Essays in Behavioral and Experimental Economics

by

Neeraja Gupta

Bachelor of Arts (honors), Shri Ram College of Commerce, 2012

Master of Arts, Delhi School of Economics, 2014

Master of Arts, University of Pittsburgh, 2018

Submitted to the Graduate Faculty of

Dietrich School of Arts and Sciences

in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2023

UNIVERSITY OF PITTSBURGH

DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Neeraja Gupta

It was defended on

March 22, 2023

and approved by

Dr. Lise Vesterlund, Department of Economics

Dr. Richard Van Weelden, Department of Economics

Dr. David Huffman, Department of Economics

Dr. Alistair Wilson, Department of Economics

Dr. Alex Imas, University of Chicago

Essays in Behavioral and Experimental Economics

Neeraja Gupta, PhD

University of Pittsburgh, 2023

This dissertation consists of three essays on behavioral and experimental economics. Chapter 1 examines if a temporary affirmative action policy can improve representation of women beyond the immediate scope of the policy in settings where employers hold biased beliefs about performance of women. I experimentally elicit employer beliefs and hiring choices for worker performance in two experimental treatments: a control with no restriction on hiring and a temporary affirmative action for women. I find that while hiring choices and beliefs are biased against women in the control treatment, temporary affirmative action treatment leads to improvement in representation of women even after the policy is lifted. Further, employers who are most likely to discriminate against women show the fastest reduction in gender bias in beliefs which in turn help explain their hiring choices. Chapter 2 presents a comprehensive review of 317 papers in the experimental economics literature studying gender differences in economic behavior to assess the empirical validity of the assertion than women are more sensitive to changes in experimental conditions. We find that there does not exist a discernible pattern with respect to whether men or women drive gender differences in responsiveness. We further find that the female-sensitivity assertion gets selective positive reinforcement in the literature which many in turn lead to over generalization of this claim. Chapter 3 presents work from a study where we compare five populations commonly used in experiments in economics and other social sciences: undergraduate students at a physical location (lab), and virtually over Zoom (V-lab), Amazon's Mechanical Turk (MTurk), Cloud Research approved MTurk workers (Cloud-R), and Prolific. Our results are threefold - first, MTurk is dominated both in terms of noise as well as elasticity of response towards a treatment intervention. Second, Prolific offers greater inferential power due to low cost and low noise but has almost zero elasticity of response. And finally, Cloud-R exhibits a similar elasticity of response to the lab samples, but the cheaper observations lead to substantially high inferential power for a simple experiment such as ours.

Table of Contents

List of Tables

List of Figures

Preface

This dissertation would never have been completed without the unwavering support of my partner Anshul Mittal who has been the light of my life through all the doubts and uncertainties. I would not have made it this far without the love and sacrifices of my parents. Working on this dissertation was fun and exciting because of my amazing network of friends, including but not limited to Mallory Avery, Kelly Hyde, Marissa Lepper, Beatriz Ahumada, Rachel Landsman, Jessica LaVoice and many many more. I would like to thank my advisor Richard Van Weelden for never making me feel stupid and encouraging my research ideas while helping me critically evaluate them. I could not have become an experimental economist without the support of my advisor Lise Vesterlund who took me under her wing early on, whose careful eye never missed any design details, and who always motivated me to become a better researcher. Finally, I would like to thank my entire dissertation committee and the faculty of the Department of Economics at the University of Pittsburgh for their advice and suggestions to make these papers the best I could make them.

1.0 Can Temporary Affirmative Action Improve Representation?

If employers hold biased beliefs about a particular group, they may be less likely to hire workers from this group, preventing them from learning and correcting their beliefs. This paper explores whether temporary affirmative action can correct biased beliefs and in turn improve representation even after the policy is lifted. I elicit employer hiring decisions and beliefs about potential employee performance in two between-subject experimental treatments: a control treatment without affirmative action and a temporary affirmative action treatment. While beliefs and hiring are biased against women in the control treatment, I find in the temporary affirmative action treatment that representation improves even after affirmative action is lifted. This increase is partially driven by employers' beliefs about performance. Further, employers who are most likely to discriminate against women show the greatest reduction in gender bias in beliefs which in turn explain the shift in hiring choices toward women. The results shed light on how temporary affirmative action policy can alleviate self-perpetuating under-representation by correcting biased beliefs.

1.1 Introduction

Employers who hold biased beliefs against a group of workers are less likely to hire candidates from that group [318, 311]. As employers only get feedback about performance of workers they hire and not the entire pool of applicants, biased beliefs may be sustained due to hiring decisions that limit belief updating. Thus, biased beliefs can trigger self-perpetuating under-representation of the group of workers against whom the employers are biased [256]. External interventions such as affirmative action may help break this pattern. Increased exposure can provide the employers an opportunity to correct biased beliefs and potentially improve long term representation.

In this paper I study an environment where employers hold biased beliefs about the performance of women. Within this environment, I explore if a temporary affirmative action

quota leads to lasting improvement in the representation of women. I track the evolution of beliefs about performance and ask if quotas reduce bias in beliefs after they have been lifted. Finally, I study the extent to which beliefs about performance help explain the potential changes in hiring.

Under-representation of women is a well-documented phenomenon. It is seen both in horizontal as well as vertical segregation where women are underrepresented in stereotypical male fields [47, 269, 322, 193, 257, 345] and in managerial positions [271, 314, 79]. This under-representation of women could be due to a variety of reasons including, but not limited to, discriminatory hiring practices.

The economics literature defines three distinct types of discrimination. Taste-based discrimination [46] provides a model wherein employers' prejudices or preferences affect hiring choices. A model of statistical discrimination [5, 306, 26] instead predicts under-representation only if a group is expected to have lower performance. In this model, discrimination arises against a group of workers when their average performance is lower than that of the other group(s). Scholars have recently formalized another model of discrimination - inaccurate statistical discrimination [172, 54]. This model also bases hiring choices on expected performance but employers inaccurately believe one group to be better performers than the other(s). Discrimination thus arises in this model against whom the employers are biased. For example, if men and women are equally good at math, but an employer inaccurately believes men to be better than women, then a woman worker will be less likely to be hired for positions that require such skills.

While affirmative action mechanically improves representation irrespective of the nature of underlying discrimination, it is of interest to explore whether we find a lasting improvement in representation even after the policy is lifted. For example, if employers exhibit taste-based discrimination against women, a mandated exposure to women under affirmative action may reduce antipathy towards them and result in improved representation of women even when affirmative action is lifted. However, note that the effect well could go in the opposite direction with exposure increasing resentment towards women. If instead employers hold biased beliefs against women and these impact hiring choices, then mandatory hiring of women under affirmative action will provide additional feedback about women's true relative

performance. Any resultant correction in beliefs can improve representation after the quota is lifted.[1]

Using a series of three online experiments, I explore a setting with biased beliefs against women and study dynamics of hiring choices and beliefs about performance to offer insight into the potential effects of temporary affirmative action. An experimental setting is ideally suited to answer the questions I pose due to three major challenges. Firstly, to study the causal impact of temporary affirmative action on representation, there is need for random assignment between hiring environments with no restrictions and a temporary affirmative action policy for women. Secondly, to examine a setting with biased beliefs one needs to find a task where there are no gender differences in actual performance but men are believed to outperform women. And finally, in order to understand learning as a potential mechanism, one needs a reliable measure of beliefs which is consistent overtime. My experiment is carefully designed to overcome these analytical challenges.

Participants in the first experiment serve as workers and answer trivia questions from a number of male-type trivia categories [113, 61, 112]. Experiment 2 elicits beliefs on gender difference in performance of participants in the previous experiment. Of key interest here is finding a trivia category where evaluators hold biased beliefs on the gender gap in performance. The first two experiments identify sports as the trivia category where there are no gender differences in average performance but where men are believed to outperform women. Workers' performance in the sports trivia quiz becomes the foundation for employer beliefs and hiring decisions which are captured in the third experiment of the study.

Participants in the employer experiment are presented with four randomly selected resumes of a gender-balanced set of workers from experiment 1. Employers' beliefs are elicited about expected performance of all presented worker resumes on the sport trivia quiz. They are then asked to hire two out of this group of four workers. Finally, employers get feedback about actual performance of both of their hired employees. Employers proceed with these belief elicitations and hiring decisions for six rounds with different workers in each round, allowing us to see how employer beliefs and hiring decisions update based on feedback. The

[1] A temporary affirmative action policy can also produce lasting effects under accurate statistical discrimination by creating economic incentives conducive for investment in skill enhancement by the underrepresented group for a well-defined set of initial parametric conditions [111].

experiment has two between-subject treatments: a control treatment in which there are no restrictions on who the employers can hire; and a temporary affirmative action treatment. In the temporary affirmative action treatment, the first three rounds have a quota policy for women wherein at least one of the two hired employees must be a woman, and this policy is subsequently removed in the last three rounds.

We find that without affirmative action, women are 13 percentage points less likely to be hired than men and this gender gap in hiring persists and slightly worsens to 16 percentage points in the last three rounds of the experiment. By design, affirmative action improves representation of women while in effect, but we find that the positive effects of the policy remain after it is lifted where women are now 6 percentage points less likely to be hired than men in the last three rounds which is much smaller than the male advantage in the control treatment. We find a positive and significant lasting effect of the temporary affirmative action treatment on representation of women.

Next, we turn to beliefs about performance and find them to be statistically similar across the two experimental treatments. The effect of increased exposure to women under a temporary affirmative action treatment is however muted due to two factors - first, some employers disregard information leading to noisy data, and second, change in beliefs depends on initial gender bias in beliefs. Accounting for either of these factors reveals that beliefs are changing in response to temporary affirmative action treatment. In the last case, there are a number of employers in this sample who are less biased in their beliefs and less likely to discriminate against women. For this subgroup of employers, quotas are not binding and their exposure to women workers not limited. As a result, we can expect to see no differential impact of the temporary affirmative action treatment on changes in beliefs and we indeed find this to be the case.

On the other hand, we find that employers who are more biased and more likely to discriminate against women exhibit a significant reduction in gender bias in beliefs due to a temporary affirmative action treatment even when quotas are lifted. Within this subgroup of employers we also find a positive and significant effect of the treatment on representation of women. And further, this positive effect of the treatment on hiring of women is partially explained by beliefs about performance. As such, this subgroup of employers who are

more biased and more likely to discriminate against women offers us the starkest possible comparison to explore dynamics of inaccurate statistical discrimination under a temporary affirmative action policy. The data from this study points to the possibility of reduction in gender bias in beliefs due exposure under a temporary quota which can in turn lead to a lasting improvement in representation of women.

