Pre-Analysis Plan
Gender Discrimination Elicited in the General Population of the United States

Ingvild Almås Serena Cocciolo Jonathan de Quidt Sebastian Fest
Anna Sandberg

June 26, 2020

Contents

This plan was written and posted prior to receiving any outcome data from Gallup.

1 Introduction

We investigate gender discrimination against women in the hiring processes in the labor market. In particular, we focus on two aspects.

First, we ask whether discriminating behavior decreases when employers have information about past performance from individual work rather than from team work, i.e., is discrimination lower when the signal about past work experience is stronger? Furthermore, we examine whether discrimination against women in team work settings depends on the gender composition of teams they work in. We investigate two policies designed to mitigate discrimination in team settings: one ("hidden gender") in which we hide the gender of the candidates, and one ("cognition") in which participants are forced to reflect on the productivity weight they give to the candidates and their co-workers prior to making a choice.

Second, we investigate whether there is an association between economic hardship during the COVID-19 pandemic and discriminating behavior. More specifically, do participants who experienced a negative shock in the labor market discriminate more?

1.1 Previous research and our contribution

Sarsons (2017) and Sarsons et al. (2020) explore whether discrimination increases if past work experience is from team work rather than individual work in the field, finding that female researchers incur a penalty when they coauthor but male researchers do not. This penalty is largest when women coauthor with men. In an MTurk sample and a sample of HR managers they again find evidence that women are penalized for team work. We will investigate whether this effect replicates, and its magnitude, in a controlled experimental setting on a representative sample from the United States; whether we can conclusively identify core mechanisms at play; and whether we can design efficient policies to mitigate these issues.

The second part of our project relating economic hardship to discrimination is more exploratory. It has been shown that racial discrimination may change in recessions and that the behavioral response to economic hardship may depend on internal motivations of the decision maker (Krosch et al., 2017). We will be able to study whether economic hardship during the COVID-19 pandemic outbreak leads to more or less gender discrimination in the labor market. We will also investigate how such discrimination relates to support for the argument that, "in times of high unemployment, men should have priority for jobs" (Hakim, 2018).

2 Experiment

We conduct an experiment structured in three stages, illustrated in Table 1. The Work stage and the Evaluation stage are conducted on Amazon Mechanical Turk (MTurk), and the Employer stage, a separate evaluation stage, and a measurement of beliefs, are conducted in collaboration with the Gallup Panel, a representative sample of the adults in the United States (we discuss sampling procedure and representativeness below).

Table 1: Timeline of the experiment

- **Work stage - Task 1**: Workers work individually and propose a slogan for an advertising campaign for a product.
- **Work stage - Task 2**: Workers work individually or in teams (depending on treatment) and propose a slogan for an advertising campaign for a different product.
- **Evaluation stage**: Evaluators evaluate the quality of the slogans proposed during Task 1 and Task 2 of the Work stage.
- **Payment of workers**: Workers from the Work stage are paid according to the evaluations received by their proposed slogans in the Evaluation stage.
- **Employment stage - Hiring decision**: Employers choose whom to hire from a set of two candidates, observing the slogan produced during **Task 2** of the Work stage.
- **Payment of employers**: Employers are paid according to the evaluations (from the Evaluation stage) of the slogan proposed during **Task 1** by the person they hire.
- **Survey**: Participants reveal whether they personally have experienced any negative labor market shocks the last 3 months.

We randomize (we will be more elaborate on the randomization when we describe the treatment branches below):
(i) whether the hiring decision is made based on a slogan produced individually or in a team (Individual work vs. Team work treatment),
(ii) the gender composition of the teams in the Team work treatment,
(iii) policies (Baseline vs. Hidden gender vs. Cognition treatment), and
(iv) products. In particular, we randomize whether the two products in Task 1 and Task 2 are both stereotypically male, or both stereotypically female.

In the main part of our analysis, we will focus on the stereotypically male products as our focus is on discrimination against women. We will also examine whether discrimination follows stereotypes, i.e. whether we observe discrimination against men when the products are stereotypically female.

2.1 Work stage

In this stage, workers recruited on MTurk are instructed to come up with slogans for two advertisement campaigns. To generate their first slogan (Task 1), all workers work individually. To generate the second slogan (Task 2), workers are assigned to work either individually (the "Individual work" treatment) or in teams of two (the "Team work" treatment, in which workers can communicate through a chat window).

Workers write slogans for one of four different product pairs:

1. *Task 1:* Sports motorcycle,
   Task 2: Drill

2. *Task 1:* Truck,
   Task 2: Package of sports TV channels

3. *Task 1:* Diet shake,
   Task 2: Shampoo

4. *Task 1:* Sewing machine,
   Task 2: Sun screen for kids

The first two product pairs represent stereotypically male products, while the last two represent stereotypically female products. The products were selected based on a pilot study conducted on MTurk in fall 2018.[1]

The tasks are incentivized: In addition to the show-up fee of $2, workers earn a bonus if their slogan receives a high evaluation from a set of independent evaluators. For each product, workers receive a bonus of $1 if their slogan is ranked, on average, in the top third by the evaluators.

Under the "Team work" treatment, the two workers in the same team can communicate through a chat window. The team members are only eligible for the bonus if they agree on one slogan.[2] Thus, while we allow workers to not reach an agreement on the slogan to propose, we provide incentives for the team to produce a joint output.[3]

2.2 Evaluation stage

Evaluators are recruited on MTurk and instructed to evaluate slogans according to their perceived quality. We show evaluators groups of 9 slogans, and we ask them first to rank the slogans and then to assign to each slogan a quality score between 0 and 10.[4] The groups of 9 slogans are formed randomly, grouping together slogans for the same product from both working conditions ("Individual work" or "Team work"). Each evaluator is shown one group of slogans from each of four randomly chosen and randomly ordered products. The evaluations are not incentivized, but we stress that the compensation of the MTurk participants in the Work stage is linked to this evaluation. We reward evaluators with a $2 show-up fee.

