Are Women Penalized for Coauthoring in Academic Economics?

- Evidence from the Lab

Vegard Sjurseike Wiborg

Master of Philosophy in Economics

Department of Economics

University of Oslo

May 2017

Are Women Penalized for Coauthoring in Academic Economics – Evidence from the Lab

Vegard Sjurseike Wiborg

http://www.duo.uio.no/

Trykk: Reprosentralen, Universitetet i Oslo

Preface

I have written this thesis with the (Advanced Method) scholarship support of the Department of Economics, University of Oslo, and it is part of a project initiated by my supervisor, Karine Nyborg, and Kjell Arne Brekke.

I wish to thank Karine Nyborg for introducing me to the topic of discrimination and experimental methods, and for excellent guidance throughout this process. Additionally, I want to thank Kjell Arne Brekke for good collaboration.

Thanks to Ragnhild Sjurseike for valuable comments. Lastly, thanks to Malin Jensen and Vegard Tørstad for insightful discussions. All remaining errors are my own.

Data sets and codes are available upon request.

Abstract

The large discrepancies in the fraction of male and female tenured academics have been subjected to a vast amount of research. Among the explanations of the male dominance are differences in preferences between men and women, structural conditions that favor men and mere discrimination. Within the field of economics, Sarsons (2017) assigns the disparity in promotions to tenured positions, to differences in recognition for joint work. Her findings suggest that male economists experience an equal raise in probability of promotion of writing an extra paper, independently of whether they collaborate on publications. Women attain the same increase in probability when publishing individually, but are substantially less credited for joint work. Similar analyses in sociology does not reveal the same coauthor penalty for women. This leads her to believe that the alphabetical order of coauthors in academic economics and ordering per contribution in sociology constitute the difference.

This thesis explores the notion of coauthor penalty by two means. I first relate coauthor penalty to existing economic theories of discrimination. Coauthor penalty fits well with the concept of statistical discrimination. That is, discriminatory behavior based on the use of heuristics under limited information. I find that coauthor penalty is largely in line with Phelps' (1972) seminal model of statistical discrimination. Furthermore, Sarsons' (2017) results simile those of empirical literature claiming to identify other cases of statistical discrimination.

Secondly and most importantly, the thesis investigates coauthor penalty and the role of alphabetical ordering of coauthors by reporting the results of a randomized experiment. By letting the participants act as employers I, jointly with Karine Nyborg and Kjell Arne Brekke, study how experiment participants' hiring behavior differ within and between information schemes. Kjell Arne Brekke programmed the experiment in Z-Tree (Fischbacher, 2007). The analyses were performed in StataSE 14.

In part 1 of the experiment, the participants performed a series of mathematical quizzes. In each quiz, they were randomly assigned a partner. We did not provide them any information on their own score, their partners name or his/her score. On two occasions, in part 2 and 3, the participants were asked to choose partners based on information from part 1. In part 2, we showed each participant a table containing the names of four other subjects, the name of their partner in each quiz in part 1 and their joint score. That is, they did not observe the individual score of the four subjects. We provided some participants with tables where the pairs were

ordered alphabetically (alphabetical treatment). The rest observed tables where each of the four subjects' pairs were listed according to who did best (first-author treatment). The best one was placed first. We asked them to choose two team members who would earn them money in the subsequent quiz. In part 3, all the participants were shown similar tables as in the previous part, but could observe the four subjects' individual score in each quiz from part 1. We asked them to pick one team member for the following quiz. In the last part of the experiment they answered two questions: who did worst and who did best in the quiz in part 3? They were allowed to choose from three candidates.

I find that subjects consistently use performance variables when assessing candidates in part 2 and 3. Furthermore, there is no evidence of coauthor penalty for females. When information on performance is absent, the subjects make decisions independent of gender.

Moreover, I find no evidence suggesting that females are less likely to be chosen relative to males in the alphabetical compared to the first-author treatment, controlling for available signals on performance. In both treatments, subjects seem to choose according to the score of the candidates and, in the first-author treatment, the number of times a candidate is listed first.

Table of Contents

1.0 Introduction

In the labor market, males dominate some sectors or industries while women constitute the largest share of workers elsewhere. One sector where the gender[1] gap has been particularly persistent is academia. In Norway, for instance, men have traditionally held the largest share of academic positions (Vabø, Gunnes, Tømte, Bergene and Egeland, 2012). Even though women constituted around half of the PhD candidates in Universities and Colleges in 2011, men held approximately 60 % of the associate professor positions and 80 % of the full professorates (Vabø et al., 2012). This pattern is also evident across disciplines (Vabø et al., 2012) and in many European countries (Goastellec and Pekari, 2013).

If we look to the US and the field of economics, the findings are quite similar to the Norwegian: the discrepancy in the share of male and female economists gets larger as one moves up the career ladder (see Vabø et al., 2012; McElroy, 2016). In 2015 at 124 American economics departments, the share of female PhD candidates was 35 % (McElroy, 2016). Moreover, women only constituted 24 % of tenured associate professors and 12 % of full professors (McElroy, 2016).

In this thesis, I address one issue that might contribute to explain why women attain fewer tenured positions in the US, and possibly in Norway. Specifically, I take a closer look at potential differences in recognition for group work between men and women in disciplines where researchers' contribution to joint work is unclear. More precisely, I address the question of how female economists may be disadvantaged by coauthoring and how it relates to the alphabetical listing of coauthors, which is the salient method of ordering coauthors.

The starting point of this work is Sarsons´ (2017) study of the effect of co-authoring on the probability of getting tenured in academic economics. She finds that female economists are less likely to get tenured if they coauthor relative to writing alone. On the other hand, men are equally likely to be promoted regardless of whether they coauthor or not. She does not find this pattern within sociology, where authors are listed per contribution. Thus, even though sociology might not be a suitable counterfactual, a possible consequence of alphabetical

[1] Note that I exclusively use the term gender until chapter 5, as the term is used when referring to social interaction (see American Psychological Association, 2010). Furthermore, using sex and gender interchangeably might confuse the reader, which is a concern that overrides potential misuse.

ordering is that employers give men more credit for joint work when size of contribution is unclear.

Jointly with Karine Nyborg (UiO) and Kjell Arne Brekke (UiO), I undertake an experiment meant to mimic assessment of joint and individual work[2]. We address the question of recognition of women's contribution in collaborative work, coauthor penalty and the role of alphabetical listing. By conducting a randomized experiment, we contribute to the understanding of coauthor penalty in at least three ways. Firstly, we see whether female subjects in the lab experience a penalty for collaborating when signals of ability are blurry or absent. That is, when the product of collaboration is the only available signal of ability or no information is displayed. Secondly, we address the role of alphabetization as opposed to listing per contribution by implementing these randomly. Randomization is naturally an important feature of our design since it allows us to treat outcomes in the two treatments as appropriate counterfactuals. Thirdly, while Sarsons (2017) used American economics departments, we use a sample of Norwegian students. There might be interesting differences.

In terms of relevance, identifying coauthor penalty is interesting as it might be one explanation to why women hold fewer positions in academic economics. Moreover, if women are less recognized for their joint work when they are listed alphabetically relative to when they are listed per contribution, one may want to change practice to the latter. At least with regard to women's recognition.

Subordinate to the aim of investigating coauthor penalty and influence of alphabetical listing – as opposed to listing per contribution – I link coauthor penalty to the economic literature on discrimination. Traditionally this strain has been concerned with identifying cases of discrimination (Guryan and Charles, 2013). That is, identifying unexplained gaps in wages or hiring rates that are likely to relate to discrimination. However, in later years there has been an increasing focus on different types of discrimination. Both theoretical and empirical researchers have mainly focused on the distinction between discrimination based on the preferences (such as animus) of the employer, taste-based, and that of using heuristics under

[2] Nyborg and Brekke are behind the idea and they have developed the design and instructions. I have contributed to the review of design and instructions, and the conduction of the experiment. In addition, I have performed the econometric analyses. Note that the experiment is a pilot-study with quite few observations from the outset. Additionally, due to a technical problem, we lost one of four sessions. I discuss challenges connected to few observations in chapter 5, 6 and 7.

limited information, statistical discrimination (Guryan and Charles, 2013). As the coauthor penalty seems to strike women when signals of intellectual contribution are blurry (or non-existent), I will mainly address the latter form.

Distinguishing these types empirically may have great policy relevance. As noted by Guryan and Charles (2013), countermeasures might differ substantially depending on the employer's motivation for discriminating. If it boils down to the difficulty of identifying workers´ skills, one would perhaps improve evaluation techniques rather than using resources trying to change the attitude of employers, if possible.

First, in chapter 2, I present some possible explanations as to why women are underrepresented in academia and Sarsons' (2017) main findings concerning coauthor penalty and the role of alphabetization. Thereafter, in chapter 3, I discuss how coauthor penalty relates to the two central theoretical concepts of discrimination, taste-based and statistical discrimination. I also exemplify how the latter type is usually identified in empirical work and potential conceptual and methodological challenges in such identification. In chapter 4, I describe the instructions and hypothesis, followed by the results in chapter 5. I devote chapter 6 to the results in relation to Sarsons' (2017) findings, and the conclusion to "wrap things up".

2.0 Gender differences in academia – some explanations

Various explanations have been put forth to explain why women hold fewer tenured academic positions in general. Hovdhaugen et al. (2004) present three possible reasons that are repeatedly suggested in the literature. Firstly, they point to traits of the hiring process. In an environment dominated by men, women's research might be less valued and considered less relevant (Hovdhaugen et al., 2004). Connected to this first point, Knights and Richards (2003) suggest that the academic environment reproduces a masculine narrative that favors men. They argue that the academic career path and assessment criteria are tailor-made for men. Moreover, De Paola and Scoppa (2015) find that men may have higher propensity towards hiring other men. Secondly, as men usually constitute the largest share of employees, women might experience more difficulty in being integrated in the work environment and consequently have less access to collaborative academic networks (Hovdhaugen et al., 2004). This may lead to fewer publications and potentially lower research quality. Lastly, Hovdehaugen et al. (2004) note that childbirth and caretaker responsibility hinder women from competing with men on equal grounds (see also Ginther and Kahn, 2006). Furthermore, female researchers may prioritize teaching and other non-research related tasks to a greater extent than men (Hovdehaugen et al., 2004). Also, since the academic environment is quite competitive, a different explanation might be that women avoid rivalry (Niederle and Vesterlund, 2007).

