasks. Before choosing whom to hire, they receive information about the past scores of pairs, not of individuals. We vary the observability of individuals' abilities by ordering pair members either according to performance, or alphabetically by nickname. We find no evidence of gender discrimination in either treatment, however, possibly indicating that gender stereotypes are of limited importance in the context of our study.

JEL Classification: C91, J71, A13, D83

Keywords: discrimination, collaboration, alphabetic, gender

Corresponding author:
Karine Nyborg
Department of Economics
University of Oslo
P.O.Box 1095 Blindern
N-0317 Oslo
Norway
E-mail: karine.nyborg@econ.uio.no

1 Introduction

If measures of individual productivity are absent or imprecise, employers may hire applicants based on their beliefs about group level productivity, potentially leading to statistical discrimination (Arrow, 1998; Phelps, 1972). One challenge when evaluating individuals' productivity is that in collaborative work, individual contributions can be hard to disentangle. If employers hold different priors about the abilities of different groups, the result may be that, in terms of future employment possibilities, the returns to collaborative work is greater for some groups than others. In such cases, more precise information on individual contributions could mitigate discrimination. For example, Sarsons (2017) finds a gender difference in returns to coauthored papers in academic economics in terms of tenure probability. While men and women are equally awarded for solo-authored papers, women receive substantially less credit for coauthored papers (Sarsons, 2017). Sarsons does not find a similar coauthor penalty in data from sociology. One possible explanation for this is that the sociological convention of listing authors per contribution, the main contributor being listed first, provides more information about each author's intellectual contribution than the alphabetical ordering typically used in economics.

We explore, in a lab experiment, whether imprecise information about contributions to a pair task leads to gender discrimination in hiring decisions in the lab. Subjects performed mathematical tasks, subsequently hiring two individual subjects to do similar tasks for them, based on information about candidates' pair performance. To vary the degree of imprecision in ability information, we implemented a between-subject design where pair members were ordered according to their performance in one treatment, and alphabetically, according to nicknames, in another.

In the Alphabetical treatment, subjects were shown the nicknames of four candidates; the pairs each candidate had been a part of in previous mathematical quizzes; and the joint score of each pair. The names of a pair were ordered alphabetically. We call the other treatment First Author. The only difference in this treatment group was that names in pairs were ordered according to their contribution to the pair's joint score, listing first the best performing individual.

Multiple studies have shown that mathematics is often perceived as a male domain (e.g. Eccles et al., 1993; Nosek et al., 2009; Nosek et al., 2002), and that such self-perceptions can affect mathematical performance of male and female students (e.g. Spencer et al., 1999; Schmader, 2002).

Two papers are especially closely related to this one. In the experiment of Sarsons et al. (2019), men and women individually perform mathematical and grammatical tasks and are afterwards paired with an individual of the opposite gender. A group of other individuals (predictors) are then showed mix-gender pairs and joint scores - the combined number of correct answers of individuals in a pair - on set of mathematical tasks. They are then asked to predict future performance. Predictors systematically anticipate that men will attain higher scores on a subsequent mathematical task. The authors conclude that the differences in predictions based on joint scores, are caused by math being stereotypically male.

In Reuben et al. (2014) subjects perform mathematical exercises. Two subjects are randomly selected to act as employees, while the rest act as employers, observing the gender of candidates. Employers make hiring decisions either based on past performance of the employees, the employees' self-reported expectation about future performance, or on no other information than gender. Both men and women discriminate against the female employee under all information schemes.

Our design differs as all our subjects act as employees and employers. Furthermore, as opposed to the aforementioned papers, our main focus is on the distinction between two imprecise signals: alphabetical order or ordering per contribution. In contrast to Reuben et al. (2014), however, we find no statistically significant differences between male and female candidates in terms of the probability of being hired, within or between treatments, implying lack of gender stereotypes regarding mathematical stereotypes. This insignificance of gender persisted when we, later in the experiment, provided subjects with candidates' individual score, and even when information on performance was absent. Our findings indicate that the context of our experiment did not trigger gender discrimination, neither due to statistical discrimination nor for other reasons (see, e.g., Guryan and Charles, 2013; Bertrand and Duflo, 2017). This cannot, of course, be interpreted to mean that gender discrimination does not take place in other contexts, and that varying the precision of ability signals could well matter in such situations. Recent economic research has documented many examples of differential treatment of men and women, within and outside of the economics profession itself (see, e.g., Sarsons, 2017; Born et al., 2018; Hengel 2018; Wu, 2018; meta study by Lane, 2016). Solving simple mathematical problems may be less associated with gender stereotypes than, for example, becoming a tenured economics professor. Moreover, it could be that low stakes do not sufficiently incentivise discrimination (see List and Levitt, 2007), or that the gender egalitarian Norwegian culture (Kjeldstad, 2001) dampen such behaviour. If repeated in a context in which discrimination is in fact present, our experiment would shed more light on the impact of ordering conventions on discrimination.

2 Experimental Design

The present experiment consists of four parts[1]. In Part 1 the subjects only performed mathematical exercises, and in part 2 and 3 they chose partners among the other subjects based on various amounts of information about performance from Part 1. Part 4 was a ranking exercise, where subjects ranked three other subjects by guessing their performance in a quiz in Part 3, without having information on performance. All subjects acted as employers and employees. Thus, in order to be clear about their roles, we refer to subjects as candidates when they are picked as partners, or being ranked in Part 4. The instructions and the chronology of the experiment were as follows.

[1] See instructions in Supplementary Material, Section B

Prior to Part 1, we asked everyone to choose a nickname: ”During the experiment, you will be provided with anonymous information about other subjects' results. This will be easier if you have nicknames”. This took care of our concern about anonymity, but we also wanted the nickname to serve as an identifier of gender throughout the experiment. The following question was read aloud: ”Imagine that you were to have a different first name. What name would you prefer to have?”. We further requested them to choose a relatively common first name, hoping this would induce them to choose names corresponding to their gender. Self-reported gender was obtained in a post-experiment questionnaire.

Part 1: In Part 1 the subjects performed five math quizzes (quiz 1-5), each lasting for 60 seconds. The exercises were variations of adding two and three ciphered numbers and subtracting two ciphered numbers. In each quiz, each subject was randomly assigned a partner. Subjects did, however, not get to know their partner's nickname. Moreover, we neither informed the subjects about the number of correct answers provided by themselves, nor by their partner. The subjects got 1 NOK per question answered correctly by the pair[2]. Hence, if two paired subjects provided 8 and 10 correct answers, they got 18 NOK each in that round.

Part 2: The next part differed between the Alphabetical and First Author treatment. Subjects in both treatments were shown a table containing the nicknames of four candidates, the names of the candidates' partners in each quiz in Part 1 and each candidate's joint score with their partners[3]. The candidates were randomly selected for each subject.

In the First Author treatment, subjects observed a table where the members of each pair of which a candidate had been a part, were listed per contribution. That is, the one who obtained the most correct answers was listed first. Subjects in the Alphabetical treatment observed pairs listed alphabetically according to nicknames. Based on this information, subjects were asked to choose two of the four candidates as team members[4]. These would earn money for them in the following math quiz (quiz 6).

For each correct answer in the following quiz, a subjects would get 1 NOK. Each correct answer provided by their teammates would earn the subject 3 NOK. Thus, if a subject had 10 correct answers and her teammates had 5 and 20 correct answers, she would get 85 NOK. Subjects then had 60 seconds to solve as many exercises as possible in quiz 6.