When eliciting quality scores, we include in the instructions:

> If you judge a slogan to be unusable (e.g. it is blank, unrelated to the product, or poor English) please give it a score of 0.

and use these "zero scores" to screen out unusable slogans, analogous to a very basic pre-screening of job applications.

2.3 Employment stage

Employers recruited through Gallup are shown information about two candidates (one female and one male) and asked to indicate which candidate they want to hire to come up with a slogan individually. Choices are

[1] In the pilot, the following 13 products were tested: motorcycle, drill, truck, package of sports TV channels, beer, barbecue, diet shake, shampoo, sewing machine, sun screen for kids, laundry detergent, dishwasher detergent, and diapers. We hired 99 "workers" to come up with slogans for 4 products each (in total, 391 slogans were produced), and 51 "evaluators" to evaluate the quality of these slogans (each evaluator was asked to rate 10 different slogans for each product on a scale between 1 and 10). We also hired 100 "employers" who each made a hiring decision for 6 different products. The hiring decision consisted of choosing one candidate to hire to create a slogan for an advertisement campaign for the product. For each product, the employers were presented with a list of six candidates, with information about their gender, age, education and geographic origin. Employers earned a bonus of $3 if the slogan proposed by the candidate they hired received a high ranking from the evaluators. In addition to this incentivized choice, employers were asked to report, for each of the 13 products, their beliefs about the relative ability of men and women to produce high quality slogans. They did this by first ranking all products in terms of whether men or women produce the best slogans, and then indicating their beliefs on a 5-point scale ranging from "Men are much better than women" to "Women are much better than men". We chose the final 8 products in two steps. First, we removed the products where gender differences in average slogan quality were significant with $p<0.15$. Then, we kept the products with the strongest gender stereotypes based on the employers' decisions and ratings. We thus ended up with the following products (in parentheses, we show the share of employers who hired a woman for producing a slogan): sports motorcycle (10.9%), drill (16.7%), pick-up truck (19.1%), package of sports TV channels (19.6%), diet shake (74.5%), sun screen for kids (81.4%), shampoo (81.8%) and sewing machine (85.7 %).

[2] We also paid bonuses to team workers who claimed that they did reach agreement, even if the coworker claimed they did not, but we do not use these slogans in the analysis.

[3] In spring 2019, we tested the team work setting in a pilot study on MTurk. The chat function worked very well; the workers used the chat window to discuss slogans, and in most teams the two team members managed to agree on one joint slogan.

[4] When asking evaluators to assign a quality score to each slogan, we show the slogans in the order based on the evaluator's own ranking.

incentivized: employers receive a $6 bonus if they select the worker who produced the best **Task 1** slogan, based on the average quality score assigned by the MTurk evaluators.

The information shown includes the candidate's name (an assigned pseudonym), education level, "relevant experience", and the slogan they produced for **Task 2** of the Work stage. The gender of the candidate (and, in the "Team work" treatment, their co-worker) is conveyed through pseudonymous names (see below) and an avatar showing a silhouette of a man or a woman. We tell the employers that we do not use workers' real names. Education is displayed in three categories: "High School or less", "Some College education", and "College degree or more". Relevant experience is shown in three categories: "Up to 1 year", "1 to 3 years" and "3 years or more". Worker gender, education, and experience are constructed from self-reported characteristics by the MTurk workers.[5]

Since we do not mention gender explicitly, and, since we show information about education and work experience, we do not believe that gender is too salient for the employers when evaluating candidates.

An overview of our treatment branches is displayed in Figure 1. We randomize:

1. Whether the slogan is produced individually or in a team ("Individual work" versus "Team work" treatment). In the "Team work" treatments, employers are informed that the teams were formed randomly.

2. Policies: (a "Baseline", a "Hidden gender" treatment, and a "Cognition" treatment). In the "Hidden gender" treatment, all information about the gender of the candidates and their co-workers is removed from the resume. In the "Cognition" treatment, participants are asked to predict the relative contribution of each team member prior to selecting a candidate, which we interpret as forcing them to explicitly reflect on the productivity weight they give to the candidate and their co-worker (for example, if they wish to assign a low weight to the candidate, they must assign a high weight to the co-worker).

3. The gender of the co-workers of the candidates in the "Team work" treatment. Co-worker gender is revealed by the co-workers' pseudonymous name. In particular, we vary whether the employer chooses between (i) a male candidate with a female co-worker and a female candidate with a male co-worker ($M_{can}F_{cw}$ vs. $F_{can}M_{cw}$), (ii) a male candidate with a male co-worker and a female candidate with a female co-worker ($M_{can}M_{cw}$ vs. $F_{can}F_{cw}$), (iii) a male candidate with a male co-worker and a female candidate with a male co-worker ($M_{can}M_{cw}$ vs. $F_{can}M_{cw}$), and (iv) a male candidate with a female co-worker and a female candidate with a female co-worker ($M_{can}F_{cw}$ vs. $F_{can}F_{cw}$).

4. Whether the candidates write slogans for stereotypically male or stereotypically female products.

To reduce the total number of treatments, and increase statistical power,

- The "Hidden gender" and "Cognition" treatments are only implemented under the "Team work, male products" treatment branch.

- All four combinations of candidate/co-worker gender are only implemented under the "Team work, baseline, male products" treatment for stereotypically male products. All other team work treatments use only mixed-gender teams.

[5] Relevant experience is constructed from their months of experience working on MTurk, split into three roughly equal-sized categories.