2.1 Publications as signals of ability

As noted above, factors leading women to engage less in research related activities may be important contributors to why women are underrepresented in academia. So are potential perceptions that assign less relevance to women's academic interests and perspectives. The quality assessment of papers and the rate at which they are published, are critical signals of productivity, and thus important determinants of whether one gets a job or is promoted. It is trivial that scholars with many publications in renowned journals have a (ceteris paribus) higher probability of getting tenured than those with fewer such publications[3].

As noted above, Sarsons (2017) addresses the gender tenure gap within academic economics in the US and links it to differences in recognition for coauthored papers and alphabetical ordering of coauthors. Her results suggest that female researchers seem to get less credit for coauthoring with men in the field of economics. One idea is that alphabetical listing of

[3] For instance, see Lynch (2006) on the value of publications in the academic sphere.

authors in economics blurs signals of ability and lead employers to use their priors of men and women´s abilities (Sarsons, 2017).

Previous investigations on the effect of alphabetical ordering has mainly been focusing on whether there is an advantage of being Professor A relative to Professor Z; that is, whether the actual order matters (see e.g. van Praag and van Praag, 2007; Einav and Yariv, 2006[4]). According to Sarsons' (2017) findings, the alphabetization might also affect women's chances of promotion. This is not connected with the order per se, but rather the fact that the order is unrelated to contribution. Thus, less knowledge about the researchers intellectual contribution to a paper may lower female economists' chances of moving up the career ladder. I use the next section to summarize her findings.

2.2 Sarsons Explained

By using CV information of economists who was up for tenure in the period 1984-2014, Sarsons (2017) investigates the relationship between; 1) number of co-authors and tenure, conditional on gender and 2) gender and tenure, conditional on coauthors. She aims at isolating these correlations by assessing the quality of the published papers and she views the results in light of number of presentations and collaborations with senior faculty. Her main results are listed below:

Influence on tenure:	Results:
Number of coauthors	Individuals with mostly solo-authored articles have a higher probability of getting tenured than people with a higher fraction of co-authored papers
Gender and coauthoring	Women with few solo-authored papers have a lower chance of getting tenured than their male counterparts. The tenure gap narrows as the signal from the solo papers begins to outweigh the penalty.
Gender of the coauthor	Women are especially punished for writing with men. That is, the increase in probability of tenure for writing

[4] Both studies find that economists with surnames ranked early in the alphabet are more credited for joint work relative to those with names ranked later. In addition to being placed first in a reference, Einav and Yariv (2006) suggest that the "et al." convention (author1 et al. (year)) play an important role.

	another paper with a man is significantly lower than that of writing alone or with other women.
Coauthoring in sociology	There is no unexplained gender gap in promotion to tenured positions in sociology (where authors are listed per contribution).

Sarsons (2017) also finds that women present their work as much as men and do not, on average, collaborate more with senior faculty – which could have lead employers to believe that they take the role as an assistant.

Note that the comparison with sociology might also lead us astray. Individuals' perception of gender difference in sociological competence might be divergent from that of gender and economic competence. Sociology may not be a fitting counterfactual. For instance, since economics is a math intensive discipline, economists may possess stereotypes connecting mathematics and males[5], while sociologists may not. Thus, suggesting that female sociologists would have experienced the same coauthor penalty had they been listed alphabetically is a weakly founded proposition. This is not to say that listing per contribution would not lead to higher tenure rates for women in economics, but it does shed light on the variety of potential differences between sociologists and economists.

Regardless of whether alphabetization is an important explanation for coauthor penalty, female economists seem to be discriminated against when the size of their intellectual contribution is not clear (Sarsons, 2017). I elaborate on the link between coauthor penalty and existing concepts of discrimination in the next chapter.

[5] For information on the link between math and males, see for instance Reuben, Sapienza and Zingales (2014) and Nosek et al (2009)

3.0 Economics of Discrimination

3.1 Theoretical foundations

Broadly defined, discrimination is the mechanisms through which individuals of different groups, be it race, gender, politics, etc., are treated differently, given that they have the same qualifications (Pager and Shepherd, 2008). More accentuated, everything else equal, an employer will base his/her decision of hiring, wage or promotion on the abovementioned characteristics.

The definition implies that employers' interpretations of signals sent out by the employee are important with regard to discrimination. We can define a signal as anything a person emits of information about him- or herself. In terms of human interaction, especially two traits about signals are important. Firstly, some signals relate to other attributes, others do not. For instance, if one observes a person with big feet, it is reasonable to assume that this person uses big shoes. On the other hand, claiming that there exists a link between having big ears and being a good listener is a bit more controversial. Secondly, the perception of signals and their meaning may vary between individuals. For example, what some perceive as dark skin or long education, others might regard as brown skin and short education (Charles and Guryan, 2011). That is, people have different relative measures when interpreting signals.

3.1.1 Taste based and statistical discrimination

Acknowledging differences in perception and connotations of signals are very important with regard to discriminatory practice. It means that quite different reasoning can cause discrimination. The observations above are trivial, but they allow us to go past the broad definition above. That is, discrimination is not only reasoned by animus towards certain groups. People interpret and act on signals differently.

In economic literature, discriminatory practice in the labor market is explained by several group characteristics, but focus has mainly been fixed upon two types of mechanisms. Discrimination based on an agent's taste or distaste for certain groups is called taste-based discrimination (Becker, 1995). Choices based on what is believed to be a correlation between certain groups and productivity – when the actual productivity is unobservable - is called statistical discrimination (Phelps, 1972; Arrow, 1973, Aigner and Cain, 1977).

Becker's (1995) seminal work "The Economics of Discrimination" goes into detail of how discriminatory behavior arises in the market and how market competition will make

discriminating firms along with the practice itself, perish. In his analytical framework, he explains what he calls "a taste for discrimination". If an employer has a taste for discrimination, he "[…] must act as if he were willing to forfeit some income in order to avoid certain transactions" (Becker, 1995, p. 16). That is, if an employer avoids hiring certain groups he acts on this taste. It is interesting to note that the concept of discriminatory tastes, in Becker's understanding, include both animus towards certain groups and decisions based on incomplete information or ignorance (as he calls it). However, in later years it has generally come to mean the aforementioned (e.g see conceptualization in Charles and Guryan, 2013).

Arrow (1973) and Phelps (1972) address discrimination that is due to limited information about abilities and productivity of workers. They suggest the possibility that instead of using costly analyses to retrieve more information, employers base decisions on easy observables such as gender or race (Arrow, 1973). That is, employers use costless signals as predictors of productivity (Phelps, 1972).

Sarsons (2017) notes that employers do not seem to have, as Becker calls it, a taste for discrimination. If they were to act on such taste or animus, one would expect the employers to recognize women's solo authoring to a lesser degree than they do men. Instead, Sarsons (2017) finds that employers only disfavor women when the size of intellectual contribution is less clear. That is, if they coauthor. Thus, even though she does not comment on it herself, statistical discrimination might be at play.

Before turning to the empirical literature, I will apply Phelps' (1972) theory of statistical discrimination to coauthor penalty. This is to show what we would have to be willing to assume to link Sarsons' (2017) findings to statistical discrimination.

3.2 A statistical theory of coauthor penalty[6]

Suppose that some economics department is considering promoting two of their employees to some tenured position. The committee observes several characteristics connected to the human capital of the applicants, such as CV, experience, education, teaching hours, publications, etc. They also observe the applicants' gender. For each applicant, they add these

[6] The basic mathematical framework can be found in Fang and Moro (2011). The two cases below are proposed by Phelps (1972).

features and end up with assessing the subject with some score for individual i,y_i, on an index of past academic performance, where

$$y_{ig} = q_{ig} + u_{ig}$$

Where q_{ig} is the applicants' actual score, u_{ig} is the error term and subscript $g = \{W, M\}$ refers to Woman and Man. Assume that the distribution of ability and error is the following

$$q_{ig} \sim N(\mu_{qg}, \sigma_{qg}^2)$$

$$u_{ig} \sim N\left(0, \sigma_{ug}^2\right)$$

Thus, the qualification and error terms are normally distributed. The distributions for men and women are independent of each other. To address the coauthor penalty in particular, suppose that both q_{ig} can be divided into research ability r_{ig} and the rest $q_{ig} - r_{ig}$, and y_{ig} into publishing p_{ig} and the rest $y_{ig} - p_{ig}$. Furthermore, assume that the committee already observes and evaluates qualities related to teaching and other tasks with high degree of precision. Conditional on these qualities suppose we are left with an assessment of research abilities and the new error term l_{ig}.

$$p_{ig} = r_{ig} + l_{ig}$$

Assume that r_{ig} and l_{ig} have the same qualities as q_{ig} and u_{ig} respectively. So the applicant's research abilities are evaluated with the produce of research as proxy. The committee evaluates the following term of expected ability (Fang and Moro, 2011):

$$E\left[r_{ig} | p_{ig}\right] = \frac{\sigma_{rg}^2}{\sigma_{rg}^2 + \sigma_{lg}^2} p_{ig} + \frac{\sigma_{lg}^2}{\sigma_{rg}^2 + \sigma_{lg}^2} \mu_{rg}$$

The expectation of research ability, conditional on publications, is a weighted average of the research an applicant has conducted and the average in each group. That is, the averages in the pool of female and male researchers.

Let us first assume that a single-authored paper is a perfect predictor of research ability: σ_{lg}^2 goes towards zero. They are rewarded according to the quality of their research. In the noisy case – coauthoring – Phelps (1972) notes that there are at least two actual or perceived differences between men and women that lead to statistical discrimination. Note that I look at the case where quality and number of papers are equal. For example, imagine that a

committee assesses two equally qualified economists only differing in terms of gender. Thus, the assessment of relative abilities of candidates depend on μ_{rg} and σ_{lg}^2.

Case 1: $\mu_{rM} - \mu_{rW} > 0$

The disturbance term, σ_{lg}^2, is relatively large unless each author's contribution is specified. Assume that it is also equal for men and women. The hiring committee will accordingly put more weight on the expectation, μ_{rg}. If the expectation of r_{iM} is higher than the expectation of r_{iW}, the committee will prefer the male candidate when signals are blurry. However, this does not explain why there are seemingly no differences in probability of tenure between men who coauthor and those who write alone. Case 2 might remedy this.