The rest of the paper is organized as follows: section 1.2 details the experiment design; section 1.3 lays out the main hypothesis from the employer experiment; section 1.4 summarizes the main results; and section 1.5 concludes and discusses the results within the context of existing economics literature and policy perspectives.

1.2 Experiment Design

The study consists of three experiments. The first experiment, detailed in section 1.2.1, collects performance data on trivia quizzes and demographic characteristics of participants. The second experiment, detailed in section 1.2.2 elicits beliefs about performance of workers with different demographic characteristics for different trivia categories. To construct a setting with biased beliefs, these two experiments help identify a trivia quiz category where men on average are perceived to be better performers than women without any actual gender difference in performance. The third and the main experiment, detailed in section 1.2.3 explores employer beliefs and hiring choices over multiple rounds and randomly selected workers from experiment 1.

An important note here is that a worker's performance measure of interest is reduced to a binary measure of success and failure. A worker is considered successful if their score in a trivia category exceeds the median score within that category.[2] For every belief elicitation, performance is measured as the likelihood of a worker being successful in a particular trivia category. This binarization is helpful in reducing the parameters in the belief elicitation to a single probability measure while abstracting away from distributional concerns. All experiments are programmed using Qualtrics and run on Prolific during the summer of

[2]The median score is 6 for business, 9 for sports and 7 for video games.

2021.[3]

1.2.1 Experiment 1: Workers

The objective of this first experiment is to identify and collect performance data on a real effort task for which there are no gender differences in the average likelihood of success of men and women. The task here is a trivia quiz from three male-type trivia quiz categories: sports, business, and video games. Previous literature shows trivia quizzes from male-type categories can invoke biased beliefs resulting from stereotypes on account of the associated domain types – male or female [113, 61, 112]. In this experiment, a gender balanced pool of 50 participants are given 5 minutes to work on 30 multiple choice questions, 10 from each trivia category.[4] Upon completion of the quiz, participants proceed to a demographic survey that collects data on 12 personal attributes of participants which are used in subsequent experiments to anonymously identify them.[5] Participants are paid a \$1.50 completion fee along with \$0.10 bonus payment for each correct answer on the quiz.[6]

We find that the magnitude of gender difference in performance is small and statistically insignificant for trivia categories of sports and business. And while the gender difference in performance is statistically insignificant for trivia category of video games as well, we nonetheless eliminate it due to a relatively larger magnitude (see Appendix A for details).[7] The next experiment proceeds with trivia categories of sports and business.[8]

[3]In a comparison study between physical laboratory, Amazon's Mechanical Turk, Prolific, and the online version of the laboratory - Prolific has been shown to be superior on the extensive margin for elicitations not involving a social tension among participants due to its low noise and substantially lower cost of collecting data [201].

[4]All questions appear on the same page and their order is randomized. Further, the order of options within each question is randomized.

[5]Demographic survey includes attributes which are commonly found on resumes like age, years of schooling, employment status, gender, geographical location as deduced from the time zone of residence, whether or not someone did any volunteer work in the past 5 years, number of languages they could speak, and some arbitrary characteristics including height, whether or not they are registered voters in the United States, whether or not they are smokers, whether they preferred cats or dogs, and what is the operating system of their phones.

[6]Detailed instructions, sample trivia quiz questions and screenshots of the experiment can be found in the online appendix: https://tinyurl.com/y34avc89

[7]Given the levels and standard deviations, the gender difference for video games is likely to become statistically significant with a small increase in the sample size. A simple power calculation in STATA for a test of means confirms this conjecture.

[8]The likelihood of success in business differed significantly by age, number of spoken languages, smoking

1.2.2 Experiment 2: Evaluators

To understand how perceptions about likelihood of success change with demographic characteristics for each trivia quiz category, a new set of participants on Prolific are recruited as evaluators. Participants are recruited independently for evaluations related to the business and sports trivia category quizzes. For these evaluations, the 12 demographic characteristics are first divided in two groups. This generates 24 demographic groups as shown in Table 1. For example, a categorical variable like time zone of residence is divided in eastern vs. not eastern while a continuous entry variable such as age is divided based on the median worker's response.

Evaluator beliefs about performance are then elicited for a representative worker from the first experiment within each demographic group. For each elicitation, the evaluators are asked to indicate how likely they think it is that a randomly selected worker from the first experiment in the respective group is successful.[9] As before, success is defined as getting a score in excess of the median score in the specific trivia quiz.

Evaluators are paid a $2 completion fee along with a chance to earn a bonus payment of up to $2. For this bonus payments, two of the submitted guesses by the evaluators are randomly selected and they are paid $1 each for the accuracy of their guess using minimum information binarized scoring rule [219, 127]. The evaluators are informed that truthful reporting will secure the maximum chances of securing the bonus. The experiment ends with a short exit survey along with payment information.[10]

A gender balanced pool of 80 and 61 evaluators are recruited for business and sports trivia category evaluations respectively.[11] There are two key results from this experiment. First, for the trivia category of business we see no gender difference in perceived likelihood

behavior, and operating system of their phone. The likelihood of success in sports on the other hand differed by age, employment status and whether someone preferred cats or dogs. This is indicative that not only common resume attributes, but also arbitrary characteristics can be explanatory of likelihood of success in the respective trivia categories.

[9]The order in which the 12 characteristics are presented to the evaluators are randomized. Moreover, the order of the two groups within each characteristic is also randomized.

[10]Attention checks are added in this study and a post hoc assessment is conducted to verify response data from these attention checks, as well as identify any outliers for time spent on the instructions page. No participants are identified to be excluded from analysis through this assessment.Detailed instructions and screenshots of the experiment can be found in the online appendix: https://tinyurl.com/y34avc89

[11]Power calculations are parameterized using [112] and can be made available upon request.

Table 1: Demographic characteristic groups for evaluator experiment

Demographic Characteristic	Group 1	Group 2
	(1)	(2)
Age	27 yrs or below	Above 27 yrs
Gender	Male	Female
Height	5 ft 6 in or less	More than 5 ft 6 in
Time Zone	Eastern	Not eastern
Education as Years of Schooling	13.5 yrs or below	Above 13.5 yrs
Employment Status	Employed	Not employed
Volunteer Work	Yes	No
Voter Registration	Yes	No
Smoker	Yes	No
No. of Spoken Languages	1	More than 1
Cat/Dog Person	Cat	Dog
Phone OS	Android	iOS

Notes - This table presents the set of 24 demographic characteristic groups for which beliefs were elicited in the evaluator experiment. For each demographic characteristic the two groups are defined based on a median worker's response from the worker experiment.

of success. Business trivia category is thus seen as unlikely to invoke biased beliefs against women and hence eliminated for use in the employer experiment. Next, we see a large and significant gender difference in perceived likelihood of success for the sports trivia category, one where men are believed to outperform women (see details in Appendix A).[12] The employer experiment thus proceeds with the sports trivia quiz as the underlying real effort task for worker performance because there are no gender differences in average likelihood of

[12]The perceived likelihood of success on sports varies by all demographic characteristics except time zone and number of spoken languages, which is indicative of not only common resume attributes, but also arbitrary characteristics being explanatory of perceived likelihood of success. These observable characteristics must then be controlled for in the subsequent analysis.

success of men and women, but men on average are believed to perform better than women.

1.2.3 Experiment 3: Employers

Experiment 3 is used to elicit performance evaluations and hiring decisions and is the main experiment of interest. Participants from the first experiment form the pool of workers over which hiring decisions are elicited. Workers' demographic characteristics of years of schooling (level of education), employment status, geographical location (as measured by the time zone of residence) and number of spoken languages, along with information about gender - male or female, are used to create resumes. Workers' performance on the sports trivia quiz is the foundation for employer decisions. As before, the performance criterion of interest is the probability of getting a score that exceeds the median worker's score i.e., get a score of 9 or above.

The employer experiment consists of six rounds with two decision stages in each round-evaluation and hiring. A round ends with feedback based on the employer decisions. Participants are paid a completion fee of $2 along with a bonus payment of up to $3 from one decision stage selected at random.

In each round, employers are presented four randomly selected worker resumes – two men and two women.[13] Employers first complete evaluation decisions for all the four worker resumes where using a scale of 0 to 100 they report on how likely they think it is that the worker is successful in the sports trivia quiz i.e., got a score of 9 or above.[14] If this stage is selected for bonus payment, one worker resume is randomly selected, and employers are paid a bonus of $3 for the accuracy of their guess for the selected worker. The elicitation is again incentivized using minimum information binarized scoring rule [219, 127] and employers

[13]For example, 24 resumes (out of 50) are drawn randomly (12 men, 12 women) and they are randomly divided in 6 groups of two men and two women each. One sequence of the six groups comprised one employer assignment wherein the first group is presented in round 1, the second group in round 2, and so on. From this one draw of 24 employees and six groups, six employer assignments are created to balance across rounds i.e., group 1 is 1st in sequence for employer 1, then permuted to be 2nd in sequence for employer 2 and so on. Nine unique draws of 24 resumes each are carried out for creating 54 assignments and these assignments are assembled in a matrix and read into Qualtrics using JavaScript. The same 54 assignments are implemented across all employers' experimental sessions.

[14]In making evaluation decisions, employers are informed that the number of correct answers ranged from 2 to 10.

are informed that truthful reporting would secure them the highest chance of winning the bonus. Those looking for more information could access a non-mathematical explanation of the payment rule at the end of the experiment [358].[15]

Next, employers make hiring decisions from the same set of four workers. They are asked to select two workers and also to rank them as their 1st and 2nd preferred candidates for hiring.[16] If this stage is selected for bonus payment, they receive \$2 if their 1st hired worker is successful on the sports trivia quiz; \$1 if their 2nd hired worker is successful; and \$0 otherwise. The round ends with feedback where the employers learn whether their two employees are successful or not. This design feature mimics the real world settings where employers only learn about performance of their hired employees and not of the workers they do not hire.