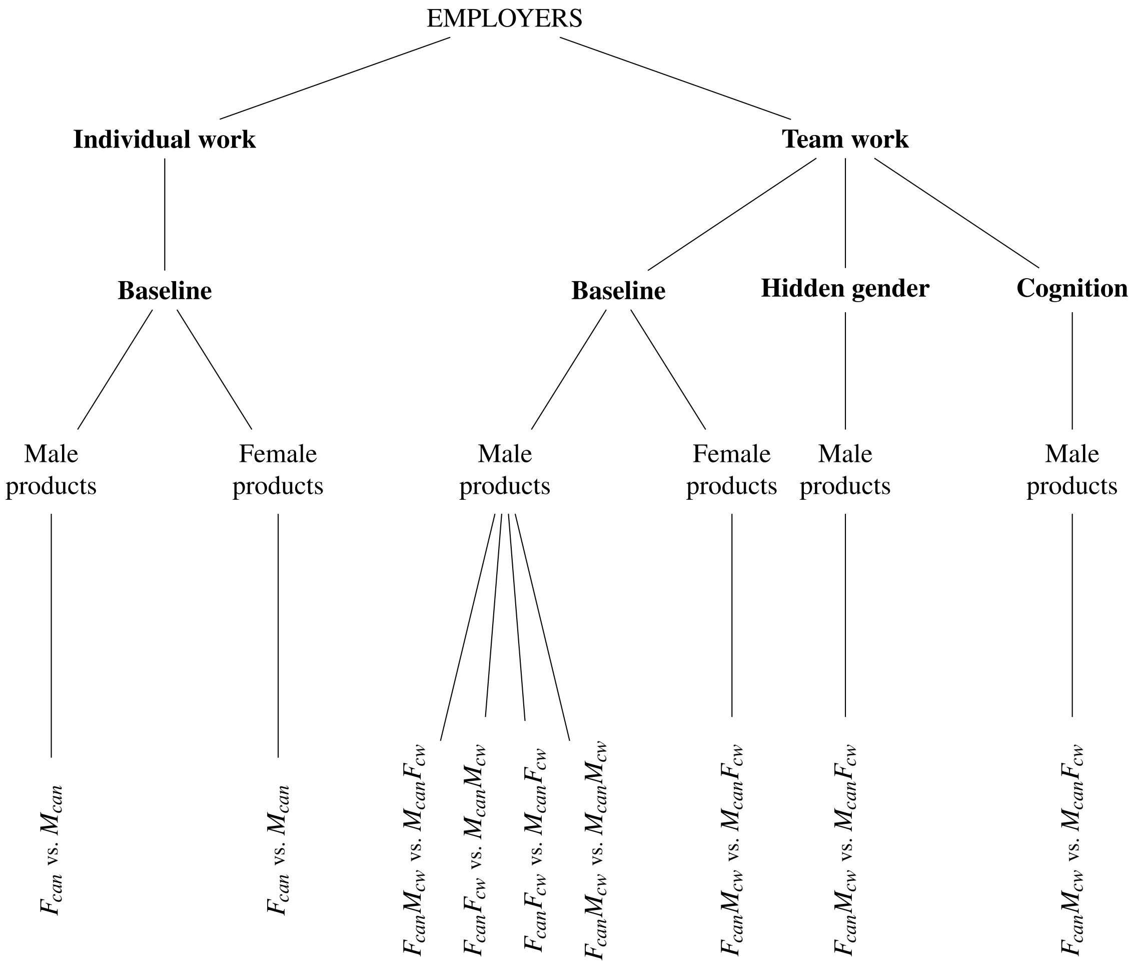
EMPLOYERS
Individual work
Team work
Baseline
Baseline
Hidden gender
Cognition
Male products
Female products
Male products
Female products
Male products
Male products
F_{can} vs. M_{can}
F_{can} vs. M_{can}
$F_{can}M_{cw}$ vs. $M_{can}F_{cw}$
$F_{can}F_{cw}$ vs. $M_{can}M_{cw}$
$F_{can}F_{cw}$ vs. $M_{can}F_{cw}$
$F_{can}M_{cw}$ vs. $M_{can}M_{cw}$
$F_{can}M_{cw}$ vs. $M_{can}F_{cw}$
$F_{can}M_{cw}$ vs. $M_{can}F_{cw}$
$F_{can}M_{cw}$ vs. $M_{can}F_{cw}$

Figure 1: Treatment Branches

3 MTurk data collection and choice set construction

In this section we first describe how we collected worker and slogan data on MTurk, from which to construct employers' choice sets. Then we describe how we construct the choice sets.

1. Work stage: We recruited a total of 1997 MTurk workers to write slogans in 5 batches on separate dates. The first batch (404 workers) was all individual work, i.e. each worker wrote two slogans by themselves. These workers were randomized into four groups corresponding to the four product pairs (102 for motorcycle–drill, 100 for truck–TV package, 100 for shake-shampoo, 102 for sewing machine–sunscreen. Batches 2–5 were team work, each for a single product pair, aiming for 400 workers per batch. This was done to maximize the number of same-product workers available to match at any point in time. Teams were matched in real time through a chat platform that connected each worker with the next available coworker, for our purposes they are as good as randomly formed. We recruited 394 workers for motorcycle–drill, 399 for truck–TV package, 400 for shake–shampoo, and 400 for sewing machine–sunscreen. At a maximum this would generate around 200 teams per product pair, but because some workers log into the task when there are no teammates available, not all workers successfully match into a team.

2. Evaluation stage: This stage consists of two parts, first we screen the slogan data to select only the slogans that are eligible for evaluation. Then, we form sets of slogans to be evaluated, and recruit MTurk workers to evaluate them.

Our primary screening was to identify workers that could potentially be used in the employment experiment (i.e. could be included in an employer's choice set), and ensure that both of their slogans were evaluated. In addition we submitted for evaluation slogans of workers that could not be used in a choice set, but who nevertheless should be assessed for a bonus (the main category here was workers who did not match with a teammate in Task 2 but could still receive a bonus for their individual work in Task 1).

To be usable, a worker had to:

- Pass a basic attention check at the start of the survey,[6]
- Submit 2 slogans,
- Not participate multiple times, or work with a teammate that participated multiple times,
- Neither worker or teammate reported "Other" gender (since we only consider Male and Female genders for this study),
- Not submit as a slogan a blank response, "slogan," the product name, or two identical slogans,
- If in teamwork: match to a teammate and report that the team agreed on a slogan,
- If in teamwork: team submissions differ by no more than 10 characters (Levenshtein distance).