Case 2: $\sigma_{lW}^2 > \sigma_{lM}^2 = \epsilon$

In the second case, Phelps (1972) suggests that the expected value of ability (or productivity) is equal, but the signal from one group is associated with less variability. Thus, we can imagine that the hiring committee is surer that the publications of men actually do correspond to their research ability. Women on the other hand suffers from collaboration as the committee does not trust that the paper reflects their real abilities. Thus, in this case the different treatment of applicants with equal resumes concerns how the committee evaluates the reliability of men and women's publications as predictors of research ability. This might also explain why men's probability of tenure seems to be independent of the number of coauthors. The quality of the paper still reflect the male economists' ability regardless of whether he is collaborating with one, two or three others.

One point that disfavor the explanation of statistical discrimination is that writing alone might not be a perfect signal of ability. That is, maybe the model assumes too little variance in terms of correlation between single-authoring and academic ability. If one were to assume lower reliability of a single authored paper and that $\mu_{rM} - \mu_{rW} > 0$, we would not expect women to get the same credit as men for a single authored paper. Suppose instead that $\sigma_{lg}^2 > 0$ of a single authored paper and $\mu_{rM} = \mu_{rW}$. Then this would also imply equal recognition for joint work. The predictions fail either way. Thus, this is a serious pitfall if concluding on statistical discrimination.

3.3 Empirical literature

In the empirical literature, a first observation is that many researchers aim at detecting labor market discrimination in the broad sense. That is, exploring whether characteristics like gender and race disfavor some individuals in certain sectors or industries (see Bertrand, Mullainathan, 2004; Cain, 1986; Reimers 1983; Riach and Rich, 2010). Still, many studies have a design that may suggest what type of discrimination is at play (see Nunley, Pugh, Romero and Seals, 2014; Kaas and Manger, 2012; Altonji and Pierret, 2001).

As in the theoretical literature, empirical studies have mainly been concerned with the distinction between taste-based- and statistical discrimination. I will direct sight on some of the contributions on statistical discrimination[7]: their use of model implications and interpretation of data. This focus is due to Sarsons' (2017) own observation that taste-based models are not aligned with her results and that the coauthor penalty seemingly concerns limited information as expressed through the model above.

Generally, the aim of type studies is to see whether different treatment of employees persist when employers receive better signals of their abilities (Guryan and Charles, 2013). The key implication of Arrow's and Phelps' models is that individuals would not have discriminated if they had clear sight on the relevant factors concerning their decision. Their discriminatory actions are in reality consequences of associations between group affinity and productivity or ability (or other characteristics if outside the labor market). Thus, observing actual productivity and not only a resume would make discrimination disappear. The studies below investigate this indirectly in the sense that the results infer something about what lies behind the decisions of the actor. Common for all of them is the idea that statistical discrimination is likely at play if agents change their behavior in the face of better information about the individuals they assess.

Altonji and Pierret (2001) study young white and black men in their first years in the labor market. They develop a test to identify statistical discrimination under the assumption that employers update their beliefs about workers as time goes. The idea is that statistical discrimination is at play if wages are increasingly correlated with characteristics that are hard to observe, such as productivity (Altonji and Pierret, 2001). Firms in their sample do set

[7] Examples of studies concerning taste-based discrimination are Charles and Guryan, 2008, Baert and De Pauw, 2014, and Mobius and Rosenblat, 2006.

wages that become increasingly correlated with productivity as time goes by. They interpret the initial discrimination as statistical of sort.

Knowles, Persico and Todd (2001)[8] develop a model of police searches for contraband in vehicles. They assume that the police are maximizing the numbers of arrests. That is, the police search groups (for example racial groups) where drugs are most likely to be found. Furthermore, they assume that the most frequently searched groups will respond by carrying less drugs. Thus, in equilibrium the probability of finding drugs should be equal across racial groups. The implications are the following: if probability of finding drugs are equal across groups and one group is searched more frequently than others are, this indicates statistical discrimination. On data from Maryland, US, they find that the pattern in vehicle searches are in line with the predictions of the model.

Kaas and Manger (2011) design a field experiment that is aimed at identifying the same mechanism, but in the employment phase. First, they randomly assign abilities to fake job applicants with either foreign or native (German) names, and advertisement for student internships. Then, call back rates give an indication of discrimination. In the cases where reference letters are not included, there is a significant positive difference in the numbers of callback between those with German- and foreign sounding names, while they are equally likely to be called back when reference letters are included. The authors find this as suggestive of statistical discrimination.

Castillo and Petrie (2010) perform a public good experiment to investigate if and how discrimination occurs in group formation. First, the participants perform several rounds where they invest an initial endowment to their private fund or a public fund shared with 4 other randomly selected subjects. Thereafter, they rank the other subjects according to with whom they want to collaborate. In this ranking process, they are either shown photographs of the other subjects, subjects' past investments or both. When only photographs are observable, the participants consistently rank black subject lower than other groups. On the other hand, when information on past behavior is observable, subjects rank according to payoff relevant information. Castillo and Petrie (2010) concludes that statistical discrimination is at play.

[8] Not a labor market model, but relevant test of implications.

3.4 Pitfalls in identification of statistical discrimination

The conclusions above do support the idea that statistical discrimination might be an explanation to Sarsons' (2017) findings. That is, it might suggest one explanation to why women appear to suffer from coauthor penalty when signals are blurry. Furthermore, with the variation in clarity of signal in our experiments, it is tempting to draw conclusions regarding the presence of statistical discrimination. However, there are some caveats, both conceptual and methodological, challenging such a temptation and questioning some conclusions made in the past.

First, statistical and taste-based discrimination are not exhaustive explanations of discrimination. There might be other compelling explanations such as less intentional types of discrimination. For instance, Bertrand, Chugh and Mullainathan (2005) explores the notion of implicit discrimination. This concept, originating from social psychology, concerns unintentional discriminatory behavior (Bertrand et al., 2005). That is, employers acting as "objectively"[9] as possible might cause outcomes of discrimination. Thus, even though the story of statistical discrimination may seem compelling, one should as an economist, be open to other explanations of discriminatory behavior.

Neumark (2016) finds in his meta study of field- and laboratory experiments that conclusions on what type of discrimination is at play, are generally not well founded. He presents two reasons for this difficulty. Firstly, there may be interplay between different types of discrimination (Neumark, 2016). That is, some employers may use easy observable traits, such as gender, as proxy for productivity and at the same time act on taste. Thus, the belief that employers make decisions based on one type of discrimination exclusively should be well founded. Secondly, they are simply hard to isolate methodologically (Neumark, 2016). Finding evidence of one form of discrimination does not necessarily exclude others. Ergo, we face a conceptual challenge in addition to the methodological one. Namely, how to further develop implications that not only identifies a type of discrimination, but also exclude other types.

[9] In the sense that he/she tries to assess the employee based on available information and not prejudice.

4.0 Instructions and Hypotheses

In this section, I describe the details of the experiment and the instructions. Thereafter, I formulate the hypotheses relevant to our investigation. Throughout the rest of the thesis, I refer to participants as candidates when they are affected by a decision.

4.1 Description of instructions[10]

First, we informed the participants about general rules: no communication and no mobile phones. They were also told the following: that their decisions were anonymous, how to proceed from one part to the next and that the reward system was performance-based. Prior to part 1, we asked the participants to choose a nickname to preserve their anonymity. They answered the following question: "Imagine that you would have a different first name. What name would you prefer to have?" We requested that they should choose a relatively common name followed by a capital letter to ensure that names were different. For example "Anne K".

In part 1 the participants performed five math quizzes in total. The exercises were variations of adding two and three ciphered numbers and subtracting two ciphered numbers. They were asked to answer as many exercises as possible within a time frame of 60 seconds. Since they had to push an "OK" button before the time ran out in order for the answers to be registered, we notified the participant when they had 5 seconds left.

In each quiz the participants were randomly assigned a partner. For instance, of 22 participants in one session there would be 11 unique pairs in each round. They undertook the quizzes individually, but were informed that they were paired with one other partner. The participants were not updated on their partners' nicknames or their individual or joint score. The payment scheme was as follows:

> Both you and your partner get 1 NOK for each correct answer you give. This applies independently of whom answers correctly, and independently of whether you answer the same answer correctly or not. For example, if you provide 10 correct answers and your partner provides 8 correct answer, you get 18 NOK each in that round.

In part 2 we implemented two treatments: first-author treatment and alphabetical treatment. Individuals in both treatments were displayed a table of four other, randomly selected participants and their joint score with their partner in each of the five rounds in part 1. They

[10] Full instructions available in the appendix.

were also shown the name of the candidates' partners. Participants in the first-author treatment observed a table where each pair was listed per contribution, while those in the alphabetical treatment observed pairs listed alphabetically. For example, the placement of Anders was dependent on his score relative to his partners' score in the first-author treatment (table 1a). If he were in the alphabetical, he would be listed conditional on his partners' names (table1b). The explanations to each treatment group were as follows:

> 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.)
>
> Alphabetical treatment: For each pair the names are ordered alphabetically.

Table 1: Picking two candidates

a

Anders S	.	.	.
Anders S & Leif P – 28	.	.	.
John K & Anders S – 30	.	.	.
Anders S & Ane L – 24	.	.	.
Nina M & Anders S – 26	.	.	.
Jens N & Anders S – 20	.	.	.

b

Anders S	.	.	.
Anders S & Leif P – 28	.	.	.
Anders S & John K – 30	.	.	.
Ane L & Anders S – 24	.	.	.
Anders S & Nina M – 26	.	.	.
Anders S & Jens N – 20	.	.	.

Notes: table 1 a and b are examples of what information the participants receive in part two. The tables display information about one of four candidates. Each pair is listed according to score in panel a (first-author treatment) and alphabetically in panel b (alphabetical treatment). The number is joint score. Four such columns were shown to the participants, as indicated by the dots. The subjects chose two of the four candidate in the upper row. See the screenshot in the instructions for exact format.

The participants were informed about their own treatment exclusively. They were asked to choose two team members who would earn money for them in the following quiz:

> 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. Then you have 5 seconds to push "OK" and thereby saving your answers.

The participants were also informed about the payment scheme. For each correct answer they provided, they got 1 NOK. Each correct answer provided by their teammates earned the participant 3 NOK. Thus, if a participant had 10 correct answers and her team members had 5 and 20 correct answers, respectively, she got 85 NOK (10+3· (25)=85).

Prior to the quiz we asked the participants questions regarding a toy table to see whether they understood how to retrieve information and the listing of the individuals within each pair. They received three questions about a table. They regarded whether they could read off the score from the table, identify who was paired with whom, and whether one could infer who did best in a pair based on the ordering.