Part 3: Part 3 was similar to Part 2 except that the subjects were asked to pick only one team member, and we presented each subject with information on four, randomly

[2] 1 EUR≈9.5 NOK.

[3] See Figure B.1 in the Supplementary Material.

[4] Prior to the quiz we asked the subjects questions regarding a toy table to see whether they understood how to retrieve information and the listing of the individuals within each pair.

selected candidates' individual scores from the quizzes in Part 1. After picking a teammate, the subjects performed a math quiz (quiz 7). As before, they got 60 seconds to solve as many exercises as possible. The payment scheme was equal to that in Part 2: 1 NOK per correct answer provided by themselves and 3 NOK per correct answer given by the teammate.

Part 4: In Part 4, the subjects were shown the nicknames of three others. We gave them no information about previous performance of the three candidates, only nicknames were displayed. Each subject was then instructed to guess which candidate had performed best and worst in the quiz in Part 3. They were rewarded 10 NOK per correct answer they provided.

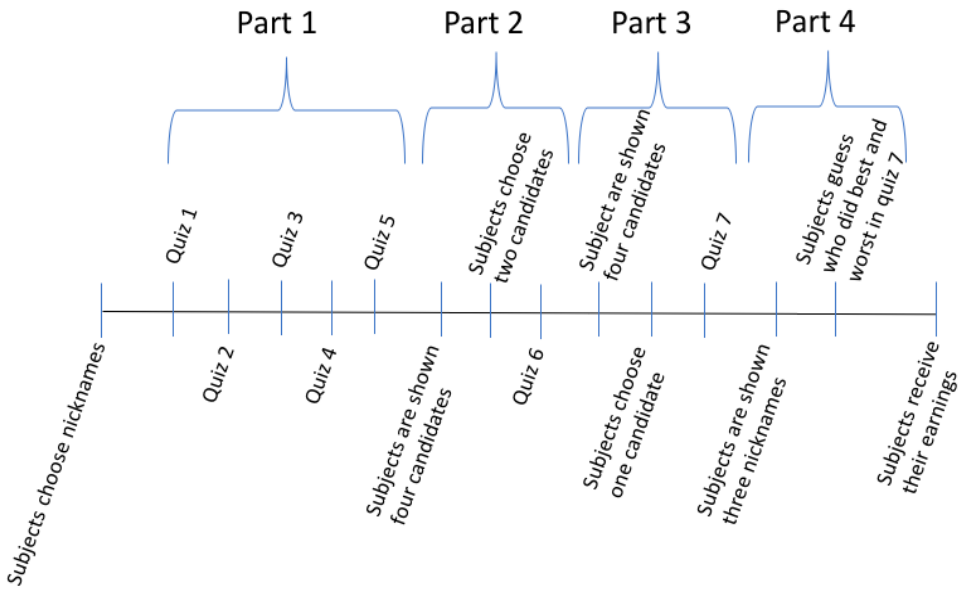

Part 1
Part 2
Part 3
Part 4
Subjects choose nicknames
Quiz 1
Quiz 2
Quiz 3
Quiz 4
Quiz 5
Subjects are shown four candidates
Subjects choose two candidates
Quiz 6
Subject are shown four candidates
Subjects choose one candidate
Quiz 7
Subjects are shown three nicknames
Subjects guess who did best and worst in quiz 7
Subjects receive their earnings

Figure 1: Overview of the experiment

The timing of the experiment is illustrated in Figure 1. We conducted the experiment in the spring and fall of 2017 at the University of Oslo. The subject pool consisted of 86 female and 62 male undergraduate students that were affiliated with different faculties and departments. Overall, we conducted eight sessions, each consisting of 16-22 subjects. In four of the sessions we implemented the Alphabetical treatment and in the other four we implemented the First Author treatment. The experiment was computerised and programmed in Z-tree (Fischbacher, 2007).

3 Results

In this section we analyse factors relevant for the choices of teammates and for evaluations in Part 4. The question of whether gender is a (statistically) significant determinant for the choice of partner, is particularly important. Since subjects did not observe self-reported sex, any discrimination against males or females should be related to nicknames.

In order to determine the gender corresponding to the nicknames of the subjects, we used the webpage Nordic Names (NN, 2017). NN returns the gender of names *used* in the nordic countries[5]. Four names did not yield results in NN. These cases were solved by having two research assistants indicating the gender. Both agreed on three of four subjects. The subject with a nickname of undetermined gender is treated as female in our analysis, as defined by one assistant and herself in the post-experiment questionnaire[6]. After determining the gender of the subjects based on their chosen nicknames, the pool of subjects consisted of 83 females and 65 males. Four of the 83 females, as defined by nickname, reported that they were males, while seven of the 65 males, defined by nickname, reported "female" as their sex[7]. From hereon "male" and "female" refers to the categorisation based on nicknames and not self-reported sex.

A first look at the arithmetic performance of male and female subjects reveals that the former, on average, performed slightly better in the quizzes in Part 1. Summing over all quizzes in Part 1 male subjects attained a mean of 41.3 correct answers while the mean score for females was 38.4. The difference is not statistically significant (p-value=0.20)[8]. With regard to choices, female candidates constitute a smaller portion of the chosen candidates relative to the share of females in the pool of eligible candidates in each part of the experiment. That is, the raw probability of a female candidate being picked is below 50%. These differences could, however, be caused by the differences in performance which we control for below.

3.1 Part 2

In Part 2 subjects were asked to pick two teammates out of four eligible candidates. We are primarily interested in three factors that potentially affected the subjects' decisions. 1) How the gender of a candidate directly affects the probability of being chosen, 2) whether such effects, if present, are different between treatments, and 3) how the gender of the candidates' partners affects the probability of being chosen.

In order to answer these three questions we employ a series of conditional logit models. These models exploit the variation between the four candidates that were presented to the subjects. Table 1 displays seven such regressions with odds ratio[9] (OR)

[5] Denmark, Faroe Islands, Finland, Greenland, Iceland, Norway and Sweden.

[6] Results are robust to treating this subject as male.

[7] Throwing them out of the analysis does not change any conclusions.

[8] Cumulative distributions and tests of statistical difference can be found in Supplementary Material, Section D.

[9] OR= $\frac{P(\text{female being picked})}{P(\text{female not being picked})} / \frac{P(\text{male being picked})}{P(\text{male not being picked})}$

as the measure of effect size. If an OR is equal to one, the probability of a candidate being chosen is independent of the corresponding variable. An OR above one indicates a positive relation between the variable and the outcome, and vice versa if OR$<$ 1.

Chosen is the dependent variable and indicates whether a candidate is picked to be a team member. *Female* is a dummy variable equal to one if the candidate is female and zero otherwise. *Cent. Female Partners* is a *centred* variable counting the number of times a candidate had a female partner in Part 1. We centre by subtracting the mean of female partners (in the whole sample of candidates) from each candiate's number of female partners[10]. *Individual Score* and *Pair-Score* are the summations of a candidate's individual and joint score, respectively, with his/her partners in Part 1. Lastly, *Listed First* is the number of times a candidate is listed first in a pair, and *Listed First in FA* is its interaction with the treatment variable. Model 1-3 and 4-6 concern subjects in the First Author (FA) and Alphabetical treatment (AL) respectively. In column 7 we run the regression on the whole sample.