In total, 1,706/1,997 workers and 2,515 slogans were submitted for at least one evaluation. Of those workers, 1,112 were potentially usable in a choice set based on the above criteria (the main reason for the difference is failure to match with a teammate, since the matching was based on a waiting room protocol and workers might not match if there was not a co-worker available).

After this screening, we randomly group same-product slogans into "evaluation sets" of nine. We mix together individual and team-produced slogans. Each slogan is added to multiple evaluation sets: for Task 1 slogans (that are not shown to the employers) each slogan belongs to 7 different sets, for Task 2 slogans 15 sets (we have fewer of them, because each team produces two Task 1 slogans but only one Task 2 slogan).[7] We end up with 2,651 evaluation sets in total.[8] We recruited 698 evaluators on MTurk and each evaluator evaluated one randomly chosen evaluation set from each of four randomly chosen products. The four randomly chosen products are selected randomly from the eight possible products (Task 1 and Task 2 together) without replacement.

3. Choice set construction: Based on the evaluation data, we construct choice sets that can be shown to employers. First, we eliminate evaluators that failed our attention check, leaving us with 631 evaluators. Next we eliminate some slogans (and thus workers) based on the evaluation data:

- The random sampling of evaluation sets (an unavoidable feature of Qualtrics, the platform through which we conducted the evaluations) means that some slogans are evaluated more frequently than others. To reduce noise, we eliminate slogans that received fewer than 3 evaluations. This leaves us with 1,075 workers.
- Next we eliminate slogans over 100 characters in length, since these are unlikely to display well on the employer interface. This leaves us with 1,047 workers.

[6] Question 1: "If you are above 15 years old, please select "Agree""; Question 2: "Elephants are smaller than ants"; Question 3: Please answer "Agree". Options were "Agree", "Not agree nor disagree", "Disagree". MTurkers have to be over 18 years of age so the correct answer to 1 is "Agree."

[7] In total we have 426 motorcycle, 208 drill, 428 truck, 226 TV package, 433 shake, 198 shampoo, 417 sewing machine, and 179 sunscreen slogans to be evaluated.

[8] In total we have 329 motorcycle, 345 drill, 329 truck, 375 TV package, 336 shake, 330 shampoo, 322 sewing machine, and 285 sunscreen evaluation sets, after dropping "remainder" sets with fewer than 9 slogans.

- Finally we eliminate slogans that received 20% or more "zero quality" evaluations (i.e. the evaluator judged them to be "unusable"). This threshold was set by visual inspection: a large fraction of slogans with 20% or more zero scores were indeed of very low quality.[9]. The goal is to screen out very low-effort submissions. This leaves us with 790 workers.

Next, we "harmonize" the number of workers, so that each product-gender combination has the same number of unique workers.

- For individual work we have 24 unique workers for "female worker, motorcycle–drill" and more than 24 for each other gender-product combination. We randomly drop workers until we have 24 men and 24 women for each product.
- For mixed teams, we first check for "incomplete" teams, where one member has survived our screening and the other has not. We have zero male incomplete teams for sewing machine–sunscreen. So we drop all mixed incomplete teams. Turning to complete teams, we have 24 unique complete teams (i.e. 24 men and 24 women) for sewing machine–sunscreen, and more for the other products. We randomly drop teams until we have 24 of each. Each team can appear as a "FM" or a "MF" combination.
- Next, we eliminate same-gender teams for female stereotypical products, since these will not be used. Then we apply the harmonization process. We end up with 5 incomplete teams per product (i.e. 5 men who are missing their male partner, and 5 women who are missing their female partner), and 9 complete teams per product (where both men or both women appear in the data). In total we have 23 workers per product-gender combination.

Next, for each of our treatment arms excluding the policy treatments, we create all possible pairwise combinations of workers that could form a choice set (in the context of mixed teams, we eliminate choice sets where both candidates come from the same team).

From these we randomly sample the desired number of choice sets, one unique choice set per targeted employer in each treatment, as detailed in section 4.1. We note, however, that Gallup will randomly sample with replacement from these collections of choice sets, so some will be drawn more than once and some not at all.

4. Names and avatars: Employers are told that candidates' names have been changed. We do this as follows.

- We take the 9 most popular male and female names from each of the 1950/60s, 1970/80s, and 1990/2000s, i.e. 27 male and 27 female names. Source `https://www.ssa.gov/OACT/babynames/decades/names1970s.html`
- We eliminate duplicates and two female names judged to be of ambiguous gender (Ashley and Taylor).
- We randomly assign male names to male candidates and coworkers and female names to female candidates and coworkers, ensuring that no name is repeated within a choice set.
- For hidden gender we randomly assign the female candidate to "P" or "Q" and the male to the other, then label teams "candidate P," "coworker P," "candidate Q," "coworker Q."

For avatars, we use three generic "female" silhouettes (1–3) and three generic "male" silhouettes (4–6), to be presented alongside candidate names. These are matched in six possible combinations: 1–4, 1–5, 2–6, 2–4, 3–5, 3–6 and randomly assigned in equal proportion. For hidden gender we have a single neutral avatar used for both candidates.

[9] E.g. "This machine are automatic needle an speed and control.....," "CLEAN AND TRUTHFUL," "DRIVE IS BEST FUTURE."

4 Gallup data collection

4.1 Employers

The Gallup panel is a probability-based panel recruited using random digit-dial (RDD) phone interviews that cover landline and cell phones and address-based sampling methods (ABS). To minimize the variance due to weighting and to account for anticipated nonresponse by demographic group, Gallup statisticians drew a stratified sample of panel members. In the stratified sample, the demographic distribution of the sample matched the population targets for United States adults obtained from the 2017 Current Population Survey.