Part 3 was similar to part 2 except that the participants were asked to pick one team member. First, we presented each participant with information on four candidates' individual scores from one of the quizzes in part 1. Table 2 serves as example. Acting as employers, the participant chose one candidate as team member for the subsequent quiz. We gave them the following instruction:

> The first thing you will do in part 3 is to pick this person [the candidate]. 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.

Subsequently, the participants performed a math quiz. As before, they got 60 seconds to solve as many exercises as possible. After 55 seconds, we informed them that they had 5 seconds left to save their answers. The payment scheme was equal to that in part two: 1 NOK per correct answer provided by themselves and 3 NOK per correct answer given by the teammate.

Table 2: Picking one candidate

Navn:	Kand1: Anders S	Kand2: Jenny V	Kand3: Lars I	Kand4: Pål L
Score runde1	10	12	11	13
Score runde2	9	10	12	14
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

Notes: Each cell corresponds to the individual score of the participants in each of the five quizzes in part 1. See figure A1 in appendix for screenshot.

In part 4, the participants were shown the nicknames of three other participants. We asked each one of the participants to pick one candidate as having performed best and as having performed worst in the quiz in part 3. We gave them no information of previous performance.

Only their nicknames were exhibited. They were rewarded 10 NOK per correct answer they provided.

Lastly, we asked the participants to report their sex, age and faculty affinity. Furthermore, we asked the participants whether candidates with names early in the alphabet were more likely to be picked.

4.2 Hypotheses

Below I formulate hypotheses that I put to the test based on the decisions in part 2, 3 and 4.

Hypothses:

Part 2

Treatment

H_0^1 : Gender is an equally important explanation of choice in both treatments.

H_1^1 : Gender is a more important determinant of choice in the alphabetical treatment relative to the first-author treatment.

Coauthor penalty

H_0^2 : Gender is not a statistically significant determinant when choosing team members.

H_1^2 : Gender is a statistically significant determinant when choosing team members.

Part 3

H_0^3 : Gender is not a statistically significant determinant when deciding on whom to choose.

H_1^3 : Gender is a statistically significant determinant when deciding on whom to choose.

Part 4

H_0^4 : $E\left[\frac{Number\ of\ females\ picked}{Number\ picked\ subject}\right] = \frac{Number\ of\ females\ displayed}{Number\ of\ names\ displayed}$, in category "Best" and "Worst"

H_1^4 : $E\left[\frac{Number\ of\ females\ picked}{Number\ picked\ subject}\right] \neq \frac{Number\ of\ females\ displayed}{Number\ of\ names\ displayed}$, in category "Best" and "Worst"

The hypotheses for part 4 regard whether the fraction of women in the categories "Best" and "Worst" is significantly different from the fraction of female candidates among all candidates that can be chosen.

5.0 Results

We recruited 76 students from the University of Oslo to participate in the experiment. The pool of participants consisted of students from different disciplines. Kjell Arne Brekke programmed the experiment in Z-tree (Fischbacher, 2007). Econometric analyses were performed in StataSE 14.

The gender corresponding to each nickname was determined using Nordic Names[11]. It indicates whether the name is a female or male name. Note that three subjects reported that they were female, but their nicknames were regarded male. Moreover, two subjects reported male as their sex, but were considered female based on their nicknames. Be aware that I use sex when referring to male and female participants, as reported by themselves, and gender about the categorization of candidates, as defined by the researcher.

In total, we conducted 4 sessions. Table 1 displays information on each session and the corresponding treatments.

Table 3

Session	Subjects	Mixed-gender quartets: Part 2	Mixed-gender quartets: Part 3	Mixed-gender trios: Part 4	Treatment
Session 2	22	19	20	16	Per Contribution
Session 3	18	15	17	15	Alphabetical
Session 4	16	16	15	13	Per Contribution

Notes: table 3 gives an overview of the number of subjects, the number of mixed-gendered groups at each decision and which sessions received one or the other treatment.

Due to a technical problem, we lost all observations from the first session. Within the timeframe of this master's thesis, we were unable to restore the results. Thus, we have few observations from the alphabetical treatment. Consequently, the hypothesis on differences between listing alphabetically and per contribution is difficult to test. Nevertheless, I will perform the analyses. I comment on potential challenges. Note also that I use "total score" when referring to the sum of candidates' joint score with their partner in quiz 1-5, part 1, and

11 Four names did not yield results in Nordic Names. Two fellow students unanimously characterized two of the names as male and female. One of the students regarded the last two names, Coffe and Petry, neutral. The other regard them male and female, respectively, which correspond to their self-reported sex. Dropping these subjects would entail dropping the subjects who could hire them in part 2 and 3. Thus, considering the lack of observations I define Coffe as male and Petry as female.

"total individual score" when referring to the candidates' sum of individual scores in the same rounds.

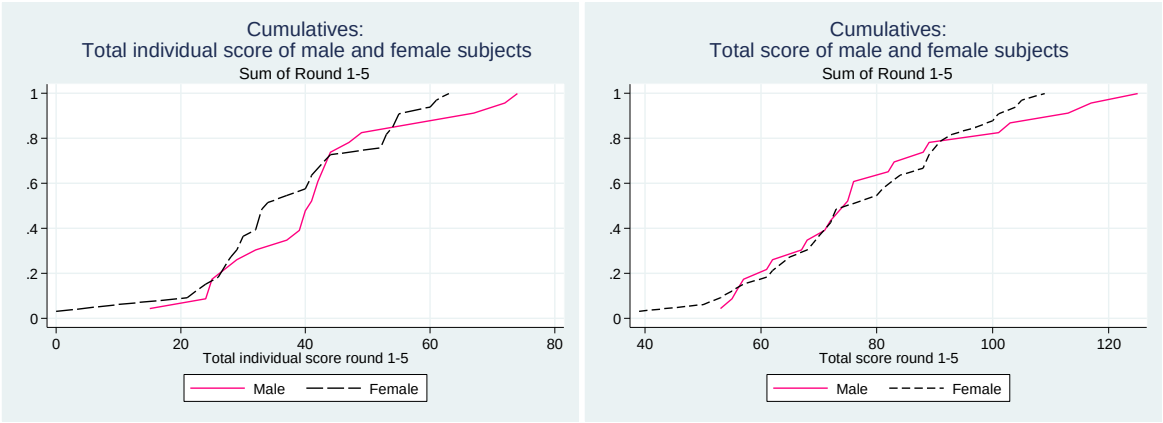
5.1 Differences in performance

Result 1: Females and males (sex defined by themselves) do not perform significantly different in part 1, neither individually or with their partner.

Figure 1 shows the cumulative distribution of the female and male subjects' total individual score when summing their score in round 1-5 in part 1. The average score of female subjects was 37.5 while men on average answered 41.7 exercise correctly. The standard deviation of men's score (15.1) is slightly higher than that of women's (14.7). The nonparametric Kolmogorov-Smirnov test, does not support the rejection of the null hypothesis that men and women have equal distributions (p-value=0,584). A Wilcoxon rank-sum test leads to the same conclusion (p-value=0.405). Thirdly, we cannot reject equality of means based on a standard two-sided t-test (t=1.0364).

The correlation between total individual scores and total score is 0.9037. Thus, I would expect the same pattern in terms of total score. To see whether there are large gender differences in the total scores – that is, when summing the scores of the pairs from round 1-5 in part 1 – I perform the same tests on these distributions. Figure 1 shows the cumulative distribution of male and female subjects. The mean for women is 77.8 while the average total score for men is 79.3. As with the individual score, the variation is slightly bigger in the male pool (std.dev.=20.5) than for females (std.dev.=17.6). Nonparametric tests of the null hypothesis of equal distributions yields p-values of 0.975 and 0.954 for the Kolmogorov-Smirnov and Wilcoxon rank-sum tests respectively. Furthermore, a two sided t-test does not give reason to reject the hypothesis of equal means (t=0.2929).

Figure 1: Cumulative distributions of score in part 1



Notes: The left panel shows the cumulative distributions of male and female participants' total individual scores in part 1. The right panel displays the cumulative distributions of the females and males total scores.

We see that male participants performed slightly better than females on the mathematical exercises in part one. However, the difference was not significant. This is not an unexpected result as female high school students perform almost equally well as their male counterpart (Grønmo, Hole and Onstad, 2015). It is important to note these distributions are based on the sex reported by the subjects.

5.2 Decisions in Part 2

Result 2: H_0^1, that gender is an equally important explanation of choice in both treatments, cannot be rejected.

Result 3: H_0^2, that gender is not a statistically significant determinant when choosing team members, cannot be rejected.

Table 4 shows the fraction of females among the chosen candidates by treatment. Firstly, overall we see that subjects choose more females (0.61) than males (0.39). Secondly, within the first-author treatment 55 per cent of the chosen candidates were females. Thirdly, a striking observation is that among the chosen candidates in the alphabetical treatment 72 per cent are female.

Table 4: Fraction of females among the chosen candidates in part 2

Treatment	Fraction of chosen females
Alphabetical treatment	.72
First-author treatment	.55
Total	.61

Note: The table present the fraction of women picked in each treatment and overall.

Looking at the means of total score of all candidates – not only the chosen ones - in the two treatment groups, yields insight. In the alphabetical treatment, the mean of females' total score is 84.5 while that of males' is 77.5. However, in the first author treatment males have mean of 75.1 while females have a mean of 74.7. Thus, differences in total score is a potential explanation of the high fraction of females picked in the alphabetical treatment. As the instructions and exercises in part 1 were equal for all subjects, the difference in performance is most likely completely random[12].

Throughout the discussion regarding choice in part 2 I utilize the candidates' total score as a measure of performance. This is due to the fact that the subjects only observed total score, not the total individual score of the candidates. Thus, potential discrimination arises from different treatment of males and females conditional on what they observe.

5.2.1 Coauthor penalty and the effect of first-author treatment

Table 5 displays OLS regressions assessing the role of treatment, sex of the subject choosing and the performance of women in terms of how many females subjects choose. *Fem/Fem Dis* is the dependent variable, measuring how many females the subject choose as a fraction of how many females are displayed to them. *First author* indicates in which treatment group the subjects are placed, taking the value one if first-author treatment and zero otherwise. *Fem Sub* is a dummy variable indicating whether the subject (choosing) is female (1) or male (0). *Num fem top 2* reflects how many female candidates were among the two best in the pool of four candidates, based on total score. That is, it is a relative measure of the females' performance.

[12] Since there were only 18 participants in the alphabetical treatment, such random differences were not unexpected.

Note that with this measure I lose information and variability. However, it preserves the performance of females relative to males[13].