Result 1: Gender is not a statistically significant determinant for the choice of two team members.

First off, there is no statistically significant difference in the fraction of males and females chosen within either of the two treatment groups when controls are not included (Table 1, Column 1 and 4). When we include controls there is still a slight tendency that females are less likely to be chosen, but it is not statistically significant. This holds regardless of whether we use *Individual Score* or *Pair-Score* and *Listed First* to control for individual merit. Specifically, ORs in column 3 and 6 would both reflect that the probability is 48% of a female candidate being chosen as teammate. When we pool both treatments we find that the OR of a female being chosen relative to a male candidate is 0.89 (column 7). This translates into a probability of around 49% of a female candidate being picked. Hence, the estimated ORs on gender comes very close to the benchmark of $OR = 1$ and the corresponding 50% probability of either gender being chosen[11].

Even though our point estimates suggest that males and females are treated equally, the estimated parameters are somewhat imprecise. For instance, the pooled regression (column 7) suggests small gender differences, but we cannot reject gender differences smaller than $OR = 0.59$ according to the 95% confidence interval (CI)[12]. This means, however, that we can reject gender differences of the magnitudes (point estimates) reported in Reuben et. al. (2014), including the one where employers observe the exact past performance of the candidates ($OR = 0.57$). In terms of power, we would reject the H_0 (no discrimination) with 99% probability if discrimination of women was as large as

[10] This procedure remedies problems with 1) huge correlation between *Female* and the interaction term *Female x Female Partners* leading to 2) variable inflation factors in the range 8-11 on both variables in the Alphabetical and First Author treatment and 3) unstable coefficient on *Female* (see Belsley, 1991).

[11] This is robust to using OLS regressions. See Table E.1 in Supplementary Material.

[12] A given significance level of 0.05 and a power of 80% imply an OR of 0.38, 0.39 and 0.52 underlying the models in columns 3, 6 and 7.

the point estimates in Reuben et. al.'s (2014) no-information and expected-performance conditions ($OR \approx 0.24$), both in the Alphabetical and First Author treatment.

Result 2: The impact of gender on decisions does not differ significantly between the Alphabetical and First Author treatment.

For completeness' sake, even though the gender of the candidates is not a significant predictor of choice within treatments, we formally test the difference between treatments: i.e. whether the coefficient on gender in equivalent models - model 1 and 4, 2 and 5, and 3 and 6 - are different[13]. None of the three two-sided t-tests suggest that we can reject the hypothesis that the coefficients are equal across treatments (p-value of 0.21, 0.61 and 0.90, respectively). Consequently, we find no evidence suggesting that females (males) are less likely to be chosen relative to males (females) in the Alphabetical compared to the First Author treatment.

Result 3: Both male and female candidates were relatively more likely to be chosen if they were matched with partners of the opposite gender.

Lastly, we examine the third question of whether the gender of the candidates' partners matters for the probability of being chosen. There is weak evidence that the number of female partners - or male partners - matters for the probability of being chosen. In column 7 we see that female candidates were more likely to be chosen if they were paired with male subjects (p-value=0.053). However, the insignificant (and positive) coefficient on *Cent. Female Partners* suggests a cross over interaction. That is, male candidates were less likely to be chosen if they were assigned a male partner. Hence, both male and female candidates received relatively less credit when paired with a subject of the same gender. This effect definitely has a gender dimension as the probability of being picked varies with the number of female (male) partners a candidate has had in Part 1. The effect does not, however, imply *consistent* evaluation in favour of males or females in mixed pairs. Also, this effect is only weakly significant and it is hard to see a good explanation for the effect. This might, of course, be a false positive.

[13] T-tests were executed after using the *suest*-command in Stata 15. The latter procedure combines estimations of different regressions and allows for comparisons of coefficients across models (Stata, 2019).

Table 1: Probability of being chosen

	FA	FA	FA	AL	AL	AL	Overall
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Dep Var: Chosen	b/se	b/se	b/se	b/se	b/se	b/se	b/se
Female	0.706	0.776	0.815	1.054	0.961	0.861	0.899
	(0.155)	(0.232)	(0.276)	(0.243)	(0.284)	(0.256)	(0.194)
Cent. Female Partners		0.970	1.062		1.173	1.138	1.151
		(0.161)	(0.180)		(0.233)	(0.212)	(0.149)
Female X Cent Fem Par		0.856	0.801		0.651	0.691	0.704^{*}
		(0.208)	(0.201)		(0.188)	(0.188)	(0.127)
Pair-Score			1.061^{***}			1.128^{***}	1.088^{***}
			(0.0108)			(0.0203)	(0.00993)
Listed First			1.685^{***}			0.907	0.895
			(0.273)			(0.0887)	(0.0802)
Listed First x FA							1.876^{***}
							(0.379)
Individual Score		1.089^{***}			1.111^{***}		
		(0.0149)			(0.0161)		
No. of candidates	288	288	288	304	304	304	592
No. of subjects	72	72	72	76	76	76	148

Notes: Clogit regressions with being chosen as the dependent variable. The table reports odds ratios and standard errors in parenthesis. Regressions in column 1-3 and 4-6 are run on observations in the First Author and Alphabetical treatment, respectively. *,** , and *** denote significance at the 10%, 5% and 1% level.

3.2 Part 3

Result 4: Gender is not a statistically significant determinant for the choice of team member when subjects observe candidates' individual score from Part 1.

In Part 3 of the experiment subjects chose one teammate and they were allowed to pick between four candidates. The only information about the candidates were their nicknames and individual scores in quiz 1-5, Part 1. Considering the subjects' choices in Part 2 we would expect that, with clearer signals of performance, individual score explains decisions at least to the same extent.

As expected, most subjects seem to be driven by the candidates' individual score. Approximately, 84 % of the subjects chose the individual with the highest rank as measured by a variable that ranks the four candidates according to total individual score in Part 1[14]. Only 10% of the subjects chose the second best candidate and 5% and 1% went with the third- and fourth-ranked candidate, respectively. Naturally, there are several potential reasons for why these subjects have not picked the top candidate, and one of them is gender. If female candidates were to be discriminated, subjects would have to pick lower ranked, male candidates instead of better performing female candidates. Hence, it would imply a negative correlation between the gender of the chosen and top ranked candidate. This correlation is close to zero ($\rho = -0.03$). While this subsample is too small for statistical inference purposes, the correlation coefficient suggests that gender is at least not a vital factor for the choice of a teammate.

Moving on to inference, Table 2 displays two conditional logit models that estimate the probability that a candidate is chosen as teammate by a subject. *Individual Score* is the total individual score of the candidates in Part 1, while *Female* indicates the gender of each candidate. The coefficient in column 1 corresponds to the result from Figure 1: statistically significantly fewer females were chosen as teammates. This reflects the fact that 20% of female candidates were chosen, relative to 30% of the male candidates. Still, this difference becomes smaller ($OR = 0.76$) and insignificant when controlling for performance[15]. Specifically, the OR corresponds to a conditional probability of approximately 45% of a female being picked[16]. While the confidence interval of the OR on gender in column 2 includes Reuben et. al.'s (2014) estimate of the probability of picking a female in their Past Performance treatment ($OR = 0.57$), power calculations suggest we would reject the H_0 (no discrimination) with a probability of 84% if 0.57 was the true OR.