At the end of the round 6 hiring decision stage, employers enter a final decision stage. Here they are offered a chance to get costless information on workers' success and to potentially revise hiring choices for round 6. They are again presented with the four worker resumes from round 6 along with their evaluation and hiring decisions indicated on each resume. They are then offered an option to "Reveal the Quiz Outcome" using a button below each resume. They could select this option for as many workers as they like or for no worker at all. On the next screen, they are given the information for workers they select the option to reveal the quiz outcome in the form of whether a particular worker was or was not successful. Employers are then given the option of revising their hiring selections and are given feedback on those ultimately hired. If they do not select the option to reveal the quiz outcome for any worker, they proceed directly to receive the feedback on their original hiring selections and skip this final decision stage.

This final stage is used to identify employers who do not opt for this costless information for any worker. Such employers are classified as being likely to disregard information about employee performance and this classification allows us to explore heterogeneous effects on hiring and beliefs. It also allows for potential reduction in noise in the data as participants

[15]The participants who sought further clarifications are directed to my email address and I did not receive any requests for more information.

[16]To aid hiring decisions, their submitted evaluation decisions are indicated along with each employee resume

who do not demand the costless information are also likely to be less attentive.

The employer experiment has two between-subject treatments- a control treatment and a temporary affirmative action treatment. Affirmative action takes the form of a soft quota where for the first three rounds in the experiment, employers have to hire at least one woman.[17] This policy is then removed in the last three rounds. A comprehension check is added in round 4 to ensure that the change in hiring policy is noted by participants in the temporary affirmative action treatment. The employers in the control treatment on the other hand do not have affirmative action policy for women in any round and nor do they learn about the possibility of affirmative action.[18]

The summary statistics for the employer sample are presented in Table 2. Panel A shows the demographic characteristics of the employers across the control and the affirmative action treatments. The employer characteristics are balanced across the two treatments and the sample comprised of 216 employers in the temporary affirmative action treatment and 214 in the control.[19] The employer sample is gender balanced by construction and the experiment was run in 8 sessions of 54 participants each with randomization into treatments implemented at session level.[20] Panel B of the table summarizes the characteristics of the resumes as presented to the employers. Since the same set of randomly selected resumes are used between the two treatments, we unsurprisingly find resume characteristics to also be balanced. Notably, within the set of resumes presented to the employers, 52% of women and 49% of men are successful in the sports trivia quiz.

[17]This policy is binding in the experiment such that on the hiring decision screens, participants could not move forward in the experiment if at least one of their hired workers is a woman.

[18]Following [288] the knowledge of affirmative action is considered to be a part of the treatment. Detailed instructions and screenshots of the experiment can be found in the online appendix: https://tinyurl.com/y34avc89

[19]2 control employers had to be excluded on account of JavaScript not loading onto their browsers. As a result, there is no record of the associated resume information that they saw during the experiment.

[20]The experiment contains multiple comprehension checks designed such that even if a participant does not read the instructions carefully, they could understand their main goals in the study from the comprehension questions. A participant could not proceed in the study without answering a comprehension check correctly. The experiment also includes simple attention checks where a submission is approved and paid automatically if all attention checks are answered correctly. Only minor mistakes of having an extra space or misidentifying a cursive "n" as "r" are overlooked. Additionally, if participants get one attention check wrong but answer all comprehension checks correctly on their first attempt, they are also approved. All others are rejected, and their spots are opened to new participants in the study. A total of 12 participants are rejected from the employer study.

Table 2: Employer experiment summary statistics

		(1) Control	(2) TempAA	(3) Mean Diff.
Panel A: Employer Demographic Characteristics				
Age (in years)		30.72	31.31	-0.59
Gender	Men	0.50	0.50	0.00
	Women	0.48	0.49	-0.01
	Other	0.02	0.01	0.01
Years of Schooling	16 years or less	0.58	0.58	0.00
(level of education)	More than 16 years	0.42	0.42	0.00
Employment status	Currently Employed	0.68	0.70	-0.02
	Not Currently Employed	0.32	0.30	0.02
Sample Size		**214**	**216**	
Panel B: Presented Resume Characteristics				
Gender	Men	0.50	0.50	0.00
	Women	0.50	0.50	0.00
Years of Schooling	13.5 years or less	0.47	0.47	0.00
(Level of Education)	More than 13.5 years	0.53	0.53	0.00
Employment status	Currently Employed	0.66	0.66	0.00
	Not Currently Employed	0.34	0.34	0.00
Time Zone of Residence	Eastern	0.45	0.45	0.00
(Geographical Location)	Central/Mountain/Pacific	0.55	0.55	0.00
No. of Spoken Languages	1	0.66	0.66	0.00
	More than 1	0.34	0.34	0.00
Avg Prop Successful	Men	0.49	0.49	0.00
	Women	0.52	0.52	0.00
	Overall	0.50	0.50	0.00

Notes - *** p<0.01, ** p<0.05, * p<0.1, for a two-tailed test of mean difference between the control and the temporary affirmative action treatment values against a null of 0.

1.3 Hypotheses

Given the employer experimental setting, gender discrimination in hiring is defined as a systematic tendency toward hiring less women. We can further determine the nature of the underlying discrimination through the decision making process. For example, if hiring decisions are determined purely by perceived likelihood of success of each worker, then the resulting discrimination is statistical and grounded in beliefs. If instead, hiring decisions are not determined by beliefs but by an innate preference against women, then the resulting discrimination will be taste-based. Note that discrimination can also be a combination of statistical as well as taste-based where an employer may be making hiring decisions partly through beliefs and partly through preferences. Since the current setting is designed to be that where women on average are inaccurately believed to be worse performers than men, any gender discrimination in hiring that we might observe can either be inaccurately statistical in nature or be taste-based or a combination of both.

If the nature of underlying discrimination is in parts inaccurately statistical, then in absence of affirmative action for woman, we can expect employers to have limited exposure to women workers. With feedback about hired employees at the end of each round, we can then expect beliefs about men to converge toward the true value while the same argument cannot be extended gender difference in beliefs due to limited feedback about women workers. As a result, we can expect discrimination against women to persist in the control treatment of the experiment. With a temporary affirmative action treatment intervention, employers are exposed to more women workers. Employers get the opportunity to learn not only about men but also about women and we can expect the gender difference in beliefs to now also converge toward the true value, which by design is 0. The effect of affirmative action for women can then extend beyond the policy instance and produce lasting improvement in representation of women as the hiring choices are based in beliefs. This argument forms the following three hypotheses for this experiment:

In environments where beliefs about performance are biased against women relative to men:

H1: Temporary affirmative action for women improves representation of women even after

affirmative action is lifted.

H2: Temporary affirmative action for women reduces the bias in beliefs about performance of women relative to men.

H3: Beliefs about performance help explain the lasting improvement in representation of women due to a temporary affirmative action policy.

1.4 Results

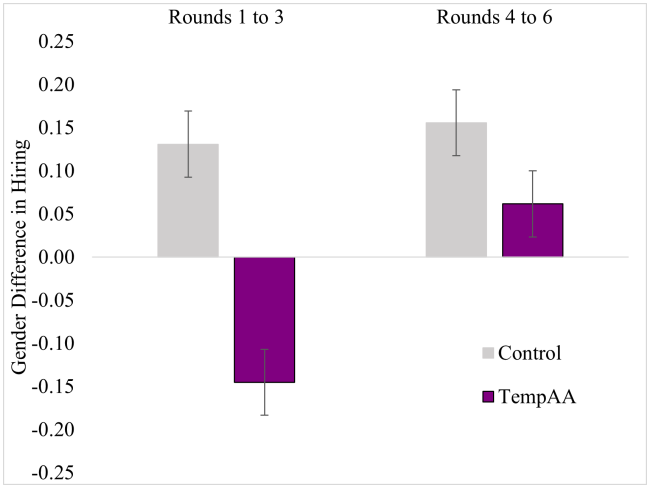
This section describes results from the employer experiment. To address the first hypothesis, I track the hiring choices of employers in the temporary affirmative action treatment relative to the control treatment in section 1.4.1. Next, I explore dynamics of beliefs about performance in both experimental treatments in section 1.4.2. And finally in section 1.4.3, I discuss beliefs as the mechanism behind changes in hiring choices. Additional analysis is presented in section 1.4.4 addressing gender difference in the distribution of hired employees as 1st and 2nd preferred candidates for hiring.

1.4.1 Evolution of Hiring Decisions

Results from the raw data of hiring decisions are summarized in Figure 1. We find that in the control treatment without affirmative action, men are 13 percentage points more likely to be hired than women ($p < 0.001$). This differences increases to 16 percentage points ($p < 0.001$) in the last three rounds though the difference is not statistically significant. In the temporary affirmative action treatment, women are 15 percentage points more likely to be hired than men ($p < 0.001$) in the first three rounds when quotas are in effect. When the quotas are lifted in the last three rounds, we again find a male advantage but at 6 percentage points it is much smaller than that seen in the control. These raw proportions give indicative evidence in favor of the first hypothesis that temporary affirmative action can improve representation of women even after the policy is lifted.

To confirm these findings we use a regression framework and model the hiring decisions

Figure 1: Difference in Proportions of Men Hired vs. Women Hired



Notes - This figure shows mean gender difference in proportions of men hired vs. women hired. The left panel shows the first three rounds where there is a quota for women in effect under the temporary affirmative action treatment, and the right panel compares the two experimental treatments for the last three rounds when affirmative action is removed. Error bars correspond to 95% confidence intervals.

as follows

$$\pi_{ijr} = x_{ir}\beta + u_{ir} + \epsilon_{ijr} \tag{1}$$

where π_{ijr} represents employer i's profit from hiring worker j in round r, with $j \in J = \{1, 2, 3, 4\}$ and $r \in \{1, ..., 6\}$. Further, x_{ijr} represents the row vector of covariates and β is the column vector of coefficients. The u_{ir} represents the group level heterogeneity where a group represents a single round where an employer makes decisions over 4 resumes, and ϵ_{ijr} is the observation-level error term. Employer i will hire the worker j in any given round if hiring that worker maximizes their profit:

$$Pr(y_{ikr} = 1) = Pr\{max(\pi_{ikr}) = \pi_{ijr}\}\forall k \in J \tag{2}$$

Table 3: Hiring results - Conditional Fixed Effects Logit Regression

Dependent Variable:	Rounds 4 to 6 I(Hired)	
	(1)	(2)
I(Female)	-0.467***	-0.416***
	(0.0684)	(0.0749)
I(TempAA)*I(Female)	0.282***	0.292***
	(0.0965)	(0.0986)
N	5160	5160
Worker controls		Yes

Notes - *** p<0.01, ** p<0.05, * p<0.1; This table presents conditional fixed logit regression results from hiring decisions where a group comprises a single round where an employer makes decisions over 4 resumes. Dependent variable is an indicator variable =1 when a worker is hired. Worker controls include demographic characteristics presented on their resume - employment status, education, number of spoken languages, and time zone of residence. All options within each resume characteristic are aggregated in two groups characterized by the median worker's characteristic from the worker experiment. Standard errors are shown in parentheses.