In the final data, Gallup will provide us with survey weights to adjust for any remaining deviations from representativeness. We will use these weights in all our regressions.[10]

In total we target 4,450 employers in the following proportions:

- 400 per product pair for individual work (1,600 in total)
- 400 per product pair for mixed teams (1,600 in total)
- 125 per male stereotypical product pair for female-male vs male-male teams (250 in total)
- 125 per male stereotypical product pair for female-female vs male-female teams (250 in total)
- 125 per male stereotypical product pair for female-female vs male-male teams (250 in total)
- 125 per male stereotypical product pair to be used in the hidden gender policy treatment (250 in total)
- 125 per male stereotypical product pair to be used in the cognition policy treatment (250 in total)

Gallup invites participants to respond to the survey, and preassigns them to one of these treatment arms. The sampling procedure effectively draws with replacement from our preselected sample of choice sets. In practice, the choice sets are duplicated multiple times and more than 4,450 panel members are invited to respond, each of whom is preassigned a specific choice set. Therefore some choice sets will be sampled more than once and others not at all. Moreover, variation in response rates across treatment arms can mean that the final sample sizes slightly differ across arms.

4.2 Evaluators

Our set of possible choice sets contains 476 unique "Task 1" slogans (142 for motorcycle and truck, 96 for shake and sewing machine) and 344 unique "Task 2" slogans (100 for drill and TV sports package, 72 for shampoo and sunscreen). We form these into evaluation sets of 9 as before and collect evaluations from 162 Gallup respondents. For our main measure of slogan quality we will standardize the MTurk and Gallup quality scores separately, then average them (weighting by the number of evaluations of each type). In a robustness analysis we will compare the distributions of MTurk and Gallup evaluations to see if MTurk and Gallup respondents perceive quality differently.

4.3 Belief elicitation

In the belief elicitation we elicit beliefs about gender differences in contribution to team work and performance in individual work. This survey has 200 respondents.

We elicit:

[10] After fieldwork, sample data are weighted to minimize bias in survey-based estimates. The Gallup panel maintains weights for all members that are based on their selection probabilities. These weights are used as baseweights for weighting the final dataset of completes. Next, post-stratification weights are created to adjust for non-response bias. Targets for post-stratification weighting are generated from the 2017 Current Population Survey (CPS) and are projectable to the total United States' adult population, aged 18+. These weights take into account age, gender, education, race, ethnicity, region and population density.

- Beliefs about the percentage of the words in the team work chat that were written by the man. The respondents answer on a scale from 0 to 100 % and they earn a bonus of $3 if the answer is within +/- 5% of the average count.
- Beliefs about how much:
  - The woman advocated for her slogan.
  - The woman listened to the input and ideas of the other team member.
  - The woman contributed to the quality of the slogan.

  This elicitation was not incentivized.
- Beliefs about quality of individual slogans for:
  - The motorcycle.
  - The truck.
  - The diet shake.
  - The sewing machine.

  More specifically we told them that we selected 100 workers at random, 50 men and 50 women, and we ask them to specify how many of the 50 best slogans were produced by men, and how many of the 10 best slogans were produced by men. They earn a bonus of $10 if all guesses are correct (within +/- 10%).

5 Hypotheses and tests

Our experiment is quite rich and our sample size is large. We can test several interesting and important hypotheses with our data. We will mainly focus on two research questions: i) does gender discrimination in the labor market increase when evidence of past work comes from team rather than individual work? and ii) is gender discrimination strengthened in economic hardship under the COVID-19 pandemic?

We plan to study these two questions separately and write two separate papers. This plan describes the main hypotheses and testing related to the first research question quite rigorously whereas the work on the second question will have a more exploratory character and is consequently described in less detail.

5.1 Definition of discrimination

Because the choice sets from which employers select candidates are constructed from real-world data, the distribution of candidate characteristics is not identical between male and female candidates. Therefore, women might be hired less frequently than men due to differences in observables. Note that there are no candidate characteristics that are observable to the employer and unobservable to us.

Since every choice set consists of one man and one woman, we will compare the female hiring rate to a 50% benchmark.

Our core specification is a regression at the employer level, where the outcome variable is a dummy for whether the employer hired a woman, and the right-hand side is a set of treatment variables and controls. We will estimate our regressions using OLS, i.e. the linear probability model, with robust standard errors since randomization is at the employer level.

Definition 1: unconditional discrimination. We will examine whether women are hired more or less than 50% of the time without controlling for candidate characteristics. Since our outcome variable equals one if the employer hires a woman, and zero if they hire a man, in general this amounts to comparing the regression constant to 0.5.

Definition 2: conditional discrimination. We will examine whether women are hired more or less than 50% of the time, conditional on controls detailed below. We set up the controls in "difference form" such that a zero value means the male candidate and female candidate are identical on this dimension, allowing us to interpret the regression constant as the female hiring rate when the alternative candidate is a man with the same characteristics.

Crucially, the difference form forces the controls to apply symmetrically, i.e. the hiring bonus that a man gets for more education should be equal to the hiring bonus a woman gets. It is equivalent to restricting the regression coefficients on male and female characteristics to be equal in magnitude (and opposite in sign, since female qualities should increase female hiring and male qualities decrease it). Thus, if employers treat men and women with the same characteristics differently, this will be absorbed into our discrimination measure.

Our primary tests will be based on the conditional measure, i.e. we ask whether women are less likely to be hired than equal (on observables) men.

5.2 Controls construction

Table 2 lists the worker characteristics visible to the employer and to us. From these variables we construct our "differenced" controls.

Table 2: Variable overview

Variables visible to the employer:

Variable name	Type	Definition
High school	Binary variable	True if candidate has High school or less education.
Some college	Binary variable	True if candidate has some college education.
Bachelor	Binary variable	True if candidate has a Bachelor's degree or more education.
Experience 0–1	Discrete variable	True if candidate 0–1 years' experience.
Experience 1–3	Discrete variable	True if candidate 1–3 years' experience.
Experience 3+	Discrete variable	True if candidate 3 or more years' experience.
Product	Binary variable	Dummy for each product group.
Name	Binary variable	Dummy for each candidate name.
Avatar	Binary variable	Dummy for each avatar pairing.
Past slogan	String	The text of the slogan written for past work

Variables not displayed to the employer:

Variable name	Type	Definition
Past slogan quality	Continuous variable	The average quality score (z-scored) given by the evaluators who observed the past work slogan.