[14] Candidates within the group of four that have equal scores are ranked equally for replication purposes. For example, if two candidates have the second best score they both get 2.5 as rank.

[15] Given our sample size, significance level of 0.05 and a power 80%, we would have detected an OR of 0.58.

[16] When we exclude the two outliers that chose the worst candidate in terms of rank, the estimate shifts to an OR of 0.93 which indicates close to no difference between male and female candidates. These two subjects were shown three female and one male candidate and both chose the lower performing male.

Table 2: Probability of being chosen

	(1)	(2)
Dep Var: Chosen	b/se	b/se
Female	0.597***	0.764
	(0.117)	(0.207)
Score		1.228***
		(0.0396)
No. of candidates	592	592
No. of subjects	148	148

Notes: Conditional logit regressions. The dependent variable is whether a candidate is chosen. The table reports odds ratios and robust standard errors in parenthesis. *,**, and *** denote significance at the 10%, 5% and 1% level.

3.3 Part 4

In Part 4 subjects were asked to guess which candidate performed worst (group W) and best (group B) in the quiz in Part 3. If male and female candidates were equally likely to be picked, the gender composition in each group (W and B) would reflect the one in the whole pool of candidates. Hence, we compare the fraction of females in group W and B to the no-discrimination benchmark of 0.57, which is the fraction of females among all candidates displayed to the subjects.

Result 5: The fractions of females chosen as ”Best” and ”Worst” are not statistically significantly different from the no-discrimination benchmark.

The share of females in group B is 0.56. Testing (two-sided) against the 0.57-benchmark yields a p-value of 0.87. In group W the share of females is slightly smaller, 0.55, but still fairly close to the benchmark. The binomial test returns a p-value of 0.74 when testing (two-sided) whether the proportion of women is equal to 0.57. Hence, statistically significant gender gaps remain absent[17]. Power calculations suggest that true fractions below 0.45 would be significant at a 5% in at least 80% of drawn samples.

4 Conclusion

In this paper we report the results of an experiment primarily designed to answer two questions: 1) Are females less likely than males to be hired to perform mathematical

[17] Note that subjects might have remembered candidates from previous encounters. Indeed, Table E.3 (Supplementary Material) shows that there is a positive relationship between candidates’ score in Part 1 and assignment to group B. Conversely, score in Part 1 is negatively related to group W assignment. The coefficient on gender is close to 1 and statistically insignificant.

exercises if subjects only observe the product of joint work? 2) How does alphabetical ordering of candidates affect the probability of females being hired, compared to ordering per contribution?

Overall, gender is not a statistically significant determinant of choice in any treatments in the experiment and the point estimates are close to no-discrimination benchmarks. While we cannot rule out small/medium sized effects in disfavour of females, we find less discrimination than suggested by the point estimates found by Reuben et. al. (2014) in the US.

In our experiment, different treatment of male and female candidates does not turn out to be statistically significant when subjects only observe the product of pair-work. When individual scores are available, performance remains the only significant variable. Lastly, choices without information about performance are very close to the no discrimination benchmark. All pieces of evidence seems to suggest that there is little or no gender discrimination in our experiment. Inasmuch as subjects behaved fairly gender neutrally when pairs were ordered alphabetically, there was little or no room for even more neutral behaviour in the First Author treatment. To answer the question of whether ordinal ordering of collaborators per contribution has mitigating effect on discrimination, our experiment would have to be replicated in a context where discriminatory behaviour actually does occur.

Acknowledgements: We thank participants at the 2018 World Conference of the Economics Science Association (Berlin), 13th Nordic Conference on Behavioural and Experimental Economics (Odense), Andreas Kotsadam for helpful comments. All errors are our own.

Data: The data for this project has been deposited in Harvard Dataverse,
doi: https://doi.org/10.7910/DVN/AZDWWT.

References

[1] Arrow, K. J. (1998). What Has Economics to Say About Racial Discrimination? *Journal of Economic Perspectives*, 12(2), 91-100.

[2] Belsley, D. A. (1991). *Conditioning diagnostics: Collinearity and weak data in regression*. United States: Wiley.

[3] Bertrand, M., & Duflo, E. (2017). Field Experiments on Discrimination. In E. Duflo & A. Banerjee (Eds.), *Handbook of Economic Field Experiments*, pp. 309-393. Amsterdam: North Holland.

[4] Born, A., Ranehill, E. & Sandberg, A. (July 3, 2018), A Man's World? The Impact of a Male Dominated Environment on Female Leadership. Available at SSRN: https://ssrn.com/abstract=3207198.

[5] Eccles, J., Wigfield, A., & Harold, R. D. (1993): Age and Gender Differences in Children's Self- and Task Perceptions During Elementary School. *Child Development*, 64(3), 830-847.

[6] Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics* 10(2), 171-178.

[7] Guryan, J., & Charles, K. K. (2013). Taste-Based or Statistical Discrimination: The Economics of Discrimination Returns to its Roots. *The Economic Journal*, 123(572), 417-432.

[8] Kjelstad, R. (2001). Gender policies and gender equality. In M. Kautto, J. Fritzell, B. Hvinden, & H. Uusitalo (Eds.), *Nordic Welfare States in the European Context*, (pp. 66-97). London: Routledge.

[9] Lane, T. (2016). Discrimination in the laboratory: A meta-analysis of economics experiments. *European Economic Review*, 90, 375-402.

[10] Levitt, S. D., & List, J. A. (2007). What do laboratory experiments measuring social preferences reveal about the real world? *The Journal of Economic Perspectives*, 21(2), 153-174.

[11] Nordic Names (2017): Accessed 12/12-2017

[12] Nosek, B. A., Smyth, F. L., Sriram, N., Lindner, N. M., Devos, T., Ayala, A., Bar-Anan, Y., Bergh, R., Cai, H., Gonsalkorale, K., Kesebir, S., Maliszewsky, N., Neto, F., Olli, E., Park, J., Schnabel, K., Shiomura, K., Tulbure, B. T., Wiers, R. W., Somogyi, M., Akrami, N., Ekehammar, B., Vianello, M., Banaji, M. R., & Greenwald, A. G. (2009). National differences in gender-science stereotypes predict national sex differences in science and math achievement. *Proceedings of the National Academy of Sciences*, 106(26), 10593-10597.

[13] Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Math=male, me=female, therefore math≠me. *Journal of personality and social psychology*, 83(1), 44-59.

[14] Phelps, E. S. (1972). The statistical theory of racism and sexism. *The American Economic Review*, 62(4), 659-661.

[15] Reuben, E., Sapienza, P., & Zingales, L. (2014). How stereotypes impair women's careers in science. *Proceedings of the National Academy of Sciences*, 111(12), 4403-4408.

[16] Sarsons H. (2017). Recognition for Group Work: Gender Differences in Academia. *American Economic Review: Papers and Proceedings*, 107(5), 141-145.

[17] Sarsons, Gërxhani, Reuben & Schram (2019). *Gender Differences in Recognition for Group Work*, (Forthcoming)

[18] Schmader, T. (2002). Gender identification moderates stereotype threat effects on women's math performance. *Journal of Experimental Social Psychology*, 38(2), 194-201.