The regression output tells us that treatment is statistically significant on a 10% when (past) performance is not included. It says that subjects in the first-author treatment group on average pick 0.15 fewer females per female displayed, relative to the alphabetical treatment. This is not surprising considering the fractions in table 4. When included, past performance is highly significant while the effect of treatment wears out. One extra female among the top two candidates displayed to a subject, on average, increase *Fem/Fem Dis* by 0.203. Thus, even though the measure of female performance is low on information and variability it still picks up that females performed relatively better than men in the alphabetical treatment.

Table 5: Females chosen per number of females displayed

	(1)	(2)
Dep Var: Fem/Fem Dis	b/se	b/se
First Author	-0.150*	-0.0965
	(0.0854)	(0.0788)
Fem Sub	-0.0598	-0.0393
	(0.0747)	(0.0691)
Num fem top 2		0.203***
		(0.0509)
Observations	55	55

Notes: OLS regressions. The dependent variable is number of females chosen per number of females displayed to each subject. The top panel reports coefficients, robust standard errors in parenthesis. *,**, and *** denote significance at the 10%, 5% and 1% level. The lower panel indicates the number of observations.

In addition to running the OLS from the subjects' point of view – investigating the determinants of how they pick candidates – I employ three probit regressions exploring how different variables affect the probability of being chosen. Instead of 56 observations, one for each participant, I have 224 (4*56) observations, representing each candidate. Thus, a

13 Other measures such as female's scores relative to males would not contain information on individual differences and it is sensitive to the number of females. Using the rank of the chosen individuals (i.e. either 1,2,3,4) would entail two variables that are rather meaningless as who is picked first and second is random. The mean of the rank of females and males would not be a meaningful variable as 2 and 3 give the same mean as 1 and 4.

participant might randomly be selected to be displayed to other participants multiple times, except from to him/herself. As this procedure is randomized, there should be no worries in terms of selection bias. This point of view allows me to measure the meaning of gender directly and to use the candidates' real score instead of the inferior *fem top 2* variable.

Table 6 shows two probit specifications[14]. *Chosen* indicates whether a name is picked to be a team member. *Relative Score* is the total score of the candidate relative to the mean of the total score of the four candidates displayed to each subject. *Female* is naturally the gender of the candidate. *First author* indicates treatment group and *Female x Fir auth* is their interaction. *Rel list First* is the number of times a candidate is listed first relative to the mean of the four candidates, while *Fir Auth x Rel List* is its interaction with treatment.

First, note that the *First Auth* variable is not interesting in itself. That is, being a candidate in either of the two treatments does not influence the probability of being picked apart from the fact that there are about half as many in the alphabetical treatment. Only the interaction terms are of interest as they assess whether other variables differ between treatments.

In specification 1, the score is significant on a 1 % level. The coefficients on *Female* is not statistically significant on any of the standard levels. Furthermore, specification 1 suggests that females are equally likely to be chosen in the first-author treatment relative to the alphabetical.

In specification 2 I include the number of times the candidates' names are listed first in tables displayed to the subjects. The coefficient on *Rel List First* is statistically insignificant on a 10% level. It indicates that the number of times a candidate is listed first is overall not an important explanatory variable. However, the interaction term *First Auth x Rel List*[15], in specification 2, shows that it does matter in the first-author treatment. The coefficient on the interaction term is significant on a 5% level. It indicates that being listed first in the first-author treatment has a positive effect on the probability of being chosen. The coefficients on

[14] In the following tables, I report coefficient instead of marginal effects. I am only interested in the direction of effects. Overall, the magnitude is not interesting to address the issues at hand.

[15] Correlation between relative score and number of times listed first is substantial but does not raise concern with regard to multicollinearity: 0.28 in the whole sample and 0.41 in the first author treatment. VIF score (for both predictors) of 1.09 in the whole sample and 1.2 in the first-author treatment, which means not substantial increase in standard errors by introducing the variables of being listed first.

the two additional regressors in specification 2 are jointly significant a 1% level (p-value=0.0036).

Table 6: Probability of being chosen

Dep Var: Chosen	(1) b/se	(2) b/se
Relative Score	4.091***	3.684***
	(0.563)	(0.655)
Female	0.305	0.344
	(0.298)	(0.296)
First Author	0.318	-0.567
	(0.290)	(0.478)
Female x Fir Auth	-0.504	-0.561
	(0.383)	(0.390)
Rel List First		0.0568
		(0.224)
Fir Auth x Rel List		0.907**
		(0.363)
Observations	224	224

Notes: Probit regressions with being chosen as the dependent variable. The top panel reports coefficients, robust standard errors in parenthesis. *,**, and *** denote significance at the 10%, 5% and 1% level. The lower panel indicates the number of observations.

Table 7 highlights (and possibly facilitates interpretation) the findings in Table 6. In both treatments the coefficient on gender is not statistically significant on any of the three standard levels. Furthermore, it shows that the number of times a candidate is listed first matters in the first-author treatment and not in the alphabetical treatment.

Table 7: Probability of being chosen in each treatment

Dep Var: Chosen	Alphabetical b/se	First-author b/se
Relative Score	7.763***	2.483***
	(2.085)	(0.679)
Female	0.267	-0.178
	(0.370)	(0.234)
Rel List First	0.129	1.030***
	(0.264)	(0.261)
Observations	72	152

Notes: Probit regressions with being chosen as the dependent variable. The top panel reports coefficients, robust standard errors in parenthesis. *,**, and *** denote significance at the 10%, 5% and 1% level. The lower panel indicates the number of observations.

If alphabetization were to be a factor contributing to a lower fraction of women or men chosen, we would naturally expect the coefficient on treatment in table 5 to be statistically significant from zero. However, the first-author treatments' negative effect on number of females chosen seems to be due to differences in performance.

The analyses in Table 6 and 7 supports this claim. Controlling for performance yields equal opportunities of being chosen. Thus, we see that, as the ordering signals performance, this seems to be used as proxy along with total scores. Put somewhat extremely, differences in treatment of males and females do not seem to be an innate trait of the alphabetical listing. Moreover, the overall impact of gender (across treatments) seems to be negligible in our sample of students.

Subjects seem to be consistent in their way of assessing the candidates according to performance variables. However, as in the results below, the analyses might suffer from the low number of observations. This remark especially concerns the treatment effect. A larger sample might reveal a small but significant effect. 18 subjects in the alphabetical treatment increase noise and it limits the possibility of identifying a potential small treatment effect.

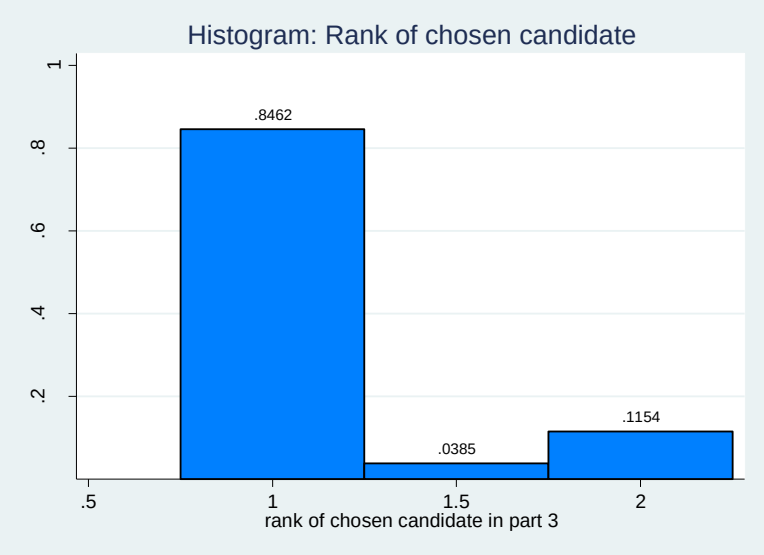
5.3 Decisions in Part 3

Result 4: H_0^3, that gender is not a statistically significant determinant when deciding on whom to choose, cannot be rejected.

The share of subjects who chose a female candidate in part 3 is exactly 0.5 among the 52 subjects who were displayed mixed gendered groups[16]. That is 50 per cent chose a male candidate and 50 per cent chose female candidates. I analyze the decisions of these 52 subjects below.

When picking team members in part 3, most subjects seem to be driven by the candidates' total individual score relative to the three other candidates score. On a variable that ranks[17] the four candidates according to total individual score in part 1, 85 % of the subjects chose the individual with the highest rank. Figure 2 shows the distribution of rank among the candidates who are chosen.

Figure 2: Distribution of rank



Notes: The bars shows the fraction of subjects choosing team members of different rank from 1 to 2. None of the 52 are rank worse than second. The bar in between the integer values indicates ties. For example, if two candidates have the best score they both get 1.5 as rank.

However, there are some differences between male and female candidates. While the mean rank of chosen females is 1.13, the mean rank of chosen males is 1.4. Thus, there are fewer female than male candidates who are picked and are ranked second.

16 The decisions of subjects who could choose from four females or four males do not yield any information to the meaning of gender. These are dropped.

17 *Rank* is constructed using the **rankrow** command in Stata. It ranks the participants according to total score in part one. 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.

Table 8[18] shows a probit regression of the probability of picking a female candidate. *Relative Score* is the total individual score of the chosen candidate relative to the mean of the four candidates from which a subject can choose. *Num Fem* is a factor variable indicating the number of females displayed to the subject. The coefficient in the table (=2, =3) are relative to being shown one female. *Fem Sub* reflects the sex of the subject (choosing).

The coefficients on the number of females displayed (*Num Fem)* and sex of the subject (*Fem Sub*) are insignificant on a 10% level. That is, more females does not change the probability of choosing a female. Nor does it matter whether the subject is male or female. Furthermore, the coefficient on *Relative Score* is not significant on a 10% level, as expected. The interpretation is that there are no statistically significant difference in *relative score* between females and males. On the other hand, if there were a substantial bias against females in our sample, we would expect statistically significant and positive effect indicating that females are consistently required to attain a higher score in order to be picked.

Table 8: Probability of choosing a female

	(1)
Dep Var: Fem Chos	b/se
Relative Score	0.785
	(0.892)
Num Fem=2	0.189
	(0.581)
Num Fem=3	0.0677
	(0.553)
Fem Sub	0.0395
	(0.381)
Observations	51

Notes: Probit regression. The dependent variable is choosing a female candidate. The top panel reports coefficients, robust standard errors in parenthesis. *,**, and *** denote significance at the 10%, 5% and 1% level. The lower panel indicates the number of observations.