Assuming a standard Type I extreme value for ϵ_{ijr}, this gives rise to the following choice model where $j = 0$ is the baseline outcome:

$$Pr(y_{ijr}|x_{ijr}, \beta) = \begin{cases} \frac{1}{1+exp(x_{ijr}\beta)} & \text{if } j = 0 \\ \frac{exp(x_{ijr}\beta)}{1+exp(x_{ijr}\beta)} & \text{if } j = 1 \end{cases} \tag{3}$$

Table 3 summarizes the findings from a conditional fixed effects logit model estimation of eq (3) using data from rounds 4 to 6 when quotas for women are lifted in the temporary affirmative action treatment. The primary set of regressors in this estimation include indicator variables for a female worker ($I(Female)$) and its interaction with the temporary affirmative action treatment indicator ($I(TempAA)$).[21]

[21]The experiment does not involve clustered sampling and there is no evidence for the possibility of randomization failure between the control and temporary affirmative action treatment, and so clustering of standard errors is seen as unnecessary as per the recommendations from [1]. However, results are robust to

In the control treatment without affirmative action, the log odds of being hired (vs. not) are expected to be 0.467 less for women than men in rounds 4 to 6 ($p < 0.001$). In the temporary affirmative action treatment and quotas for women are removed, the log odds of a woman being hired are expected to be 0.185 lower than men ($p = 0.006$). The temporary affirmative action treatment significantly improves log odds of a woman being hired by 0.282 ($p = 0.003$) These results are robust to controlling for characteristics of workers as presented to the employers on the resumes.

As such the first result to emerge from this analysis is that women are significantly less likely to be hired relative to men without affirmative action. Quotas mechanically improve the representation of women and these positive effects on hiring of women continue even after the policy is withdrawn. Temporary affirmative action thus improves representation of women even after it is lifted.

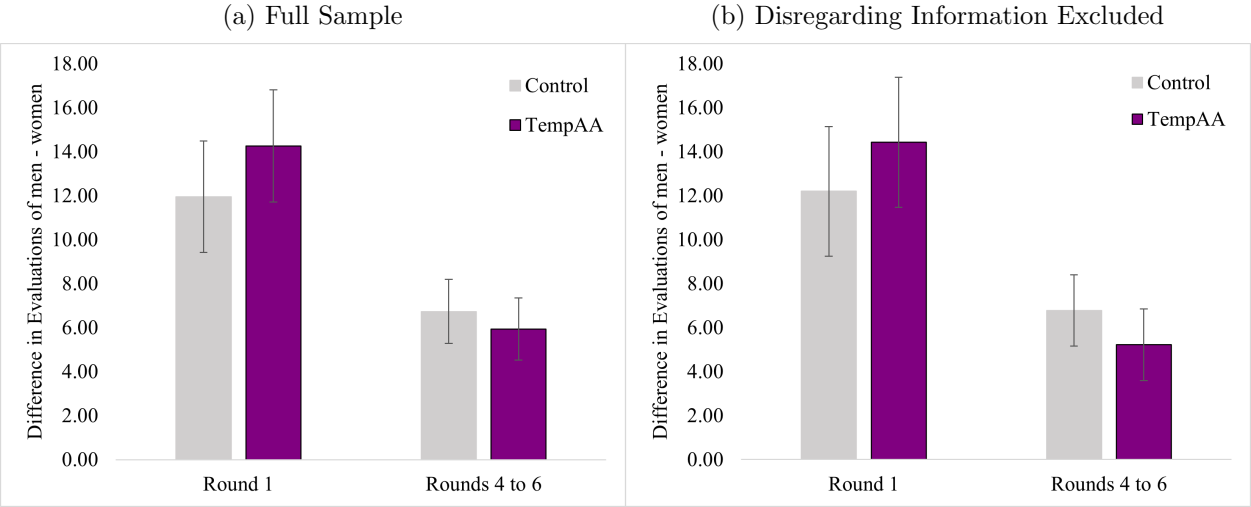
1.4.2 Evolution of Beliefs About Performance

With choice over hiring and thus which candidate an employer is getting feedback about, we find in rounds 1, 2, and 3 of the control treatment that approximately 43% of observed signals are about women employees. By comparison in rounds 1, 2, and 3 of the temporary affirmative action, about 57% of signals are about women employees. Thus, affirmative action increases signals about women by 32%. Does this increased exposure to women lead to employers learning that their beliefs are inaccurate and results in belief updating?

Figure 2 panel (a) plots the raw mean gender difference in evaluations split by the two treatments - first for just round 1 and then jointly for rounds 4 to 6. At the outset in round 1 of the control treatment, men are believed to be 12 percentage points ($p < 0.001$) more likely to succeed than women. This gender bias reduces to 7 percentage points ($p < 0.001$) in the last three rounds. In the temporary affirmative action treatment, employers start with a belief that men are 14 percentage points ($p < 0.001$) more likely to succeed than women. The similarity between the two treatments is to be expected given that the round 1 belief elicitation is carried out before participants learn about the hiring environments.

clustering standard errors at session level using wild bootstrapping procedures that test the null hypothesis that the coefficient on the interaction term is zero. This extends to all following analysis in the paper.

Figure 2: Difference in Evaluations of Men and Women



Notes - This figure shows mean gender difference in evaluations of men and women. Sub-figure (a) plots the graph for the full sample while sub-figure (b) excludes employers who disregard information about performance of workers as identified by the final stage of the experiment. The left panel in both figures represents round 1 evaluations, and the right panel compares the two experimental treatments for the last three rounds where quota is removed under the temporary affirmative action treatment. Error bars correspond to 95% confidence intervals.

When quotas are removed in the temporary affirmative action treatment, the gender bias is reduced to 6 percentage points ($p < 0.001$), a difference that is not statistically different from the belief held in the last three rounds of the control treatment.

This lack of a treatment effect on gender bias in beliefs is confirmed in a regression analysis shown in Table 4. The evaluations are modeled as follows:

$$Evaluation_{ijr} = x_{ir}\beta + u_{ijr} + \epsilon_{ijr} \tag{4}$$

where $Evaluation_{ir}$ represents employer i's evaluation of jth worker in round r, with $j \in \{1, ..., 4\}$ and $r \in \{1, ..., 6\}$. x_{ijr} represents the row vector of covariates, β the column vector of coefficients, u_{ir} is the group level heterogeneity where a group comprises a single round where an employer makes evaluation decisions over 4 resumes and ϵ_{ijr} is the observation-level error term. The main regressors include indicator variables for a female worker and its

Table 4: Evaluations results - OLS with fixed effects

Dependent Variable:	Rounds 4 to 6			
	Evaluation			
	(1)	(2)	(3)	(4)
I(Female)	-6.739***	-6.533***	-6.785***	-6.739***
	(0.559)	(0.595)	(0.633)	(0.674)
I(TempAA)*I(Female)	0.795	0.782	1.564*	1.590*
	(0.789)	(0.783)	(0.879)	(0.875)
Constant	68.74***	65.48***	68.49***	65.98***
	(0.279)	(0.570)	(0.311)	(0.634)
N	5160	5160	3816	3816
Worker controls		Yes		Yes
Disregard excluded			Yes	Yes

Notes - *** p<0.01, ** p<0.05, * p<0.1; This table presents OLS regression results on evolution of beliefs about performance elicited as evaluations on a scale of 0 to 100. The estimation uses a fixed effects model where each group comprises a single round where an employer makes decisions over 4 resumes. Worker controls include demographic characteristics presented on their resume - employment status, education, number of spoken languages, and time zone of residence. All options within each resume characteristic are aggregated in two groups characterized by the median worker's characteristic from the worker experiment. Employers disregarding information are identified through the final stage of the experiment as those who do not opt in to get information on worker performance for any worker. Standard errors are shown in parentheses.

interaction temporary affirmative action treatment indicator. Data from rounds 4, 5 and 6 has been used for this estimation using a fixed effects model. We find that women are believed to be 6.74 percentage points ($p < 0.001$) less likely to succeed in the control treatment. Exposure to more women under temporary affirmative action treatment reduces the gender bias in beliefs by 0.79 percentage points which is statistically insignificant ($p = 0.314$).

This finding could be seen as evidence against the hypothesized evolution of beliefs, but one must be careful in this interpretation as there are two factors through which this effect is muted. First, the limited response in beliefs is found to be mechanically driven by the presence of employers who disregard the feedback given to them on actual performance of

hired employees. And second, the effects are muted by a high number of employers who hold little initial gender bias in beliefs and who hire women despite having no restrictions on their hiring decisions. I expand on each of these factors separately in the next two sections.

1.4.2.1 Employers Disregarding Information

Recall that the final decision stage in the experiment helps identify the employers who are likely to disregard information about employee performance. After the employer participants submit their hiring choices in the last round and before they learn about the outcomes of their hired employees, they go through the final decision stage. In this stage, they are given an opportunity to costlessly demand information on whether a worker is successful or not, by pressing a button for as many worker resumes as they like. On the next page, they see the actual performance of those workers they selected the option to reveal the outcome and have an opportunity to revise their hiring choices. If they do not select this option for any worker, they skip to the end of the round. Employers who skip to the end of round without getting any information about the workers are characterized as those likely to disregard information about employee performance during the experiment. We find a total of 112 employers (out of the total 430) as those who disregard information - 51 in the control treatment and 61 in the temporary affirmative action treatment. Presence of these employers in the sample can potentially suppresses the belief dynamics in two ways - first, because these employers are likely to disregard the information about employee performance, they will be likely to ignore the feedback about performance of hired employees provided at the end of each round. And second, they will also be likely to be inattentive and contribute to noise in the data.