[19] Spencer, S. J., Steele, C. M. & Quinn, D. M. (1999). Stereotype Threat and Women's Math Performance. *Journal of Experimental Social Psychology*, 35(1), 4-28.

[20] Stata. (2019). suest. Retrieved from https://www.stata.com/manuals13/rsuest.pdf. Accessed 31/1-2019

[21] Wiborg, Vegard; Kjell Arne Brekke; Karine Nyborg (2020). "Replication Data for: Collaboration, Alphabetical Order and Gender Discrimination". Harvard Dataverse, doi: https://doi.org/10.7910/DVN/AZDWWT.

[22] Wu, A. H. (2018). Gendered Language on the Economics Job Market Rumors Forum. *American Economic Association Papers and Proceedings*, 108, 175-179.

Supplementary Material

A Treatments and Sessions

Table A.1 gives an overview of sessions and the number of subjects in each session.

Table A.1: Overview

	N	Treatment
Session 1	22	First-Author
Session 2	18	Alphabetical
Session 3	16	First-Author
Session 4	20	Alphabetical
Session 5	16	First-Author
Session 6	16	Alphabetical
Session 7	18	First-Author
Session 8	22	Alphabetical

B Instructions

[Translated from Norwegian]

Welcome to this experiment. The results will be used in a research project. Hence, it is important that you follow certain rules. Do not communicate with other subjects during the experiment. Mobile phones must be turned off or switched to silent mode and be put away. You are not allowed to use any other software on the computer during the experiment.

The experiment is completely anonymous. None of the other subjects will know any of the decisions you have made. Nor will it be possible for any other individual to link decisions to any single subject. You will be told when the experiment starts, and when you can start typing your answers on the computer in front of you. If you have any questions during the experiment, raise your hand, and one of us will be assisting you.

As compensation for your participation you will receive money. How much money you make depends on the choices you and others make during the experiment.

Multiple times during the experiment you will see an "OK" button on the screen in front of you. It is important that you click this when you are ready to move on. If you do not do so, everyone else will be waiting for you.

Before we start
The experiment consists of four parts. What you do in one part will not affect how much money you can make in the next part.

During the experiment you will be provided with anonymous information about other subjects' results. This will be easier if everyone have nicknames.

You will soon see a question on the computer screen. Your answer determines your nickname in the experiment.

The question is as follows: "Imagine that you were to have a different first name. What name would you prefer to have?" Your answer will be your nickname throughout the experiment. We ask you to pick a relatively ordinary first name consisting of 3-8 letters. To avoid that multiple subjects choose the exact same name, we ask you to add any capital letter (for example: "Anne K"). Because of technical reasons, it is important that the first letter in your first name is CAPITAL. We ask that you do not share your chosen nickname with others after the experiment is completed.

Part 1

In Part 1 of the experiment you and another subject, whom we will call your partner, will constitute a pair. The software draws the pairs randomly.

When you are ready to start, click the "OK" button. You will be presented with a series of simple mathematical exercises. You have 55 second to solve as many as possible. The same accounts for your partner. We will let you know when these 55 seconds have passed. Then you have 5 seconds to push the "OK" button to save your answers. Note that you have to push "OK" before the time is out – if not, your answers are not registered. After you have pushed "OK" you will not be able to solve more exercises.

Both you and your partner get 1 NOK for each correct answer you provide. This applies independently of whom answers correctly, and independently of whether you answer the same exercise correctly or not.

For example, if you provide 10 correct answers and your partner provides 8 correct answers, you get 18 NOK each in that round.

You will not know your partners nickname, and you will not observe his/her answers.

This procedure will be repeated four more times. For each round, the software will draw a new partner at random. Thus, Part 1 has five rounds and you will have a new partner in each round.

Raise your hand if you have any questions. The experiment starts when everyone have pushed the button "OK, I am ready to start".

Part 2

This part is similar to Part 1, but there is just one round, and the payment scheme is a bit different.

As before, you will get a series of simple mathematical exercises on the screen. You shall solve as many as possible within 55 seconds. Thereafter, you have 5 seconds to push "OK" and thereby save your answers. For each correct answer, you will get 1 NOK.

In addition, you will choose a team consisting of two other subjects. These will work and earn money for you. For each correct answer they provide, you get 3 NOK. These earnings are added to the money you earn by answering correctly.

The first you will do in Part 2, is picking this team.

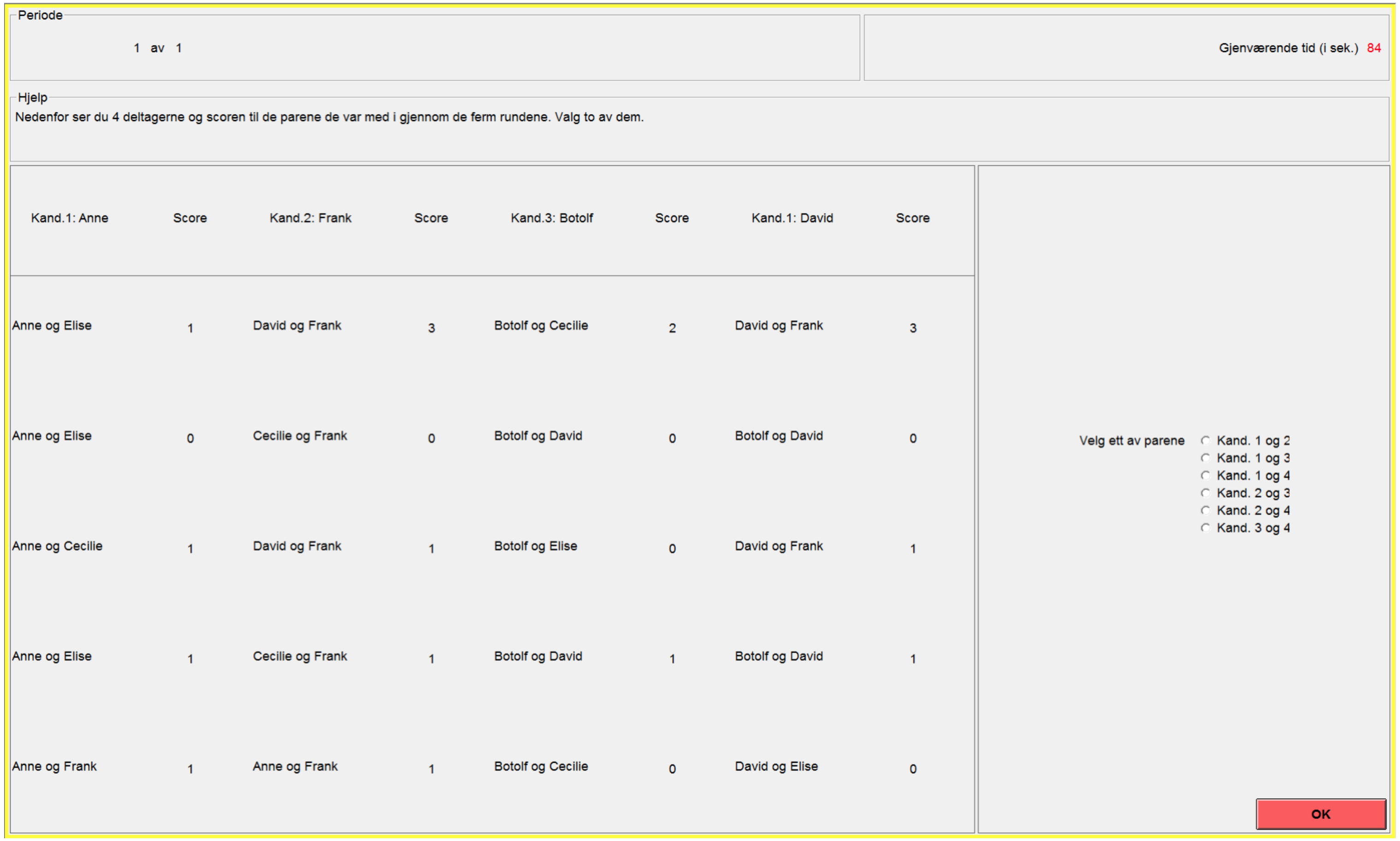
On the screen in front of you will see a table with four nicknames in the upper row. Each of them has a candidate number. You are going to choose two of these four candidates.