For our preferred specification, we will use the following set of controls:

1. Difference in z-scored past slogan quality between female and male candidate

2. Interaction between (1) and the teamwork treatment dummy, since the past slogan is differently informative under teamwork.

3. Two variables for the difference in education, each of which takes values in $\{-1, 0, 1\}$ (the omitted category is "Some College"):

   (a) High school difference: Female high school dummy - male high school dummy

   (b) College difference: Female college dummy - male college dummy.

4. Two dummy variables for the difference in relevant experience, each of which takes values in $\{-1,0,1\}$ (the omitted category us "1–3 years"):

   (a) 0–1 year difference: woman "up to 1 year" - man "up to 1 year"

   (b) 3+ years difference: woman "3 years or more" - man "3 years or more"

Thus we will have 6 control variables in total.

5.3 Paper 1: Teamwork

Our main hypothesis is H.1, on discrimination and teamwork, but before we test this hypothesis, we will first establish whether or not there is gender discrimination when employers see past work experience from individual work. In order to do this, we will run the following regression using data from our two main treatment arms (stereotypically male products with either individual work or team work with mixed gender teams):[11]

$$E_i = \beta_0 + \beta_1 TW_i + \gamma X_i + \varepsilon_i, \tag{1}$$

where i indexes employers. E_i is an indicator equal to one if the employer selects the female candidate, TW_i is a dummy for the team work treatment, and X_i is a vector of controls (see below).

We will report three estimations in our main table. We consider the 3rd to be our primary specification, which identifies conditional discrimination. The construction of control measures is explained in section 5.2.[12]

1. Without controls.

2. With controls for differences in candidate education and experience,

3. **[Primary Specification]** With controls for differences in candidate education, experience, our measure of the quality of the past slogan, and this quality measure interacted with TW.[13]

4. With controls for the education, experience, and quality (interacted with teamwork) of the male and female candidate separately. This specification allows us to examine whether the reward that men and women receive for each characteristic is different. The reference categories will be "Some College" and "1–3 years' experience."

In this regression, β_0 is interpreted as the hiring rate for female candidates under individual work. Unconditional discrimination is revealed by $\beta_0 < 0.5$ in specification 1, Conditional discrimination is revealed by $\beta_0 < 0.5$ in specification 3.

Specification 4 will allow us to describe how patterns of discrimination vary by candidate characteristics and ability (measured by their work quality).

[11] On the diagram in Figure 1, these are:

1. Individual work - Baseline - Male products - F_{can} vs. M_{can}
2. Team work - Baseline - Male products - $F_{can}M_{cw}$ vs. $M_{can}F_{cw}$

[12] In a robustness analysis we will also report estimates controlling for a product dummy, candidate names, avatars, and employer characteristics.

[13] We interact the slogan quality measure since this is necessarily differently informative between individual and team work, whereas the experience and education measures are equally informative.

5.3.1 Main hypothesis: more discrimination under team work

The next step is to test our main hypothesis, namely whether discrimination against women is larger or smaller when past work experience is from team work. As the quality signal is stronger for individual work, we hypothesize that discrimination is weaker. More specifically, we pose the following hypothesis:

Hypothesis 1 (H.1) *Gender discrimination in employment is larger when employers observe performance from team work rather than individual work.*

We test this hypothesis on the sample of individual workers and mixed-gender team workers, for male-stereotypical products.

We reject the null that teamwork does not increase discrimination if:

- $\beta_1 < 0$,

Multiple hypothesis testing We consider H1 our main hypothesis. The next sections discuss three nested hypotheses, the first concerning how discrimination varies within the team work setting, depending on team composition. The second concerns policies to address discrimination. We consider these independent sub-hypotheses to the main one. Since we test two policy remedies, we will adjust for multiple testing across those.

5.3.2 Nested Hypothesis 1: Interventions. Hidden gender and Cognition

After establishing whether there is more discrimination in team work settings in our context, we want to consider two different interventions that are i) feasible in many contexts where workers are hired and evaluated, ii) hypothesized to decrease discrimination in a team work setting, iii) possible to test in a controlled way. First, we will test a hiding intervention where gender of workers is not revealed to those making the hiring decision or evaluation ("Hidden gender"). Second, we will test the effect of asking the relevant people to explicitly state their prediction of the contribution of the team-members involved before they make the hiring decision or conduct the evaluation ("Cognition").

We will only proceed with the policy analysis if we find evidence of discrimination in the first stage of the analysis.

In the policy analysis we pose two hypotheses:

Hypothesis 2 (H.2) *Discrimination in team work settings is reduced when gender is hidden.*

Hypothesis 3 (H.3) *Discrimination in team work settings is reduced when the employer has to explicitly consider the contribution of each team member before making decisions.*

We will test how to reduce discrimination in the team work setting, so we restrict this analysis to the team work treatment.

1. Team work - Baseline - Male products - $F_{can}M_{cw}$ vs. $M_{can}F_{cw}$
2. Team work - Hidden Gender - Male products - $F_{can}M_{cw}$ vs. $M_{can}F_{cw}$
3. Team work - Cognition - Male products - $F_{can}M_{cw}$ vs. $M_{can}F_{cw}$

Using the sub-sample above, we estimate the following equation:

$$E_i = \gamma_0 + \gamma_1 H_i + \gamma_2 C_i + \delta X_i + \varepsilon_i, \tag{2}$$

where H_i is an indicator for being in the hidden gender treatment and C_i is an indicator for being in the cognition treatment.

We will again estimate according to specifications 1–4 with specification 3 as the primary one. We reject the null stating that there is no effect of hiding gender and find support for H2 if:

- $\gamma_1 > 0$

We reject the null stating that there is no effect of the cognition treatment and find support for H3 if:

- $\gamma_2 > 0$

We will adjust for multiple testing across these two hypotheses, controlling the False Discovery Rate (FDR) using sharpened q-values (Benjamini et al., 2006; Anderson, 2008).

Power and fixed effects for the policy analysis The sample for this analysis consists of the 800 employers assigned to the main treatment (mixed gender teams and male products, 400 per product pair), the 250 assigned to hidden gender (125 per product pair) and the 250 assigned to cognition (125 per product pair).