Panel (b) of Figure 2 shows the mean gender difference in beliefs about performance and restricts sample of employers by excluding those who disregard information. In round 1 of the control treatment, men are believed to be 12 percentage points ($p < 0.001$) more likely to succeed than women. This gender bias reduces to 6.8 percentage points ($p < 0.001$) on average in rounds 4 to 6. The employers in the temporary affirmative action treatment have statistically indistinguishable starting point from the control group (14 percentage points with $p < 0.001$). The bias however reduces to 5.2 percentage points ($p < 0.001$) which is

still statistically indistinguishable from the control mean. However, the reduction in gender bias in beliefs under temporary affirmative action treatment after excluding employers who disregard information is more pronounced when compared to the full sample. As such in a regression framework shown in Table 4 column (3), the difference-in-difference coefficient doubles in magnitude and is now statistically significant ($p = 0.075$). We can thus conclude, that when we account for the possibility that some employer participants are inattentive or ignoring the information provided in the experiment, we find evidence in favor of the hypothesis that temporary affirmative action reduces gender bias in beliefs even after the policy is lifted.

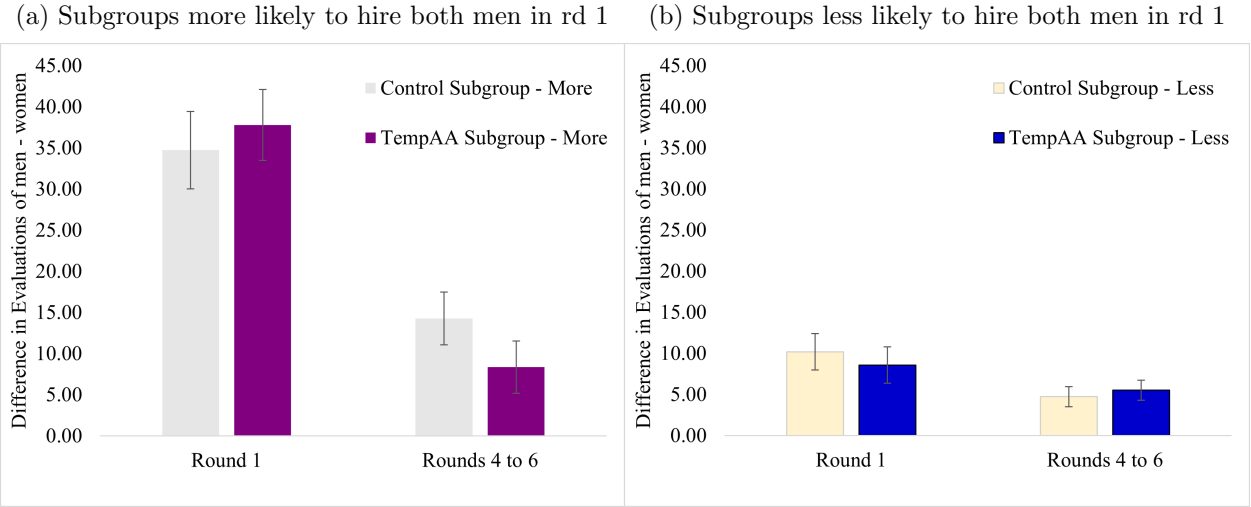
1.4.2.2 Heterogeneous Effects on Evolution of Beliefs

I now explore the second possibility that the observed effects on beliefs can be muted due to a high number of employers who do not hold biased beliefs and hire women despite having no restrictions on their hiring decisions. The conjecture here is that greater exposure to women with affirmative action, who are otherwise underrepresented, forces employers to get more feedback about women employees and thereby enables reduction in gender bias in beliefs. In order to test this in the starkest possible comparison, I divide the sample into subgroups - one that is more biased and more discriminatory towards women and other less so.

I first use data from the control treatment to estimate a logit regression where dependent variable is an indicator which takes value 1 if two men are hired in round 1, and explanatory variable is the gender difference in beliefs about performance of workers in round 1. Using this estimation, I then predict the probability of hiring two men for the control group of employers and determine a cutoff point based on the top 25% of employers most likely to hire two men in round 1 (0.32).[22] This cutoff is then used to classify two subgroups within the control group of employers- "More" ("Less") as those whose predicted likelihood of hiring both men in round 1 is more (lesser) than 0.32. To achieve a comparable classification for employers in the temporary affirmative action treatment, I use the previously estimated

[22]The logic behind choosing this cutoff is that about 25% of control groups of employers hire two men in round 1. However I also show that all results I present henceforth are not sensitive to this one cutoff point.

Figure 3: Difference in Evaluations of Men and Women - Heterogeneous Effects



Notes - This figure shows mean gender difference in evaluations of men and women. Sub-figure (a) shows the gender difference by experimental conditions for the subgroup of employers more likely to hire both men in round 1 and sub-figure (b) show this difference for employers less likely to hire both men. The cutoff to divide the two subgroups is at 32% which leaves about 25% of the sample for sub-figure (a). The left panel in both figures represents round 1 evaluations, and the right panel compares the two experimental treatments for the last three rounds where quota is removed under the temporary affirmative action treatment. Both figures exclude employers who disregard information about performance of workers as identified by the final stage of the experiment to minimize noise. Error bars correspond to 95% confidence intervals.

coefficients to predict a probability of hiring two men in round 1. Finally, I classify the two subgroups within the temporary affirmative action treatment group of employers based on the predicted probability of hiring two men in round 1 around the cutoff point of 0.32.

Raw gender difference in evaluations is shown in Figure 3. Panel (a) represents employers from the "More" subgroup, and panel (b) from the "Less" subgroup. Within the employers more likely to hire both men i.e., more likely to discriminate against women, we find that in the control treatment, men are believed to be 35 percentage points more likely to succeed than women ($p < 0.001$). This gender bias reduces to 14 percentage points ($p < 0.001$) in rounds 4 to 6. In the temporary affirmative action treatment however, the evolution of beliefs follows a faster learning trajectory where it starts with a gender bias of 38 percentage points ($p < 0.001$) which then reduces to 8 percentage points in rounds 4 to 6 when quotas are

Table 5: Heterogeneous effects in evaluations- OLS with fixed effects

Dependent Variable:	Rounds 4 to 6			
	Evaluation			
	Subgroup - More		Subgroup - Less	
	(1)	(2)	(3)	(4)
I(Female)	-14.11***	-14.69***	-4.251***	-3.831***
	(1.297)	(1.384)	(0.592)	(0.630)
I(TempAA)*I(Female)	4.844***	4.962***	-0.263	-0.268
	(1.755)	(1.740)	(0.850)	(0.843)
Constant	68.03***	64.24***	69.01***	65.88***
	(0.618)	(1.292)	(0.300)	(0.608)
N	1428	1428	3732	3732
Worker controls		Yes		Yes

Notes - *** p<0.01, ** p<0.05, * p<0.1; This table presents OLS regression results on evolution of beliefs about performance elicited as evaluations on a scale of 0 to 100. The estimation uses a fixed effects model where each group comprises a single round where an employer makes decisions over 4 resumes. Subgroup - More (Less) represents the subgroups of employers within the two experimental treatments who are more (less) than 32% likely to hire both men in round 1. Worker controls include demographic characteristics presented on their resume - employment status, education, number of spoken languages, and time zone of residence. All options within each resume characteristic are aggregated in two groups characterized by the median worker's characteristic from the worker experiment. Standard errors are shown in parentheses.

removed ($p < 0.001$). For the complementary subgroup which is less likely to discriminate against women (panel (b)), we find a small gender bias to begin with and that it doesn't change significantly over the course of the experiment for both treatments.

Regression estimates confirm the patterns observed from the raw data and are shown in Table 5. Columns (1)-(2) ((3)-(4)) estimate the effect of the temporary affirmative action treatment on beliefs for the "More" ("Less") subgroup of employers. Within the "More" subgroup, we find that without affirmative action in the control treatment women are evaluated to be 14 percentage points less likely to succeed than men in rounds 4 to 6 ($p < 0.001$). When quotas are removed however in the temporary affirmative action treatment, this gen-

der bias reduces by 4.8 percentage points ($p = 0.006$). We do not find similar effect for the "Less" subgroup in column (4) where there is a smaller and statistically significant gender bias against women ($p < 0.001$) in the control treatment which doesn't reduce at all under the temporary affirmative action treatment (-0.263 coefficient with $p = 0.757$).

Thus, within the employers less biased and less likely to discriminate against women, quotas are not binding and their exposure not limited. As a result we cannot expect beliefs dynamics to play out differently between the two experiment treatments and we indeed find this to be the case. On the other hand, in the starkest possible comparison within employers who are more biased and more likely to discriminate against women at the beginning of the experiment we find evidence in support of the second hypothesis of this study that temporary affirmative action for women reduces gender bias in beliefs about performance. This can further be generalized to say that as we restrict the sample to employers being progressively more biased and more likely to discriminate against women, we find that temporary affirmative action treatment produces stronger effect on reduction in gender bias in beliefs about performance in periods after the policy is removed (see appendix table B2).[23]

1.4.3 Mechanisms

Given the evolution of hiring choices and beliefs, the final piece in the puzzle is be to determine the extent to which beliefs affect changes in hiring choices. Revisiting the heterogeneous effects for changes in beliefs, we cannot expect inaccurate statistical discrimination within the "Less" subgroup of employers as these employers are less biased and less likely to discriminate against women without any policy intervention. On the other hand, the "More" subgroup is more biased and more likely to discriminate against women, thus creating conditions that make inaccurate statistical discrimination possible. We have shown that a temporary affirmative action treatment reduces gender bias in beliefs for this subgroup of employers. This section explores whether this reduction in gender bias can in-turn lead to improved representation of women even after the quotas are lifted.