The candidates have been in five different pairs in Part 1, just like you. The table provides an overview of the candidates' partners in the five round in Part 1, and the candidates' joint score with their partners' in each round.

[Only for first-author treatment]: For each pair the names are ordered according to score so that the one with the highest score is listed first. (If both have equal scores, the computer randomly draws the order.)

[Only for alphabetical treatment]: For each pair the names are ordered alphabetically. The table will look something like this:

Figure B.1: Choose two candidates



Periode

1 av 1

Gjenværende tid (i sek.) 84

Hjelp

Nedenfor ser du 4 deltagerne og scoren til de parene de var med i gjennom de ferm rundene. Valg to av dem.

Kand.1: Anne	Score	Kand.2: Frank	Score	Kand.3: Botolf	Score	Kand.1: David	Score
Anne og Elise	1	David og Frank	3	Botolf og Cecilie	2	David og Frank	3
Anne og Elise	0	Cecilie og Frank	0	Botolf og David	0	Botolf og David	0
Anne og Cecilie	1	David og Frank	1	Botolf og Elise	0	David og Frank	1
Anne og Elise	1	Cecilie og Frank	1	Botolf og David	1	Botolf og David	1
Anne og Frank	1	Anne og Frank	1	Botolf og Cecilie	0	David og Elise	0

Velg ett av parene
○ Kand. 1 og 2
○ Kand. 1 og 3
○ Kand. 1 og 4
○ Kand. 2 og 3
○ Kand. 2 og 4
○ Kand. 3 og 4

Choose two candidates from the upper row in the table, by ticking off the pair with candidate numbers you wish to choose (at the right). This pair will be your chosen team.

When you have made your choice and you are ready to move on to the next part, click the “OK” button. You will then get a series of simple mathematical exercises and have 55 seconds to solve as many as possible. Afterwards, you have 5 seconds to push “OK” and thereby saving your answers.

As mentioned, you get 1 NOK for each correct answer provided by yourself, and 3 NOK for each correct answer provided by your team (the number of correct answers your chosen team members have combined).

For example, if you have 10 correct answers and the two candidates have 5 and 20 correct answers, respectively, you earn 85 NOK (10+3*(25)=85).

To be sure that we have explained this sufficiently well, we ask you to answer some questions that appear on the screen. Raise your hand if you have a question. Part 2 will start after everyone have answered the questions and pushed “OK”.

Part 3

Part 3 is similar to Part 2. As before, you will get a series of simple mathematical exercises on the screen, and you shall solve as many as possible within 60 seconds. For each correct answer, you get 1 NOK.

In addition you will now pick one other subject who will work and earn money for you. For each correct answer provided by this person, you get 3 NOK. This is in addition to the money you earn by answering correctly.

The first thing you will do in Part 3 is to pick this person. You will be shown a table with the nicknames of four candidates from which you can choose, and the number of correct answers they provided in each round in Part 1. You shall pick one of these candidates.

After having picked one candidate and you are ready to proceed, push the “OK” button. You will be presented with a series of simple mathematical exercises, and you have 55 seconds to solve as many as possible. We will let you know when these 55 seconds have passed. Thereafter, you have five seconds to push the “OK” button in order to save your answers.

As mentioned, you get 1 NOK for each correct answer provided by yourself and 3 NOK for each correct answer provided by your chosen candidate.

For example, if you have 10 correct answers and your chosen candidate has 20 correct answers, you earn 70 NOK (10+3*20).

Push “OK” when you are ready to start.

Part 4
In this part you are going to answer a few questions.

The nicknames of three other psubjects will appear on the screen. Your task is to guess which of these obtained the highest number of correct answers in Part 3, and who provided the fewest correct answers.

You get 10 NOK per each correct answer. (If some of the three subjects answered the same number of exercises correctly, the order in which you place them is not relevant for your payment)

Afterwards, you will get some additional questions about yourself.

C Categorization of Names

Table C.1 shows all of the names successfully connected to gender using Nordic Names. Table C.2 displays the names determined by the two research assistants (RA). Gender is equal to 1 if defined as female and 0 if defined as male.

Table C.1: Gender determined by Nordic Names

Name	Gender	Name	Gender	Name	Gender
Einar G	0	David O	0	Karen A	1
Siri M	1	Tuva	1	Per A	0
Nora S	1	HaNna	1	Lars K	0
Joe R	0	Winnie L	1	Kine K	1
Emilie E	1	Emma L	1	Martin H	0
Kim P	0	Markus K	0	Sofia L	1
Sofie S	1	Martin L	0	Emma W	1
Sean P	0	Sara J	1	Sofie S	1
Solveig S	1	Sofie S	1	Petry B	1
MARCUS	0	Sofie J	1	Oline H	1
Sondre E	0	Linnea H	1	Andre Q	0
Nora S	1	Emma H	1	Ida V	1
Mia S	1	Sofie L	1	Roald D	0
Sofia L	1	Ariana	1	Mikael A	0
Frida K	1	Tina B	1	Iselin L	1
Bjørn L	0	Eva R	1	Rikard S	0
Fredrik J	0	Karl I	0	Cecilie B	1
Balder K	0	Hans E	0	Andrea H	1
Natalie H	1	Bård U	0	Kaja K	1

Thomas K	0	Tarald O	0	Trond P	0
Alex D	0	Ada A	1	Alex K	0
Oddvar M	0	Leif B	0	Line F	1
Anna B	1	Ada B	1	Jan M	0
Silje S	1	Rick S	0	Lea K	1
HelgeG	0	Anja N	1	Arne J	0
Josef K	0	Tina H	1	Viktoria R	1
Jesper K	0	Henrik	0	Eline S	1
Ella X	1	Thea B	1	Emma S	1
Mike H	0	Pia R	1	Petter S	0
Vera S	1	Amalie K	1	Harald H	0
Anders T	0	Alice	1	Thea N	1
Josef K	0	Emma L	1	Oda H	1
Arne T	0	Jesper K	0	Maria H	1
Emma B	1	Lasse K	0	Martin K	0
Mette G	1	Sindre B	0	Lise T	1
Ole O	0	Erik S	0	Benjamin H	0
Sanna P	1	James B	0	KATJA K	1
Molli E	1	Esten G	0	Isak X	0
Linnea L	1	Elise M	1	Eirik B	0
Helene B	1	Erik B	0	Angelika L	1
Lilly S	1	Julie B	1	Knut F	0
Henrik B	0	Solveig M	1	Bengt K	0
Martha E	1	Thor A	0	Hanne S	1
Sofus B	0	Marie E	1	Klara N	1
Thea D	1	Irja K	1	Jonas P	0
Julia K	1	Sara M	1	Theo S	0
Thomas B	0	Henny S	1	Thomas Z	0
Sara N	1	Louise H	1	Ada Q	1

Notes: The table shows gender categorisation of names by using the database Nordic Names (2017).