Figure 2 shows that the relative score of the four candidates matters in the decision making of the subjects. On average, they choose whoever has performed best in terms of total individual

[18] The observations are subjects who are displayed mixed-gendered groups.

score. In terms of gender differences, the discrepancy in means of rank might suggest that males are not required to perform as well as females in order to be picked. However, the result on relative score in Table 8 indicates that such a difference is not statistically significant. Furthermore, the insignificance of gender composition (*Num Fem*) supports the explanatory power of performance. Still, considering the sample size, it would be a mistake to bombastically rule out gender in these kinds of decisions. The differences in mean of rank might indicate a tendency worth exploring.

5.4 Decisions in Part 4

Result 5: H_0^4, that $E\left[\frac{Number\ of\ females\ picked}{Number\ picked\ subject}\right] = \frac{Number\ of\ females\ displayed}{Number\ of\ names\ displayed}$, in category "Best" and "Worst", cannot be rejected.

Since the random draw of candidates displayed to the subject was not stratified, some tables in part 4 included males or females exclusively. Naturally, these observations are not valuable when investigating the role of gender. Thus, these are dropped from the analyses, leaving us with 43 mixed gendered trios[19]. I perform binomial tests on this sample, testing whether the probability of picking a female is significantly different from 0.527 which is the fraction of women displayed to the 43 subjects.

Among these 43, the fraction of female candidates chosen as having performed best in the quiz in part 3, is 0.49. Thus, approximately equally many subjects chose males and females. Testing (two-sided) against the 0.527-benchmark yields a p-value of 0.65. One cannot embrace the alternative hypothesis that the fraction is statistically significantly different from 0.527.

The fractions are slightly different when choosing whom they believed performed worst in the quiz in part 3. A fraction of 0.44 of the subjects picked a female. The binomial test gives no reason to doubt this hypothesis. It returns a p-value of 0.29 when testing (two-sided) whether the proportion of women is equal to 0.527.

Therefore, it seems like the subjects are, on average, not choosing based on the gender of the candidates. Facing utter uncertainty, the proportions of females assigned to the categories

[19] Naturally, as the subjects are displayed 3 candidates some are left with no females or no men after making the first pick. However, with randomized decisions we would expect the two categories to reflect the gender balance in the pool of candidates that are displayed to the 43 subjects.

"best" and "worst", are not significantly different from the fraction of females the subjects were allowed to choose from. Thus, it fits into a randomization pattern. Note, as in the previous discussion, that the amount of observations is limited.

5.5 (Fe)males picking (fe)males?

Result 6: Sex of the subjects does not affect the gender composition in part 2 or decisions in part 3 and 4.

In addition to testing the hypotheses formulated above, I also address whether there is an association between sex of the subject (choosing) and the gender of the chosen candidates. The coefficient on sex in the table 5 is statistically insignificant at 10% level. The same story applies for regression on data from part 3, Table 8, where the coefficient on sex is insignificant. Similarly, the tests performed on data from part 4 showed no statistically significant difference in how the males and females chose candidates of different gender. A two-sided t-test on the difference in fraction of females chosen by female and male subjects returned p-values of 0.30 and 0.77 in terms of best and worst performance respectively[20].

Literature concerning this mechanism is quite inconclusive in terms of how gender composition of hiring committees affect hiring or promotion. Some find results suggesting that either one or both sexes are more prone to promote or hire individuals of the same sex (see Bagues, Sylos-Labini and Zinovyeva, 2017; De Paola and Scoppa, 2015; Zinovyeva and Bagues, 2011). Others find that there is a negative correlation between number of females in the hiring committee and the probability of females being hired (see Bagues and Esteve-Volart, 2010). If we look at a similar hiring experiment[21], Reuben et al. (2014) find that female subjects are discriminated against irrespective of whether the hiring is performed by a man or a women. Consequently, the finding that the decision makers' sex is not associated with the actual choice of picking a female or a male, does not stand in stark contrast to related literature.

5.6 Synthesis of results and their limitations

We implemented two parts that forced the participants to randomize or use proxies in order to assess candidates, part 2 and part 4, and one part with clear signals of (past) performance, part 3. The results do not suggest that the participants were concerned with the gender of the

[20] Approximately same p-values when running a linear probability-, logit- and probit model. See appendix for probit models.
[21] Lab experiments using math quizzes and subsequent hiring.

candidates throughout the experiment. Treating "no gender bias" as the null hypothesis, I do not find enough evidence that support discrimination under any information scheme.

When faced with the individual scores of the candidates in part 3, subjects use the candidates' relative performances to pick team members for the subsequent quiz. The candidates who are ranked first – attaining the highest individual total score in quiz 1-5 – are almost exclusively chosen as team members. Furthermore, there is no significant differences in the return to performance.

Facing utter uncertainty in part 4, subjects choose approximately 50 % women and 50 % men as "best" and equivalently when picking the "worst". Thus, subjects do not seem to pick according to gender. Rather it fits into a randomization pattern. Furthermore, in part 2, there is no compelling evidence that the gender of the candidates is important to the decision maker. The influence of gender is insignificant between and across treatment. Rather, subjects' decisions are driven by a candidates' total score. In addition, the subjects in the first-author treatment navigate according to the number of times a candidate is listed first. Consequently, in our sample subjects seem to be consistent in their way of assessing the candidates according to performance variables.

Combined, it would be hard reaching the conclusion that gender bias (or discrimination) is present in our lab. Subjects use the information we provide to assess candidates. When information is scarce, they do not seem to substitute score with gender as a signal of ability. In other word, equally (mathematically) qualified male and female candidates are equally likely to be chosen.

As noted above, the number of observations is quite low and there is a substantial lack of observations in the alphabetical treatment. While there were 38 subjects in the first-author treatment, there were only 18 subjects in the alphabetical treatment. Thus, there may well be a treatment effect or gender biases, but the number of observations limits the models' opportunity to detect it. Ergo, evaluating the results with some skepticism is of paramount importance. None of the findings suggests rejection of the null hypothesis I formulated prior to the results, but larger samples might yield alternative findings. As an example, consider the findings on data from part 4. To detect a difference from 0.5 on a 5% significance level with a power[22] of 0.8 – by using a binomial test on 44 observations – we would need to find a

[22] The probability that a false null hypothesis is rejected.

fraction of females of roughly 0.5 ± 0.3 or larger. Qualitatively this is a large difference and it goes to show the need of substantial effect sizes in small samples.

6.0 So why are our results different from those of Sarsons?

Apart from the sample size issues, I believe there are several reasons as to 1) why we do not find a coauthor penalty in our experiment and 2) why the results differ from Sarsons' (2017) findings.

To address the first issue, we should consider the usual concerns when evaluating decisions in the lab. The difference in severity of choosing team members and promoting an individual to a tenure position might matter. As List and Levitt (2007) suggest, the stakes in an experimental setting versus stakes in real life, naturally, could influence the comparisons between them. While subject in our lab make decisions leading to monetary gains around 100 NOK, hiring committees make decisions that potentially influence the reputation of the institution. Furthermore, the amount of time spent on the hiring decision in the lab is less than time spent promoting an economist. Thus, if the participants had more time, they might have assessed the candidates differently. Thirdly, the "Hawthorn effect" cannot be ruled out: There is a possibility that subjects behave differently when monitored.

Another concern is that mathematics might not be a subject that is related to perceptions of gender differences. This would help to explain the lack of differences in outcome in our experiment. In general, findings do suggest that mathematics is perceived as a male domain among students while female abilities are often connected to humanities, social science and music, across countries (Plante, Théoret and Favreau, 2009; see also e.g.: Eccles, Wigfield, Harold and Blumenfeld, 1993; Stake, 1992; Nosek et al., 2002). In fact, Nosek et al. (2008) found that 70% of approximately 1.5 million implicit associations tests from 34 countries, revealed such stereotypes. Norway is one of the countries where the association between men and mathematics is prevalent (Nosek et al., 2002). Consequently, mathematics should be a fitting subject to use. Still, these tests have not been widely used in Norway[23] and the ones that have been conducted are not particularly recent.

A third issue is that the nicknames of the participants might signal other traits than gender. For instance, Bertrand, Duflo and Mullainathan's (2004) well known correspondence study on the employability of white and non-white Americans, has been criticized for using names that not only signal race (Guryan and Charles, 2013). Such a concern is always relevant when

[23] 1502 Norwegian high school students.

signaling certain characteristics by name. On the other hand, a committee assessing candidates for promotion will most probably also observe this through other channels.

Fourth, group dynamics might lead to other outcomes than individual decisions (Yetton and Bottger, 1982). This is a particularly important concern. In a hiring or promotion committee, the gender balance could influence the outcome, as noted above. Moreover, discriminating behavior may be correlated with taking charge or other relevant personality traits. Thus, the view of some individuals can come to represent the group a whole.

Fifth, aside from laboratory issues there might actually be transatlantic differences. A highly comparable study in that regard is that of Reuben et al. (2014) on US undergraduate students. They explore why substantially fewer women than men are found in Science, Technology, Engineering and Mathematics (STEM) - related professions by using mathematical exercises and subsequent hiring. The results reveal that both men and women discriminate against females under all information schemes. Even knowledge of past performance did not wipe out such a bias against females. These results differ from ours and suggest that there might actually be differences in how Norwegian and US subjects assess females' mathematical abilities in the lab. If one allows for extrapolation to real life hiring, gender bias in math-intensive disciplines in the US, on average, may well supersede biases in Norway.

Lastly, suppose the experimental design and choice of mathematics are well suited in the first place. Then there might be some fundamental differences between our sample of students and the faculty members at economics departments. While academic economists, in general, may perceive female economists as inferior to males, on average, recent high school graduates might have a quite different view on gender differences in school related abilities. Maybe they do not recognize that there are differences at all. Furthermore, in line with such an explanation, the environments in which economists and Norwegian University students are located, are quite different. While men constitute the largest share in economics departments, women represent approximately 60 per cent of Norwegian students in higher education (Statistics Norway, 2017). Consequently, the perceptions might differ.

Hence, there may be some natural reasons to why we do not find a treatment effect. Still, it is important to be aware of the lack of observations.

7.0 Conclusion

Primarily, we set out to investigate two questions:

Are females in an economic laboratory less credited for joint work when signals of their efforts are unclear or absent?

Is the explanatory power of gender in a hiring situations substantially different when collaborators are listed per contribution as opposed to alphabetically?