Table 6 explores the hiring dynamics within the "More" subgroup (columns (1)-(2)) and

[23]The affect of temporary affirmative action is found to be even stronger if we further exclude employers who disregard information from the analysis (see appendix tables B1 and B3)

Table 6: Heterogeneous effects in evaluations as mechanism - Conditional Fixed Effects Logit Regression

Dependent Variable:	Rounds 4 to 6			
	I(Hired)			
	Subgroup - More		Subgroup - Less	
	(1)	(2)	(3)	(4)
I(Female)	-0.630***	0.0225	-0.413***	-0.207**
	(0.136)	(0.160)	(0.0791)	(0.0853)
I(TempAA)*I(Female)	0.322*	0.135	0.280**	0.314***
	(0.184)	(0.204)	(0.113)	(0.122)
Evaluation		0.0552***		0.0626***
		(0.00453)		(0.00355)
N	1428	1428	3732	3732

Notes - *** p<0.01, ** p<0.05, * p<0.1; This table presents conditional fixed logit regression results from hiring decisions where a group comprises a single round where an employer makes decisions over 4 resumes. Subgroup - More (Less) represents the subgroups of employers within the two experimental treatments who are more (less) than 32% likely to hire both men in round 1. Dependent variable is an indicator variable =1 when a worker is hired. Standard errors are shown in parentheses.

the "Less" subgroup (columns (3)-(4)). This estimation decomposes the effect of the temporary affirmative action treatment between the effect driven by beliefs about performance and any residual effect not explained by beliefs.

Within the "More" subgroup in the control treatment, the log odds of hiring a woman are 0.630 ($p < 0.001$) less than men. Temporary affirmative action treatment leads to an improvement of 0.322 units in the log odds of hiring of women ($p = 0.080$). The effect of beliefs on hiring decisions operates through two channels here. First, after controlling for beliefs the baseline gender discrimination in hiring disappears (log odds of 0.0225). And second, beliefs suggestively absorb the effect of the temporary affirmative action treatment where the moving column (1) to (2) the coefficient on the interaction term reduces in size

and is no longer statistically significant ($p = 0.508$). In the complementary "Less" subgroup on the other hand, we find a relatively smaller magnitude of gender discrimination against women in the control treatment where log odds of hiring a woman are 0.413 ($p < 0.001$) less than men . When we control for beliefs, this effect reduces but only slightly (to 0.207 log odds, $p = 0.015$). However, there is a pronounced effect of the treatment in this subgroup (0.280 log odds, $p = 0.011$) which sustains and in fact becomes slightly stronger when we control for beliefs (0.314 log odds, $p = 0.010$).[24]

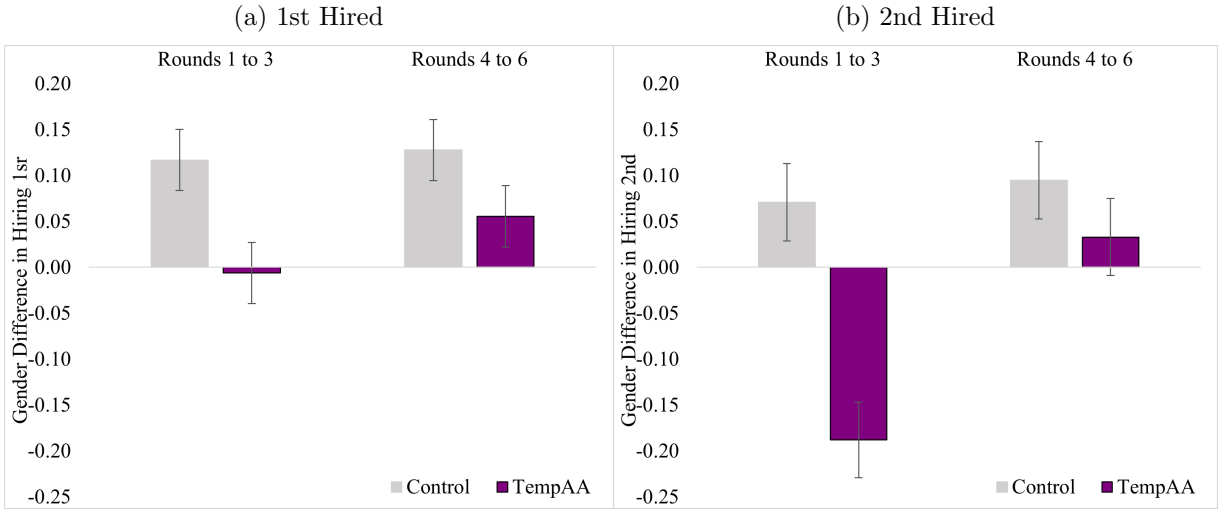
These is evidence in favor of hypothesis 3 wherein for the employers who are more biased and more likely to discriminate against women, we find that a temporary affirmative action treatment leads to a lasting improvement in representation of women. Moreover, beliefs about performance partially explain this improvement in representation of women and thereby we see rectification of inaccurate statistical discrimination within the "More" subgroup of employers. On the other hand, employers who are less biased and less likely to discriminate against women, they are not susceptible to changes in beliefs due to a temporary affirmative action treatment. We do however find there to be a positive and significant effect of the temporary affirmative action treatment on representation of women which is not operating through beliefs within this subgroup of employers.

1.4.4 Additional Analysis - Distribution within Hiring Choices

To further analyze the hiring choices, I explore the ranking distribution of hired employees and ask if temporary affirmative action treatment can induce effects powerful enough for women to not only be included as a second candidates but also become the first choice for hiring by employers. Figure 4 (a) shows the gender difference in proportions of the 1st hired employees. In the control treatment without affirmative action we find a large male advantage where men are 12 percentage points ($p < 0.001$) more likely to be hired than women. This male-advantage worsens overtime to 13 percentage points ($p < 0.001$) in the last three rounds of the experiment. In the temporary affirmative action treatment on the other hand, there is equitable hiring of men and women when quotas are in effect. When

[24]Results become stronger when we exclude employers who disregard information B4

Figure 4: Difference in Proportions of Men Hired 1st and 2nd vs. Women Hired 1st and 2nd



Notes - This figure shows mean gender difference in proportions of men hired vs. women hired. Sub-figure (a) shows this gender difference for the candidates who are hired first and sub-figure (b) shows this difference for the candidates who are hired second excluding those hired first. The left panel in both figures show the first three rounds where there is a quota for women in effect under the temporary affirmative action treatment, and the right panel compares the two experimental treatments for the last three rounds when affirmative action is removed. Error bars correspond to 95% confidence intervals.

quotas are lifted in the last three rounds, men are 6 percentage points more likely to be hired than women, which although statistically different from 0 is significantly less than the control treatment.

Hiring choices of the 2nd candidate among the remaining 3 workers is shown in panel (b). We again find a substantial male advantage in the control treatment where women are 7 percentage points ($p = 0.001$) less likely to be hired than men in rounds 1 to 3. This male advantage increases to 9 percentage points ($p < 0.001$) in rounds 4 to 6. With the temporary affirmative action treatment on the other hand, women are 19 percentage points ($p < 0.001$) more likely to be hired while quotas are in effect and the gender difference disappears when affirmative action is lifted in rounds 4 to 6 ($p = 124$). Overall, we find positive and significant effect of the treatment for both 1st and 2nd hired candidates. In terms of representation of women, they are not only included as a 2nd candidate but also become the preferred

candidate for hiring where the later effect is in fact marginally stronger. These results are also confirmed in a regression analysis (see appendix table B5).

In terms of mechanism behind lasting improvement in hiring of women as 1st and 2nd candidates, we find similar heterogeneous dynamics as the overall hiring result (see appendix table B6). Baseline gender discrimination in hiring reduces once we control for beliefs about performance for the employers more biased and more likely to discriminate against women. And, the effect of the temporary affirmative action treatment is suggestively absorbed by beliefs for this subgroup of employers for both 1st and 2nd candidates for hiring and thereby pointing towards reduction in inaccurate statistical discrimination.

1.5 Discussion and Conclusions

In a setting where employers' beliefs and hiring choices are biased against women, this paper uses a series of experiments to examine if a temporary affirmative action policy can correct these biased beliefs and thereby lead to a lasting improvement in representation of women. While the data reveals that a temporary quota significantly improves representation of women even after the policy is lifted, the effect of this intervention is muted on reduction in gender bias in beliefs due to presence of some employers who are disregarding the information about employees. Further, in the starkest possible comparison with employers who are more biased and more likely to discriminate against women at the beginning of the experiment, we find a significant reduction in gender bias in beliefs as well as a lasting improvement in representation of women due to the temporary affirmative action treatment. Moreover, within this subgroup of employers the changes in hiring choices in favor of women are explained by beliefs thereby supporting the conjecture that increased exposure to women under a temporary quota provides opportunity for correction in biased beliefs which in turn improves the representation of women even when the policy is lifted.

Findings from this paper contribute to the broad literature on lasting effects of temporary affirmative action. Economic theory [111] and lab experiment [140] on long-term effects of temporary affirmative action have so far focused exclusively on studying accurate statistical

discrimination wherein employer beliefs are correct. This project is the first to examine the effect of a temporary affirmative action in a setting with biased beliefs. I do so by identifying and eliciting employers' responses over a real effort task where there is no gender differences in actual performance but women are believed to be worse performers than men. Ruling out the possibility of accurate statistical discrimination is also my motivation for using a controlled setting of an online experiment for this study.

Prior studies have suffered from an inability to determine whether the true performance of the over and underrepresented groups are indeed the same while only the beliefs are biased. This then becomes a confounding factor in identifying changes in beliefs about performance to drive any lasting changes in representation. For example, [43], in a field experiment find that quotas in the long run improve the likelihood of women contesting and winning political elections even after they are lifted. However they also find the survey measure of effectiveness of the leader to become worse after first exposure to a female leader but improve thereafter. Their argument is that the nature of underlying discrimination could have been accurately statistical with initial conditions conducive for less effective female leaders to invest in becoming more effective overtime while the belief assessments of the participants are correct throughout the experiment.[25] Similarly, [298] also find a lasting positive effect on representation of women in politics due to temporary quotas but could not identify learning about ability as the underlying mechanism. On the other hand, [49] find positive effect of an affirmative action policy to subside once the policy is lifted and argue the lower ability of the underrepresented group to be driving their results.

Many other notable studies also document longer term effects of temporary affirmative action in various contexts and find mixed results [142, 277, 18, 17, 136, 360, 213, 278, 39]. This paper contributes to this literature by identifying beliefs as the mechanism behind changes in hiring choices. While in this study we find evidence for existence of inaccurate statistical discrimination and that it can be remedied through exposure using a temporary quota, the hiring dynamics in the real-world can change depending upon the nature of the underlying discrimination. Future work can address how to distinctly identify inaccurate

[25]Through an additional experiment involving hypothetical leaders and vignettes they show malleability in beliefs and changes in attitude are possible to achieve through exposure to female leaders.

statistical discrimination in the field from other types of discrimination to inform policy predictions.