Table C.2: Gender determined by research assistants

Name	Gender
Coffe K	0
FADIS F	1
Megara H	1 (0)
Hercules Z	0

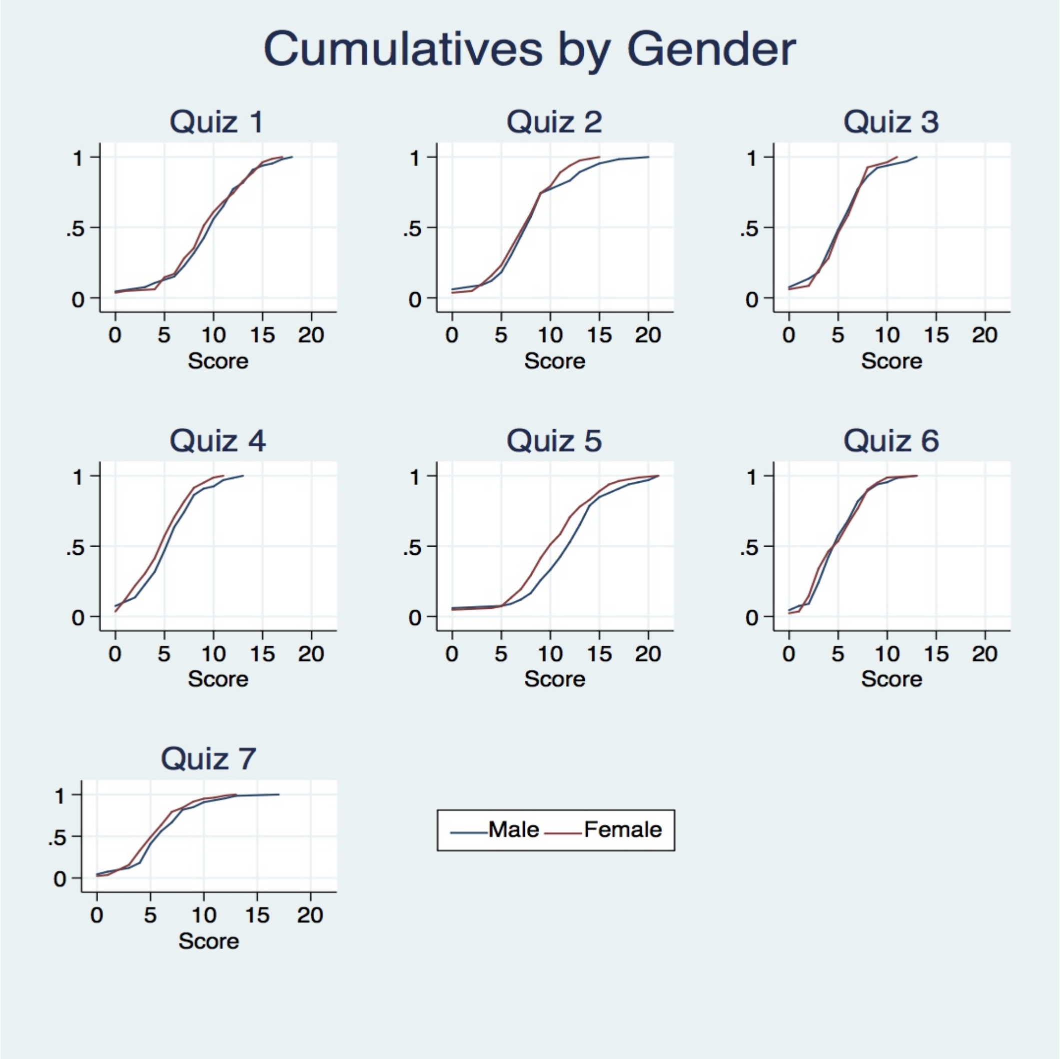
Notes: This table shows how two research assistants (RA) categorised the names that were not included in the database Nordic Names (2017). Both RAs had to agree. (0) indicates that one RA defined the name as male as opposed to the other RA and the candidate's self reported gender. Results are robust to treating Megara as male.

D Descriptive Statistics

D.1 Differences in Mathematical Performance

Figure D.1 shows the cumulative distribution of score in each quiz in the experiment, by gender (as defined by nicknames), while figure D.2 and D.3 illustrate the cumulative distributions of individual and pair-score, respectively, when summing subjects' score in round 1-5 in Part 1.

Figure D.1



Notes: The figure depicts cumulative distributions of the number of correct answers in each quiz, by gender. Gender is defined by the Nordic Names (2017).

Figure D.2

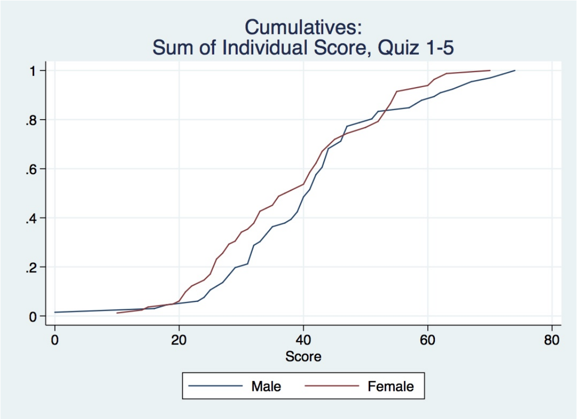


Figure D.3

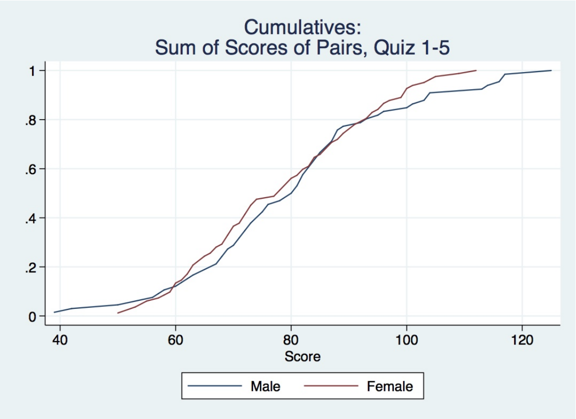


Figure D.2 shows cumulative distributions of the sum of correct answers in quiz 1 - 5, by gender. Figure D.3 shows cumulative distributions of the sum of correct answers each subject had with their partners in quiz 1 - 5, by gender. Gender is defined by the Nordic Names (2017) database.

Table D.1 and D.2 report tests on differences in distributions by gender (as defined by nicknames) and sex (as reported by themselves), respectively. Column 1 and 2 contain p-values of Komolgorov- Smirnov-test (left column) and Wilcoxon ranksum-test (right column).