As described above, the choice sets used in the policy treatments were selected randomly from the set of choice sets used in the main treatment. However, because of the way Gallup sample (with replacement) from the set of choice sets we provide them, some choice sets will end up being used more frequently or not at all in some treatment arms.

In addition to estimating the regression with and without main controls as described above and similar to the other tests, we could here restrict ourselves to the subset of choice sets that appear in both the main treatment and policy treatments, with or without controls. Or we could exploit the fact that the same choice sets appear in all three treatments and include choice-set fixed effects. This potentially improves our ability to control for slogan quality, at the cost of degrees of freedom and reducing the sample size (especially in the main treatment).

We conducted some power simulations under the assumptions that a) employers condition their choice on slogan quality (we simulated different coefficient magnitudes) and that our measures of slogan quality might include slogan-specific idiosyncratic noise (benchmark is no noise, i.e. our quality control is perfect, up to noise with twice the variance of the slogan quality itself). The more important is quality in the choice, the more power we gain by controlling for it. The more important is slogan-specific idiosyncratic noise, the more power we gain through choice set fixed effects.

Given these assumptions, the minimum detectable effect was smallest in almost all scenarios when we use (full sample + controls), i.e. the sample size and degrees of freedom gains appear to dominate. Therefore we plan to use this as our primary specification, but will also report the fixed effects specification.

5.3.3 Nested hypothesis 2: Team gender composition

Our next hypothesis is related to co-worker gender. Our baseline teamwork treatment has four different combinations of candidate and coworker gender:

1. Team work - Baseline - Male products - $F_{can}M_{cw}$ vs. $M_{can}F_{cw}$
2. Team work - Baseline - Male products - $F_{can}F_{cw}$ vs. $M_{can}M_{cw}$
3. Team work - Baseline - Male products - $F_{can}F_{cw}$ vs. $M_{can}F_{cw}$
4. Team work - Baseline - Male products - $F_{can}M_{cw}$ vs. $M_{can}M_{cw}$

We will restrict the sample to these four groups and report estimates of the following regression:

$$E_i = \beta_0 + \beta_1 FFMM_i + \beta_2 FFMF_i + \beta_3 FMMM_i + \gamma X_i + \varepsilon_i, \tag{3}$$

where the omitted category is team combination 1 (FM vs MF), $FFMM_i$ indicates combination 2 (FF vs MM), $FFMF_i$ indicates combination 3 (FF vs MF) and $FMMM_i$ indicates combination 4 (FM vs MM).

We will estimate the equation under control specifications 1–4 from above, and our main focus will be on controls specification 3.

We focus on one testable hypothesis in this analysis, that conjectures discrimination is larger under mixed-gender teams than same-gender teams. This is based on the prediction that working with a man decreases the likelihood a woman is hired (because she gets less of the credit), and conversely working with a woman increases the likelihood a man is hired (because he gets more of the credit).

We therefore test whether there is less discrimination in gender combination 2 than combination 1:

Hypothesis 4 (H.4) *Discrimination in team work is smaller under same-gender than mixed-gender teams.*

We will reject the null hypothesis that coworker gender does not matter, and find support for H4 if:

- $\beta_1 > 0$

We will conduct this test if we find evidence of discrimination in the main analysis.

The coefficients on $FFMF_i$ and $FMMM_i$ allow us to examine whether any difference we find is driven more by the female's co-worker or the male's.

5.3.4 Nested Hypothesis 3: Female stereotypical products

Throughout, we have focused on the male stereotypical products. We will use the treatment arms with stereotypically female products to examine symmetry/asymmetry, i.e. whether discrimination is consistently against women, or follows gender stereotypes.

Our basic regression specification becomes:

$$E_i = \beta_0 + \beta_1 TW_i + \beta_2 FemaleStereotype_i + \beta_3 TW_i \times FemaleStereotype_i + \gamma X_i + \varepsilon_i, \tag{4}$$

We estimate this regression on individual workers and mixed teams, for both the male and female stereotypical products. $FemaleStereotype_i$ is a dummy for female stereotypical products. We follow controls specifications 1–4 as before.

Now,

- β_0 estimates the (conditional) female hiring rate under *individual* work with *male* stereotypes,
- $\beta_0 + \beta_1$ the (conditional) female hiring rate under *team* work with *male* stereotypes,
- The total effect of teamwork on the female hiring rate under male stereotypes is estimated by β_1
- $\beta_0 + \beta_2$ estimates the (conditional) female hiring rate under *individual* work with *female* stereotypes,
- $\beta_0 + \beta_1 + \beta_2 + \beta_3$ estimates the (conditional) female hiring rate under *team* work with *female* stereotypes.
- Finally, the total effect of teamwork on the female hiring rate under female stereotypes is estimated by $\beta_1 + \beta_3$

We will report all of these estimates, which describe the landscape of discrimination under our male and female stereotypical products.

Our hypothesis of interest here asks whether gender discrimination under teamwork is different when workers write slogans for female stereotypical products. In particular, we will study whether discrimination is symmetric, in the sense that any discrimination against women under the male stereotypical products is mirrored by discrimination against men in the female stereotypical product.

Our main hypothesis in this part of the analysis is:

Hypothesis 5 (H.4) *Female hiring is larger when employers observe performance from team work rather than individual work in the female stereotypical product.*

Before turning to the testing of the hypothesis, we will examine whether discrimination against men under individual work with female stereotypical products mirrors discrimination against women in individual work with male stereotypical products. They mirror one another if $\beta_0 = (1-(\beta_0+\beta_2))$.

Turning to the hypothesis test, we will reject a null hypothesis that female hiring is not larger when employers observe performance from team work rather than individual work when the products are stereotypically female if:

- $\beta_1 + \beta_3 > 0$

There are a range of possible alternative findings

- It could be there is no teamwork bias against women in the male stereotypical products, and no teamwork bias against men in the female stereotypical products
- It could be there is a teamwork bias against women for both male and female stereotypical products. For example, if women are perceived to contribute less in teamwork whatever the gender stereotype of the slogan task.
- Teamwork bias against **men** for female stereotypical products exactly mirrors that against women for male stereotypical products (i.e., it is of the same total magnitude).