Another interesting result to emerge here is that among the employers who are less biased in their beliefs and are less likely to discriminate against women at the beginning of the experiment, we also find temporary affirmative action policy to produce a lasting improvement in representation of women. With the setting being that of biased beliefs this effect can be thought of as reduction in taste-based discrimination against women. A thorough study of identification of reduction in taste-based discrimination as well as mechanism behind this change in preferences has been left for future work.

In an experimental setting this paper documents the presence of inaccurate statistical discrimination in hiring against women and presents evidence that it can be remedied through a temporary affirmative action policy. From a policy perspective, results from this paper motivate the use of a temporary affirmative action quota to correct biased beliefs and break a pattern of self-perpetuating under-representation. Even the advocates of affirmative action policy believe it to only be a temporary fix for social inequalities [339] while we find that it is capable of creating fundamental changes in the society which can fix social inequalities even beyond the policy instance. Also reflecting on the unpopularity of affirmative action as a policy measure, it might become possible for people to be more accepting of quotas if they are only going to be temporary.[26] With a high chance that affirmative action will be banned in many parts of the world, this paper is a small step forward in understanding what to expect in a post affirmative action world.

[26]Gallup 2016 poll showed that 65% Americans disagree with Supreme Court's decision to allow race to be factor in college admissions.

2.0 On Gender Differences in Responsiveness to Experimental Conditions (Joint with Felipe A. Araujo and Lise Vesterlund)

2.1 Introduction

The experimental literature on gender differences in economic behaviors is large and growing. Researchers have used the laboratory to study gender differences in competition [190, 290, 186], negotiation outcomes [334, 253], risk and time preferences [145, 101], social preferences [15, 159, 78], among many other topics.[1] Over time, a common understanding has emerged in this literature, namely that women are more responsive than men to changes in experimental conditions. These are changes in the design and conduct of experiments that are either unrelated to the substantive research question or that are ancillary to it.

The exact origin of this proposition is uncertain, though early studies in psychology have provided support for it. The influential book by Carol Gilligan [187], for instance, argues that women are more sensitive to social cues in determining appropriate behavior, while other highly cited papers have suggested that women experience emotions more strongly than men [181, 261]. These results on gender-specific responsiveness seem to have become popular in experimental economics following the publication of a review paper in the Journal of Economic Literature [118]. This paper summarizes the experimental literature on gender differences in social preferences, risk preferences, and preferences for competition. Importantly, the authors suggest that women's greater responsiveness to changes in experimental conditions offers an organizing principle for the empirical findings in the literature. This comprehensive review has been very influential, accumulating over 6,600 Google Scholar citations as of this writing.

In this paper, we carry out an empirical test of the proposition that women are more sensitive to changes in experimental conditions. Although earlier studies have purported to test the hypothesis by conducting experiments under a limited number of different conditions (see Section 2.4 for details), the proposition that women are more responsive to experimental

[1] See [291] for a review of literature on gender differences in competition.

conditions in general, without qualifications and irrespective of the topic of study, cannot be tested on any one individual study. To the best of our knowledge, ours is the first comprehensive assessment of the female-sensitivity conjecture.

We do so by, first, conducting a detailed review of the experimental literature and, second, by complementing it with data from a recent study by DellaVigna and Pope [139] that allows for multiple direct comparisons of men and women's responses to changing conditions. Our review of the experimental literature focuses on identifying changes to experimental conditions and their corresponding gender-specific effects. It includes both published articles and recent working papers on topics such as competition, negotiation, social preferences, discrimination, non-promotable tasks, risk and time preferences, among others. The literature review was completed in the Summer of 2021 and comprises 317 papers.[2] As a complement to the literature review, we then analyze data from the experiments reported in [139]. The paper reports on real-effort experiments spanning sixteen treatments across five different conditions. Their main research question is about the stability and predictability of the ranking, in terms of subjects' efforts, of each of the sixteen treatments across conditions. Crucially for our purposes, the structure of the study allows for multiple pairwise comparisons of the gender-specific effect of changes in experimental conditions.

Our findings do not support the female-sensitivity hypothesis. In 86 of the 317 of papers in our literature review (27%), we find that both men and women are similarly responsive to changes in conditions. We moreover identify 112 papers that do contain significant gender differences in responsiveness. Of those, 57 report higher sensitivity of men, while the remaining 55 papers find higher sensitivity of women.[3] Our analysis of the experimental data in [139] conforms with this finding. Specifically, women are not more likely than men to significantly change their effort as a response to changes in the experimental conditions. If anything, men are slightly more responsive in this sample. We conclude that the gender specific sensitivity to changes is highly dependent on the particulars of the experimental design and topic of study.

Having failed to find supportive evidence for the hypothesis, we set out to understand

[2]See Appendix D for a description of the review process

[3]The remaining 119 papers either did not have variation on experimental conditions or did not report sufficient data for an unequivocal classification. See Section 2.2 and Appendix E for more details.

whether the dissemination of the female-sensitivity conjecture might have influenced researchers' interpretation of their own results. To do so, we restrict attention to the papers that specifically mention the hypothesis. At this stage, we also consider studies that did not meet the criteria to be included in our literature review, but that contain literal quotes from the review paper that first advanced the hypothesis [118]. We analyze how each paper evaluates the female-sensitivity claim and whether evidence is presented either in favor or against it. Our intention here is to shed light on the reinforcement dynamics of an influential academic thesis. That is, we ask if once a broad claim gets traction in the literature, subsequent results that do or do not conform with it receive different interpretations and emphasis.

We find that papers are much more likely to cite the female-sensitivity hypothesis when their own results conform with it. Among the 63 studies that either cite the hypothesis or contain literal quotes from [118], only 6 argue that their own evidence does not support it. This is in sharp contrast with our literature review, in which 72% of the studies that involve changes in the experimental conditions have results that are at odds with the hypothesis – i.e., either men are more sensitive than women or there are no gender differences. A casual reader of the literature is thus likely to encounter many more statements of agreement with the female-sensitivity conjecture, even though a careful assessment of the evidence does not support it.

The remainder of the paper is organized as follows. Section 2.2 gives a precise definition of experimental conditions, while our literature review is presented in section 2.3. Section 2.4 presents the results for our analysis of the experimental data in [139]. And finally, section 2.5 discusses the influence of the hypothesis on subsequent study's interpretation of results and concludes.

2.2 Defining Experimenta Conditions

Before presenting the evidence from our literature review and data analysis, we need a clear understanding of what exactly experimental conditions are. Specifically, we need a

definition that is able to distinguish between substantive treatment variations, which are not the subject of the female-sensitivity hypothesis, from lesser changes in the design or implementation of the experiment. For simplicity, we will adopt the same definition as in [118], which highlight two types of settings in which women are hypothesized to be more responsive to than men: experimental context and social cues.

Experimental context refers to features of the experimental design that could vary while the main treatment of interest remains unchanged. Examples are face-to-face versus computerized interaction and strategy versus game method elicitation. Social cues, on the other hand, are related to the experiments' incentives, monetary or otherwise, and to other more meaningful design choices. Examples include the "size of payoffs, price of altruism, or the repetition of the game, and psychological variables like the amount of anonymity between the participant and the experimenter, and the way the situation is described" – [118], p. 463. Since the hypothesized gender difference goes in the same direction – women are more responsive – in both cases, we combine both into a single component, which we refer to as experimental conditions.

It's important to note that even this definition still leaves room for ambiguity. In some studies, changes in the design are readily and unambiguously identified as changes in experimental conditions. In other cases, however, different researchers could reasonably disagree as to if a given change should be interpreted as an experimental condition or as a substantive treatment variation.

Consider, for example, [25]. This study finds that increased monetary rewards can negatively impact performance for tasks requiring concentration and creative thinking. And it also examines the effect of increased social incentives on performance. In one condition, subjects were tasked with solving anagrams either privately or in front of a 10-person group. This change from private to public effort falls squarely into our definition of experimental conditions (degree of anonymity). The paper by Gneezy, Niederle, and Rustichini [190] on the other hand, is an example of a harder-to-classify study. They study gender differences in performance of participants across various competitive environments. The treatments varied the structure of the competitive incentives in that the benchmark case offered a piece rate payment, tournament case paid the highest scorer among a group of 3 men and 3 women, and

in final treatment one person was selected from the group for payment at random. Since this involves significant variation in the treatments and not a simple variation in the conditions, we classify this paper as "NA" for our current consideration.

Considering any residual ambiguity in the classification, our literature review includes a robustness exercise in which we examine the gender-specific responsiveness for every treatment variation in a study, as opposed to focusing on those changes we identify as being over experimental conditions. We next turn to our literature review proper.

2.3 Literature Review

Our literature review focuses on experimental studies that explicitly investigate gender differences, but also includes papers that report gender-specific results in the context of a different research question. We identified 317 papers and classified each one with respect to the female-sensitivity hypothesis.[4] Specifically, we indicate whether it presents evidence that (a) women are more responsive ($Women > Men$), (b) men are more responsive ($Men > Women$), or (c) men and women are equally responsive ($Men \sim Women$). We also classify papers that did not contain any variation in experimental conditions (NA) or for which, although variation was present, the data reported did not allow for a classification (a)-(c) (*Insufficient Information*).[5,6]

Table 7 presents the results of our literature review. It provides a count of papers classified in each category. The first two columns of Table 1 restrict attention to studies that contain changes in experimental conditions as defined earlier. As can be seen from columns (1) and (2), 17.3 percent of the papers included in our literature review substantiate the assertion that women are more sensitive than men to changes in experimental conditions, whereas 17.9 percent of the papers we review illustrate the opposite effect, namely that men

[4]See Appendix C for the complete list of papers.

[5]A complete version of the literature review, including a description of the design and main results, is available from the authors upon request.

[6]The Insufficient Information classification in addition to including cases where information is insufficient to do this classification also includes cases in which only the difference-in-difference (DiD) coefficient is reported. A significant change in the DiD coefficient could be the result of either men or women (or both) responding significantly to the experimental conditions.