Table D.1: Test of differences in distributions of score by gender

	P-Value	
	KS-test	WR-test
Quiz 1	0.84	0.50
Quiz 2	0.78	0.41
Quiz 3	0.99	0.69
Quiz 4	0.93	0.17
Quiz 5	0.15	0.03
Quiz 6	0.92	0.70
Quiz 7	0.47	0.16
Quiz 1-5	0.50	0.20
Quiz 1-5 (of the pair)	0.74	0.44

Notes: The table shows two tests of the differences in the distributions of the number of correct answers in each quiz, by gender, as defined by Nordic Names (2017). The null hypothesis is that the distributions are equal. The second and third column display p-values of the Komoglorov-Smirnov test and Wilcoxon Rank-Sum test, respectively.

Table D.2: Test of differences in distributions of score by sex

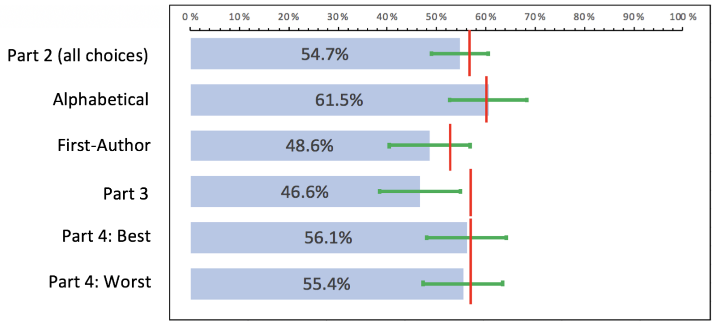
	P-Value	
	KS-test	WR-test
Quiz 1	0.41	0.25
Quiz 2	0.69	0.18
Quiz 3	0.99	0.51
Quiz 4	0.79	0.16
Quiz 5	0.15	0.05
Quiz 6	0.59	0.20
Quiz 7	0.67	0.26
Quiz 1-5	0.15	0.11
Quiz 1-5 (of the pair)	0.54	0.34

Notes: The table shows two tests of the differences in the distributions of the number of correct answers in each quiz, by sex, as reported by themselves. The null hypothesis is that the distributions are equal. The second and third column display p-values of the Komoglorov-Smirnov test and Wilcoxon Rank-Sum test, respectively.

D.2 Fractions of Females Among Chosen Candidates

There is a tendency that female candidates constitute a smaller portion of the chosen candidates relative to the share of females in the pool of eligible candidates in each part. Figure D.4 displays this pattern. The discrepancy is particularly big in Part 3 where the fraction of female candidates, vertical segment, is not covered by the 95% confidence interval of the share of females among the chosen candidates, horizontal segment. In Part 2 and 4, the proportions match relatively well, with the largest difference being found in the First Author treatment.

Figure D.4: Fraction of females among chosen candidates



Notes: Bars show the percentage share of females in the pool of chosen candidates in each part of the experiment. The horizontal segments indicate 95% confidence intervals. Standard errors are clustered on subject level for decisions made in Part 2. The vertical segments correspond to the portion of females among all eligible candidates in the respective parts and treatments.

E Alternative Models of Choices in Part 2, 3 and 4

E.1 Choices in Part 2 - Linear Probability Model

Table E.1: Probability of being chosen in Part 2

	First-Author	Alphabetical	Overall
	(1)	(2)	(3)
Dep Var: Chosen	b/ci95	b/ci95	b/ci95
Female	-0.0313	-0.00699	0.00310
	[-0.141,0.0788]	[-0.113,0.0988]	[-0.105,0.112]
Cent. Female Partners	0.0133	0.0274	0.0359
	[-0.0524,0.0790]	[-0.0470,0.102]	[-0.0182,0.0901]
Female X Cent Fem Par	-0.0544	-0.0416	-0.0621*
	[-0.172,0.0634]	[-0.125,0.0422]	[-0.132,0.00768]
Pair-Score	0.0101***	0.0178***	0.0150***
	[0.00700,0.0133]	[0.0153,0.0204]	[0.0130,0.0171]
Listed First	0.0795***	-0.0291*	0.00581
	[0.0403,0.119]	[-0.0585,0.000209]	[-0.0194,0.0311]
Female x First Author			-0.0257
			[-0.192,0.141]
First Author			0.0649
			[-0.0551,0.185]
Candidates	288	304	592
Subjects	72	76	148

Notes: OLS regressions with being chosen as the dependent variable. Regressions in column 1 and 2 are run on observations in the First-Author and Alphabetical treatment, respectively. Column 3 includes all observations. 95% confidence intervals are in parenthesis. Standard errors are robust and clustered on candidate ID to control for the fact that subjects are presented as candidates multiple times. *,**, and *** denote significance at the 10%, 5% and 1% level.

E.2 Choices in Part 3 - Linear Probability Model

Table E.2: Probability of being chosen in Part 3

	(1)	(2)
Dep Var: Fem Chos	b/ci95	b/ci95
Female	-0.105*	-0.0723*
	[-0.212,0.00195]	[-0.147,0.00223]
Score		0.0169***
		[0.0141,0.0197]
Constant	0.310***	-0.369***
	[0.227,0.392]	[-0.497,-0.241]
Observations	592	592
Subjects	148	148

Notes: OLS regressions with being chosen as the dependent variable. Since the candidates' individual scores from Part 1 was displayed to the subjects, we include the sum of the score on quiz 1-5 in column 2, to control for merit. 95% confidence intervals are in parenthesis. Standard errors are robust and clustered on candidate ID to control for the fact that subjects are presented as candidates multiple times. *,**, and *** denote significance at the 10%, 5% and 1% level.

E.3 Choice in Part 4

Table E.3 contains two conditional logit models on choices made in Part 4. The dependent variable in model 1 and 2 is whether a candidate is chosen as having performed worst or best, respectively, in the last quiz. *Female* indicates whether the candidate is female (1) or male (0) and *Score Part 3* reports the candidates' number of correct answers in the quiz in Part 3. *Pair-Score* is a candidate's sum of joint score with his/her partners in Part 1. These performance variables do predict choices. However, gender does not.

Table E.3: Probability of being categorised as "Best" or "Worst" in Part 4

	(1)	(2)
Dep Var: Chosen	b/se	b/se
Female	0.981	0.948
	(0.208)	(0.214)
Score Part 3	0.996	1.089**
	(0.0382)	(0.0450)
Pair Score	1.019**	0.977***
	(0.00757)	(0.00709)
Candidates	444	444
Subjects	148	148

Notes: Clogit regressions. The dependent variable is *chosen*. Model 1 is category Best and model 2 is category Worst. Effect sizes are in odds ratios and standard errors are reported in parenthesis. *, **, and *** denote significance at the 10%, 5% and 1% level.

F Ex-post Optimality of Choices in Part 2

When we compare the ex-post optimal choices in Part 2, in terms of earnings, to the actual choices, subjects in the Alphabetical treatment are on average closer to the optimal choice. Specifically, the average earnings of subjects in the Alphabetical treatment were about 4.14 NOK lower than what the ex-post optimal choice of partners would have generated. Comparably, subjects in the First Author treatment earned about 6.25 NOK less than this benchmark. This discrepancy in the ex-post optimality of the choices seems to be a result of the bad predictability of ordering (in the First Author treatment) on subsequent performance. Testing (t-test) whether the distance to the ex-post optimality measure differ significantly between treatments returns a p-value of 0.06. Hence, ordering the pairs according to individual performance did, at least, not bring subjects in the First Author treatment closer to the unobserved individual score of the candidates. Nor did it result in higher earnings.