Sarsons (2017) raises these questions in her exploration of recognition for group work among male and female economists in the US. She finds that female economists are less credited for coauthoring than males are when evaluated for promotion to tenure. Furthermore, she compares these results with evidence from sociology, where she finds that male and female sociologists are equally credited for joint work. Sarsons (2017) point out that this difference may be due to the salient conventions of listing coauthors in these disciplines. Primarily, economists are listed alphabetically, while sociologists are listed per contribution. As the alphabetical order is independent of contribution, Sarsons (2017) suggests it might disfavor female economists. That is, employers give males the ”benefit of the doubt”.

As noted, the lack of data in our experiment is a concern. The data available suggest, however, that subjects consistently use performance variables when assessing candidates. When such variables are absent, the subjects still make decisions independent of gender.

In part 2 of the experiment, I find that gender is not a significant explanation of choice of team member. Rather, subjects choose according to the candidates' total score. Furthermore, in the first-author treatment subjects use the number of times a candidate is listed first as proxy of mathematical ability. In part 4, subjects pick the “best” and “worst” candidates as if randomizing. That is, there is no significant difference in gender balance between the pool of all candidates and the candidates picked in each category. Thus, females do not seem to be less credited when collaborating with another individual. Nor are they treated differently from males when information is absent.

Moreover, I find no evidence suggesting that females are less likely to be chosen in the alphabetical treatment relative to the first-author treatment, controlling for available signals on performance. Thus, the potential effect of alphabetization suggested by Sarsons' (2017) is not detectable in our lab.

In terms of the association between statistical discrimination and coauthor penalty, I argue that these concepts are related. To a certain degree, Phelps' (1972) model of statistical discrimination can explain the latter. However, the assumption that a single-authored paper is a perfectly clear signal of research ability is debatable. Furthermore, as pointed out, methodological and conceptual issues need to be addressed in order to justify stronger empirical claims concerning this link. In our experiment the results does not comply with the implications of these two notions.

It is important to note that our main results might suffer from the loss of the first alphabetical session. Fewer observations make identification harder. Consequently, if effects were small it would be somehow harder to identify such an effect in the analyses I have performed. Even if we had retrieved data from the first alphabetical session, we would still have few observations in this pilot.

Future experimental research should address the obvious question of differences between countries. That is, even though I do not find evidence of coauthor penalty, this might be a country or region specific effect. New designs could also incorporate the group dynamism in hiring committees. A different novelty would be to design an experiment that observes individuals comparably through multiple periods.

References

Aigner, D. J., & Cain, G. G. (1977): Statistical Theories of Discrimination in Labor Markets, *Industrial and Labor Relations Review*, *30*(2), 175–187

Altonji, J. G., & Pierret, C. R. (2001): Employer Learning and Statistical Discrimination, *The Quarterly Journal of Economics 116*(1), 313–350

American Psychological Association (2010): *Publication manual of the American Psychological Association* (6th ed). Washington, DC: American Psychological Association.

Arrow, K. (1973): The Theory of Discrimination, In O. Ashenfelter & Rees (Eds.), *Discrimination in labor markets* (pp. 3–33). Princeton, NJ: Princeton University Press.

Baert, S., & De Pauw, A.-S. (2014), Is ethnic discrimination due to distaste or statistics? *Economics Letters 125*(2), 270–273

Bagues, M. F., & Esteve-Volart, B. (2010): Can Gender Parity Break the Glass Ceiling? Evidence from a Repeated Randomized Experiment, *Review of Economic Studies 77*(4), 1301–1328

Bagues, M., Sylos-Labini, M., & Zinovyeva, N. (2017): Does the Gender Composition of Scientific Committees Matter? *American Economic Review 107*(4), 1207–1238

Becker, G. S. (1995): *The economics of discrimination* (2. ed., 6), Chicago: Univ. of Chicago Press.

Bertrand, M., Chugh, D., Mullainathan, S. (2005): Implicit Discrimination, *American Economic Review* 95(2), 94-98

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

Cain, G. G. (1986): The economic analysis of labor market discrimination: A survey, In *Handbook of Labor Economics*. Amsterdam: North Holland.

Castillo, M., & Petrie, R. (2010): Discrimination in the lab: Does information trump appearance? *Games and Economic Behavior 68*(1), 50–59

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

Charles, K. K., & Guryan, J. (2011): Studying Discrimination: Fundamental Challenges and Recent Progress, *Annual Review of Economics 3*(1), 479–511

De Paola, M., & Scoppa, V. (2015): Gender Discrimination and Evaluators' Gender: Evidence from Italian Academia, *Economica 82*(325), 162–188

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

Einav, L., & Yariv, L. (2006): What's in a Surname? The Effects of Surname Initials on Academic Success, *Journal of Economic Perspectives 20*(1), 175–188

Fang, H., & Moro, A. (2011): Theories of Statistical Discrimination and Affirmative Action: A Survey, In J. Benhabib, M. Jackson, & A. Bisin (Eds.), *Handbook of Social Economics*. Amsterdam: North Holland

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

Ginther, D. K. & Kahn, S. (2004): Women in Economics: Moving Up or Falling Off the Academic Career Ladder? *Journal of Economic Perspectives, 18*(3), 193–214

Goastellec, G., & Pekari, N. (2013): Gender Differences and Inequalities in Academia: Findings in Europe. In U. Teichler & E. A. Höhle (Eds.), *The Work Situation of the Academic Profession in Europe: Findings of a Survey in Twelve Countries*. Dordrecht: Springer Netherlands

Grønmo, L. S., Hole, A., & Onstad, T. (2016): *Ett skritt fram og ett tilbake: TIMSS Advanced 2015. Matematikk og fysikk i videregående skole*, Oslo: Cappelen Damm Akademisk.

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

Hovdhaugen, E., Kyvik, S., & Olsen, T. B. (2004): "Kvinner og Menn – like muligheter? om kvinners og menns karriereveier I akademia" (In Norwegian. Women and men – equal opportunities? On women's and men's career paths in academia), NIFU skriftserie, No. 2004:25

Kaas, L., & Manger, C. (2012): Ethnic Discrimination in Germany's Labour Market: A Field Experiment, *German Economic Review 13*(1), 1–20

Knights, D., & Richards, W. (2003): Sex Discrimination in UK Academia, *Gender, Work and Organization 10*(2), 213–238

Knowles, J., Persico, N., & Todd, P. (2001): Racial Bias in Motor Vehicle Searches: Theory and Evidence, *Journal of Political Economy 109*(1), 203–229

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

Lynch, K. (2006): Neo-Liberalism and Marketisation: The Implications for Higher Education, *European Educational Research Journal* 5(1), 1-17

McElroy, M. (2016): Committee on the Status of Women in the Economics Profession (CSWEP), *American Economic Review 106*(5), 750–773

Mobius, M. M., & Rosenblat, T. S. (2006): Why Beauty Matters, *American Economic Review 96*(1), 222–235

Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2012): Science faculty's subtle gender biases favor male students, *Proceedings of the National Academy of Sciences 109*(41), 16474–16479

Neumark, D. (2016): "Experimental Research on Labor Market Discrimination", NBER working paper, No. 22022.

Niederle, M., & Vesterlund, L. (2007): Do Women Shy Away From Competition? Do Men Compete Too Much? *The Quarterly Journal of Economics*, *122*(3), 1067–1101

Nordic Names (2017): https://www.nordicnames.de/wiki/Main_Page, accessed 24/4-2017

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

Nunley, J. M., Pugh, A., Romero, N., & Seals, R. A. (2014): "An Examination of Racial Discrimination in the Labor Market for Recent College Graduates: Estimates from the Field", Auburn University Department of Economics Working Paper Series, No. AUWP2014-06

Pager, D., & Shepherd, H. (2008): The Sociology of Discrimination: Racial Discrimination in Employment, Housing, Credit, and Consumer Markets, *Annual Review of Sociology 34*(1), 181–209

Phelps, E. S. (1972): The Statistical Theory of Racism and Sexism, *American Economic Review 62*(4), 659–661

Plante, I., Théoret, M., & Favreau, O. E. (2009): Student gender stereotypes: contrasting the perceived maleness and femaleness of mathematics and language, *Educational Psychology* 29(4), 385-405

Reimers, C. W. (1983): Labor Market Discrimination Against Hispanic and Black Men, *The Review of Economics and Statistics 65*(4), 570-579

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

Riach, P. A., & Rich, J. (2010): An Experimental Investigation of Age Discrimination in the English Labor Market, *Annals of Economics and Statistics* (99/100), 169-185

Sarsons, H. (2017): *Recognition for Group Work*, Working paper. Retrieved from http://scholar.harvard.edu/sarsons/publications/note-gender-differences-recognition-group-work

Stake, J. E. (1992): Gender Differences and Similarities in Self-Concept Within Everyday Life Contexts, *Psychology of Women Quarterly 16*(3), 349–363

Statistics Norway (2017): "Student in higher education. Retrived from https://www.ssb.no/en/utuvh, accessed 26/4-2017

Vabø, A., Gunnes, H., Tømte, C., Bergene, A. C., & Egeland, C. (2012): "Kvinner og menns karriereløp i norsk forskning - En tilstandsrapport", NIFU skriftserie, No. 2012:9

Van Praag, C. M., & Van Praag, B. M. S. (2008): The Benefits of Being Economics Professor A (rather than Z), *Economica 75*(300), 782–796

Yetton, P. C., Bottger, P. W. (1982): Individual versus group problem solving: An empirical test of a best-member strategy, *Organizational Behavior and Human Performance* 29(3), 307-321

Zinovyeva, N., & Bagues, M. (2011): "Does gender matter for academic promotion? Evidence from a randomized natural experiment", IZA Discussion Paper Series, No. 5537

Appendix

A1 Sex and choice in part 4

Model 1, in table A2, shows that there is no statistically significant relation between the sex of the subject and the gender of the candidate picked as having performed worst in part 3. Nor is there such a statistically significant association as subjects choose whom they believed performed best in part 3, as shown in model 2. *Female* reflects gender of the chosen candidate: one if female and zero if male. *Fem Sub* is the sex of the subject (choosing): one if female and zero if male.

Table A1: (Fe)males picking (fe)males

Dep Var: Female	(1) b/p	(2) b/p
Fem Sub	-0.417	-0.120
	(0.296)	(0.762)
Observations	43	43

Notes: Probit regressions. The dependent variable is gender of chosen candidate. Model 1 is category Worst and model 2 is category Best. The top panel reports coefficients, p-values in parenthesis. Standard errors are robust. The lower panel indicates the number of observations.