5.3.5 Exploratory analyses

Beliefs Our analyses of beliefs about gender differences in team work ability and style will be exploratory, with the purpose of shedding light on mechanisms.

Welfare As part of the exploratory analysis, we will discuss how actual hiring behavior deviates from payoff maximizing behavior, and whether there is any systematic deviation based on experience from team work, gender of the worker, and gender of the employer. We will study this by using the payoff-relevant quality assessments from the evaluators and actual observed behavior and characteristics of workers and employers.

Who is discriminating? We will also investigate whether teamwork-based discrimination is related to employer characteristics, with particular focus on employer gender and exposure to negative labor market shocks.

Aggregate female hiring rate For design reasons most of our treatments involve mixed-gender teams. If teams and choice sets were formed at random, we would expect to see 25% all-female teams, 25% all-male teams, and 50% mixed teams. If these met opponents at random, the opponents would be formed according to the same distribution. Using our smaller treatment arms with alternative team compositions we will estimate the female hiring rate for each possible matching, and report the implied aggregate female hiring rate. Note that when a female faces a female opponent, a female is hired with probability 1. We will do the same assuming random matching of individual work opponents.

Structural model We intend to estimate a structural model with the following basic structure:

- Employers believe that male and female ability are Normally distributed. The standard deviation of ability is allowed to differ by gender. Mean ability differs by gender and stereotype, i.e. men might have higher expected ability for male-stereotypical products.

- Employers believe that individual slogan quality is determined by worker ability plus Normally distributed mean-zero noise, with a fixed standard deviation.
- Employers believe that team slogan quality is determined by a weighted average of team-member abilities, where the weights are gender-specific and do not depend on the product (i.e. teamwork contribution is assumed to be a fixed feature of gender). This means that teamwork is a noisier signal of individual ability, particularly so if one gender is believed to contribute less in teams.
- We assume slogan quality is observable to us, measured by our evaluation data.
- Employers observe workers' past work and form a posterior belief about their ability. The employer tries to choose the worker with the highest expected ability, since this person will in expectation write the best slogan for the product on which employer incentives are based. In other words, we assume no taste-based discrimination. Employers do not need to have correct priors and taste-based discrimination will manifest as if they have more pessimistic priors about one gender.

5.4 Paper 2: Discrimination in a time of a pandemic outbreak characterized by economic and social unrest

In the second part of this research we will study whether any gender discrimination is related to the participants' and/or its relatives' exposure to a negative shock in the labor market during the COVID-19 outbreak.

Our main hypothesis for this part of the project is:

Hypothesis 6 (H.6) *Gender discrimination in employment is more prevalent for people who have experienced economic hardship during the COVID-19 pandemic outbreak.*

To test whether economic hardship can predict the tendency to discriminate, we estimate whether there is any difference in the probability to hire a woman between those that have experienced a negative economic shock during the COVID-19 pandemic outbreak and those that have not. More specifically, we run a regression:

$$E_i = \beta_0 + \beta_1 C_i + \gamma X_i + \varepsilon_i, \tag{5}$$

where C is a measure of whether or not the employer has experienced economic hardship during the COVID-19 outbreak. We will test hypothesis 6 by analyzing the data at the employer level. We will restrict the sample to the male stereotypical products excluding the policy treatments, pooling individual and teamwork together.

Our main specification will be a robust OLS regression that controls for dummies for each treatment included in the pooled sample, for each candidate's education and experience, and for the quality of the individual slogan or team slogan, respectively, where team quality is interacted with the teamwork dummy as in the first paper.

We will use the responses to the following survey questions to build an index of economic hardship experienced during the COVID-19 outbreak:

- Do you personally know someone who has been laid off or furloughed in the past 3 months? Select all that apply.

  1. Yes, I have
  2. Yes, a close friend or family member
  3. Yes, an acquaintance
  4. No

- Do you personally know someone who has had their pay reduced in the past 3 months? Select all that apply.

    1. Yes, I have
    2. Yes, a close friend or family member
    3. Yes, an acquaintance
    4. No

- (Only applies to the subsample who have an employer:) To your knowledge, in the past month has your employer done any of the following? Please select all that apply:

    1. Cut jobs
    2. Reduced hours or shifts available
    3. Frozen hiring

We reject the null hypothesis H6 that discrimination is not different for those who experienced hardship if:

- $\beta_1 \neq 0$

Finally, we will relate the findings on discrimination to the survey question on whether we should prioritize jobs for men in times with high unemployment:

- Please share your level of agreement with the following:
  In times of high unemployment, men should have priority for jobs.

    1. Agree
    2. Agree to some extent
    3. Disagree to some extent
    4. Disagree

We will also look at heterogeneity. Particularly we will investigate whether different responses to the question of whether men should have priority for jobs in times of high unemployment, is associated with different responses to economic hardship. We will similarly examine heterogeneity between male and female employers.

References

Anderson, M. L. (2008, Dec). Multiple inference and gender differences in the effects of early intervention: A reevaluation of the abecedarian, perry preschool, and early training projects. *Journal of the American Statistical Association 103*(484), 1481–1495.

Benjamini, Y., A. M. Krieger, and D. Yekutieli (2006, Sep). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika 93*(3), 491–507.

Hakim, C. (2018). *Models of the Family in Modern Societies: Ideals and Realities: Ideals and Realities*. Routledge.

Krosch, A. R., T. R. Tyler, and D. M. Amodio (2017). Race and recession: Effects of economic scarcity on racial discrimination. *Journal of personality and social psychology 113*(6), 892.

Sarsons, H. (2017, May). Recognition for group work: Gender differences in academia. *American Economic Review 107*(5), 141–145.

Sarsons, H., K. Gerxhani, E. Reuben, and A. Schram (2020). Gender differences in recognition for group work. *Journal of Political Economy, forthcoming*.