A2 Instructions in Norwegian (Original)

Velkommen til dette eksperimentet. Resultatene vil bli brukt i et forskningsprosjekt. Det er derfor viktig at du følger visse regler. Du skal ikke snakke eller på annen måte kommunisere med andre deltakere mens eksperimentet pågår. Mobiltelefoner skal være avslått eller satt på lydløs og lagt bort. Det er ikke tillatt å bruke annet enn anvist programvare på datamaskinen.

Det vil være full anonymitet i eksperimentet. Ingen av de andre deltagerne i rommet får vite hvilke avgjørelser akkurat du tar. Det vil heller ikke være mulig for noen andre å knytte avgjørelsene som tas underveis i eksperimentet tilbake til enkeltpersoner. Du vil få beskjed når eksperimentet begynner, og når du kan begynne å taste inn dine svar på maskinen foran deg. Hvis du har spørsmål underveis i eksperimentet, rekk opp hånden, så vil en av oss komme bort og svare deg.

I kompensasjon for din deltagelse vil du motta penger. Hvor mye penger du får utbetalt, kommer an på de valgene du og andre tar underveis.

I løpet av eksperimentet vil du flere ganger se en knapp på skjermen foran deg der det står «OK». Det er viktig at du klikker på denne når du er klar til å gå videre. Hvis du glemmer det, vil alle bli sittende og vente på deg.

Før vi begynner

Eksperimentet består av fire deler. Hva du gjør i én del, påvirker ikke hvor mye du kan tjene i de neste delene.

I løpet av eksperimentet kommer du til å få anonymisert informasjon om andre deltakeres resultater. Dette blir enklere dersom alle har et kallenavn.

Du vil straks få opp et spørsmål på skjermen. Svaret ditt bestemmer kallenavnet ditt i eksperimentet.

Spørsmålet lyder: «Tenk deg at du skulle hatt et annen fornavn enn det du faktisk har. Hvilket fornavn kunne du da ha tenkt deg å velge?» Svaret blir ditt kallenavn i eksperimentet. Vi ber deg velge et relativt vanlig fornavn på 3 – 8 bokstaver. For å hindre at flere får nøyaktig samme kallenavn, ber vi deg om å føye til en hvilken som helst stor bokstav etter navnet (for eksempel: «Anne K»). **Av tekniske grunner er det viktig at fornavnet skrives med STOR FORBOKSTAV.** Vi ber om at du etter eksperimentets slutt ikke lar andre deltakere få vite hvilket kallenavn du brukte.

Del 1

I del 1 av eksperimentet skal du og en annen deltaker, som vi her vil kalle partneren din, jobbe i par. Parene trekkes tilfeldig av dataprogrammet.

Når du er klar til å starte, trykker du på knappen der det står «OK». Du vil da bli presentert for en serie med enkle matteoppgaver. Du har inntil 55 sekunder på deg til å forsøke å løse flest mulig av disse. Det samme gjelder partneren din. Vi sier fra når 55 sekunder er gått. Du har da ytterligere fem sekunder til å trykke «OK» for å lagre svarene dine. Merk at du **må** trykke på «OK» før tiden er ute – hvis ikke, blir ikke svarene dine registrert. Når du har trykket «OK», får du ikke løst flere oppgaver.

Både du og partneren din får 1 kr for hvert riktige svar paret har avgitt til sammen. Dette gjelder uavhengig av hvem som avga svarene, og uavhengig av om dere svarte riktig på de samme spørsmålene eller ikke.

Hvis du for eksempel har 10 riktige svar, og partneren din har 8 riktige svar, får dere 18 kroner hver i den runden.

Du får ikke vite partnerens kallenavn, og vil heller ikke kunne se svarene til partneren din.

Alt dette vil så bli gjentatt fire ganger til. For hver runde vil dataprogrammet trekke en ny partner til deg tilfeldig. Del 1 av eksperimentet har altså til sammen fem runder, der du vil ha forskjellig partner i hver runde.

Rekk opp hånden hvis du har spørsmål. Eksperimentet begynner når alle har trykket «OK, jeg er klar til å starte».

Del 2

Denne delen likner Del 1, men det er bare én runde, og reglene for betaling er litt annerledes.

Som før vil du få en serie enkle matteoppgaver opp på skjermen. Du skal løse flest mulig av disse i løpet av 55 sekunder. Du får deretter fem sekunder på deg til å trykke «OK» og slik registrere svarene dine. For hvert riktige svar du selv har, vil du få 1 kr.

I tillegg skal du nå selv velge et lag bestående av to andre deltakere. Disse skal jobbe og tjene penger for deg. For hvert riktig svar disse to har til sammen, får du 3 kr. Dette kommer i tillegg til pengene du tjener på å svare riktig selv.

Det første du skal gjøre i Del 2, er å velge dette laget.

På skjermen vil du få opp en tabell med fire kallenavn i øverste linje. Hver av dem har fått et kandidatnummer. Disse fire er kandidatene du kan velge mellom.

Kandidatene har vært med i fem forskjellige par i Del 1, akkurat som deg. Tabellen gir en oversikt over hvem som har vært kandidatens partner i de ulike rundene av Del 1, og hvor mange riktige svar hvert par fikk til sammen i runden.

[Bare Treatment A:] For hvert par er navnene ordnet alfabetisk.

[Bare Treatment B:] For hvert par er navnene ordnet etter poengsum, slik at den av de to som fikk høyest poengsum er nevnt først. (Hvis begge har like mange poeng, trekker datamaskinen rekkefølgen tilfeldig.)

Tabellen vil se omtrent slik ut:

(Figure A1: Choose two candidates)

Velg to av kandidatene fra den øverste linjen i tabellen, ved å huke av for paret med kandidatnumrene du ønsker (til høyre i skjermbildet). Dette paret blir ditt valgte lag.

Når du har valgt lag og er klar til å gå videre, trykker du på knappen der det står «OK». Du vil da få en serie med enkle matteoppgaver, og har inntil 55 sekunder til å løse flest mulig av disse. Du får deretter fem sekunder på deg til å trykke «OK» og slik lagre svarene dine.

Som nevnt får du 1 kr for hvert riktig svar du har selv, og 3 kr for hvert riktig svar laget ditt har (antall riktige svar de to kandidatene du valgte har til sammen).

Hvis du for eksempel har 10 riktige svar, mens de to kandidatene du valgte har henholdsvis 5 og 20 riktige svar, får du 85 kr ($10 + 3 \cdot (25) = 85$).

For å være sikker på at vi har forklart dette godt nok, vil vi nå be deg svare på noen spørsmål på skjermen. Rekk opp hånden hvis du har spørsmål. Del 2 av eksperimentet vil starte når alle er ferdige med spørsmålene og har trykket «OK».

Del 3

Del 3 likner Del 2. Som før vil du få en serie enkle matteoppgaver opp på skjermen, og skal løse flest mulig av disse i løpet av 60 sekunder. For hvert riktige svar du selv har, får du 1 kr.

I tillegg skal du nå selv velge én annen deltaker som skal jobbe og tjene penger for deg. For hvert riktig svar denne personen har, får du 3 kr. Dette kommer i tillegg til pengene du tjener på å svare riktig selv.

Det første du skal gjøre i Del 3, er å velge denne personen. Du vil få opp en tabell med kallenavnet til fire kandidater som du kan velge mellom, og deres antall riktige svar i hver av rundene i Del 1. Du skal velge en av disse kandidatene.

Når du har valgt din kandidat og er klar til å gå videre, trykker du på «OK»-knappen. Du vil da bli presentert for en serie med enkle matteoppgaver, og har inntil 55 sekunder til å løse flest mulig av disse. Vi sier fra når de 55 sekundene er gått. Du får deretter fem sekunder på deg til å trykke «OK» og slik lagre svarene dine.

Som nevnt får du 1 kr for hvert riktige svar du har selv, og 3 kr for hvert riktige svar din valgte kandidat har.

Hvis du for eksempel har 10 riktige svar, mens kandidaten du valgte har 20 riktige svar, får du 70 kr ($10 + 3 \cdot 20$).

Trykk «OK» når du er klar til å starte.

Del 4

I denne delen skal du bare svare på noen spørsmål.

På skjermen vil du få opp kallenavnene til tre andre deltakere. Din oppgave er å gjette hvem av dem som hadde flest riktige svar i Del 3, og hvem av dem som hadde færrest riktige svar i Del 3.

Du får 10 kroner for hvert riktig svar. (Hvis noen av de tre hadde helt likt antall riktige svar, vil det ikke spille noen rolle for utbetalingen din hvilken rekkefølge du har plassert disse i.)

Deretter vil du få noen enkle tilleggsspørsmål om deg selv.

A3 Instructions in English

[Translated from Norwegian]

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

The experiment is anonymous. None of the other participants will know what decisions you have taken. Nor will it be possible for any other individual to link decision to any single participant. You will be told when the experiment starts, and when you may 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 gain 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 commence

The experiment consists of four parts. What you do in one part will not affect how much you could gain in the next part.

During the experiment, you will be provided with anonymous information about other participants' results. This will be easier if you have a nickname.

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 would 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 participants 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 let other participants know you nickname after the experiment has finished.

Part 1

In part 1 of the experiment you and another participant, which 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 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 kr for each correct answer you give. This applies independently of whom answers correctly, and independently of whether you answer the same answer correctly or not.

For example, if you provide 10 correct answers and your partner provide 8 correct answer, you get 18 kr 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 randomly. Thus, part 1 has five rounds and you will have a new partner each round.

Raise your hand if you have a question. The experiment start 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 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 kr.

In addition , you will choose a team consisting of two other participants. These will work and earn money for you. For each correct answer they provide, you get 3 kr. This comes in addition 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 A2: Choose two candidates)

Choose two candidates from the upper row in the table, by ticking off the pair with candidate number 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. Then you have 5 seconds to push "OK" and thereby saving your answers.

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

For example, if you have 10 correct answers and the two candidates have 5 and 20 correct answers, respectively, you earn 85 kr (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 kr.

In addition you will now pick one other participant who will work and earn money for you. For each correct answer provided by this person, you get 3 kr. 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 to 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 kr for each correct answer provided by yourself and 3 kr 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 kr (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 participants will appear on the screen. Your task is to guess who answered the most correct answers in part 3, and who provided the fewest correct answers.

You get 10 Kroner 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)

Thereafter you will get some additional questions about yourself.

A4 Screenshot

Figure A3: Choosing one candidate in part 3