Table 7: Classification of Papers According to Gender Differences in Responsiveness to Changes in Experimental Conditions

	Experimental Conditions Only		All Treatment Variations	
	N	% Total	N	% Total
	(1)	(2)	(3)	(4)
Men > Women	57	17.9%	61	19.2%
Women > Men	55	17.3%	60	18.9%
Men ∼ Women	86	27.1%	100	31.5%
NA	96	30.3%	64	20.2%
Insufficient Information	23	7.4% 32	10.2%	
Total	317	100%	317	100%

Notes - Notes: Men > Women (Women > Men): data shows men (women) being more responsive to changes than women (men). Women ∼ Men: data shows men and women being equally responsive to changes. NA: there is no change in experimental conditions. Insufficient Information: can't infer gender differences based on published results alone.

are more sensitive to the experimental conditions. Almost 27 percent of the studies revealed that both men and women were equally responsive.

Consider the paper by Andreoni and Vesterlund [15] where we find men to be more responsive than women (Men > Women). They study participants' response to variation in price of giving in a modified dictator game and find that the price elasticity of demand (for giving) for men to be higher than that of women. In the paper by Heinz, Juranek, and Rau [209] on the other hand, we find women to be more responsive to changes in experimental conditions than men. They use a lab experiment to study gender differences in reciprocity in a dictator game where they vary how the endowment is obtained between a windfall lottery win and the recipient's performance on a real effort task determining the size of the endowment. They find that average taking rate of women varies significantly between the two treatments (74.02% in windfall and 63.30% in real effort) while it remains fairly stable among men (73.81% in windfall and 75.39% in real effort).

[31], using an experiment study gender differences in volunteering for low-promotability tasks and we find both men and women to be responsive to changes in the gender composition of the groups they play the game with. Within each group, subjects are given a chance to complete a task that benefits the entire group, but that is relatively more costly to the group member who actually completes the task. In mixed gender groups, they find that women are significantly more likely to volunteer to complete the task. But going into single gender groups, men increase volunteering by the same rate as women decrease volunteering. A point to note here is that these papers do not necessarily focus on this result as an organizing principle. The classification is based on our own assessment of the drivers of gender differences using the results presented in the paper.

Columns (3) and (4) of Table 7 detail the same classification, but now considering every treatment change reported in the paper. We view this as a robustness exercise considering the difficulty in classifying some changes in the experiment as either substantial treatment variations or simple changes in experimental conditions. In the latter two columns of Table 7, the only situation in which a paper is classified as "NA" is if there are no treatment variations at all – i.e., a paper that elicits risk preferences by gender and socioeconomic status.

Approximately one third of the studies classified as "NA" in columns (1) and (2) are now classified in columns (3) and (4). Most of those papers (24 out of 32) providence evidence of men and women being equally responsive to changes. And overall, the classification patters do not change significantly when expanding the sample in this way, with 19.2% of the papers reporting results in line with the hypothesis and 18.9% with results opposed to it. As before, a large fraction of studies (31.5%) has results indicating a similar responsiveness for men and women.

The evidence is from our comprehensive literature review is clear. A simple count of paper whose evidence is in favor or against it does not provide support for the female-sensitivity hypothesis. Even considering the (very real) possibility that we either overlooked relevant papers, or incorrectly classified some of the results, it is arguably extremely unlikely that revised numbers would change as much as necessary to provide support to the hypothesis.

2.4 Analysis of data from DellaVigna and Pope 2022

The female-sensitivity hypothesis is described in [118] as an organizing principle to aid in the interpretation of the empirical findings in the literature. As stated, the hypothesis would encompass gender-difference results in areas as disparate as preferences for competition, charitable giving, risk preference, or excuse-seeking behavior. Although the conjecture is clearly too broad for any one experimental design to convincingly refute or support, data reported in [139] provide an ideal petri dish in which to partially assess the hypothesis. Specifically, it allows for multiple pairwise comparisons of men and women's responses to changes in experimental conditions within the same experimental framework and using the same measure.

The paper is concerned primarily with understanding the stability and predictability of results in behavioral economics. It analyzes data from real-effort, online experiments with a total of more than 18,000 participants. Each subject performs an experimental task (either repeatedly typing the letters "a" and "b" or coding World War II conscription cards) under one of sixteen different behavioral treatments across five conditions (design changes). The outcome of interest is the ranking of the sixteen treatments in each of the five conditions with respect to subject's average effort, which is measured via rank order correlations across the design changes. Additionally, the authors survey academic specialists, graduate students, and participants in online experiments to investigate how predictable are the effects of changes in the experimental design.

The sixteen behavioral treatments are meant to capture several previously studied ideas to motivate effort. For example, four of the treatments involve varying the type of payment (fixed payment versus piece rate) as well as varying the piece rate amounts, from very low to very high. Other treatments adopt either a time preference incentive (payment in two weeks versus in four weeks), a probabilistic incentive (50% chance of a small piece rate versus 1% chance of a very high piece rate), or a charitable giving incentive (low versus high piece rate for the Red Cross). Finally, there are also several treatments over purely psychological incentives. For example, providing social comparisons, relative rankings, or by highlighting the task significance. The changes in the design, on the other hand, are meant

to reproduce modifications in the study protocol that, while often observed in practice, are not directly related to the main research question. These encompass the type of task used, the output measure, the presence or not of a consent form, the demographic characteristics of the participants, and the geographic and cultural background – i.e., participants from India versus the US.

Crucially for our purposes, both variation in their main treatments and in the design give rise to variation in experimental conditions as defined here. For example, by comparing men's and women's change in effort between the 1-cent and the 4-cent piece rate, or men's and women's effort changes across different types if psychological effort motivators. One could also consider comparing within treatments and between task types (typing keys versus coding WWII conscription cards), though the difficulty here is that the tasks have different outcomes and, potentially, different levels of noise, which would muddle the inference on gender differences in effort response.

Unfortunately, data in DellaVigna and Pope (2022) does not contain a gender identifier for every condition, so we are restricted to comparisons between the 16 main treatments within two of the conditions (typing task and WWI cards). We conduct pairwise comparisons of the effort change of male and female participants within treatments of the same type (i.e., piece rates, charitable giving, time preference, etc.), and separately for the typing and WWII-cards tasks.[7] Our final sample comprises 56 unique tests of differences in mean effort for both men and women.

For an example of a pairwise comparison consistent with the female-sensitivity hypothesis, consider the task of coding WWII cards under the treatments of no piece rate and high piece rate (5 cents per coded card). Women exerted significantly more effort on the high piece rate treatment compared to the treatment with no piece rate. Men, on the other hand, coded more WWII cards under no piece rate compared to high piece rate, though the difference is not significant.[8] This is a clear example where women responded more strongly than men to a change in experimental conditions.

[7] See Appendix E for a description of all the 56 pairwise comparisons.

[8] Women coded on average 55.2 cards in the treatment without piece rate and 62.1 cards in the treatment with high piece rate ($p = 0.038$, two-sided t-test). Men coded 51.7 and 48.5 cards, respectively, for treatments with no piece rate and with high piece rates ($p = 0.475$, two-sided t-test).

Let us now consider the difference in average effort in the WWII-cards task between the treatment with a high piece rate and the treatment with a very low piece rate (1 cent per 20 coded cards). Men's average effort in the high piece rate treatment was significantly lower than the average effort in the very low piece rate treatment. For female participants, average effort was similar under both conditions, indicating that men are more responsive to this change in experimental conditions.[9]

Finally, there are several cases for which men and women are equally responsive to changes. For example, both men and women's average effort in the a/b typing task increase significantly from the treatment with no piece rate to the treatment with a very low piece rate (1 cent per 1,000 points).[10] That is, both female and male participants are equally responsive to this specific change in experimental conditions.

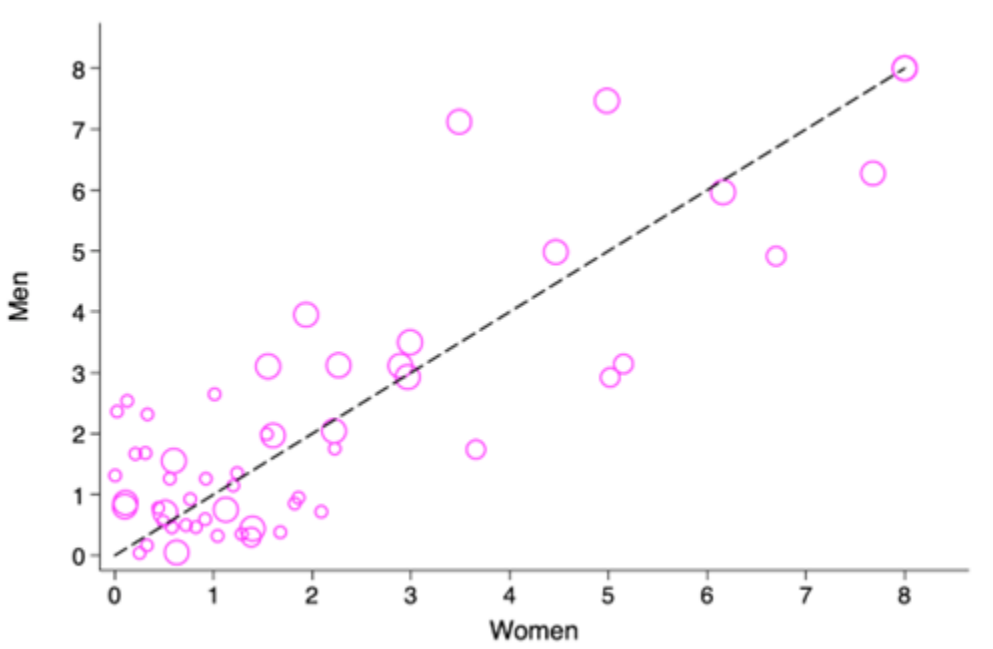
Figure 5 plots the absolute value of the test statistic of a t-test on mean effort for each of the 56 comparisons for men (y-axis) and women (x-axis). The sizes of the dots are proportional to the total sample size for each pairwise comparison. If women were indeed more sensitive to changes in experimental conditions, we would expect most of the points to lie below the 45-degree line. As Figure 5 shows, however, the points are close to evenly distributed in the graph, with a concentration of data points close to the 45-degree line.

As is becoming typical in laboratory and online experiments, women are oversampled in the [139] experiments. As such, the distribution of the test statistics could be potentially misleading. To account for that, we also compute the effect sized using the Cohen-d measure, which we report on Figure 2. As with Figure 1, we do not observe a concentration of data points below the 45-degree line, which is what we would expect if the female-sensitivity hypothesis was true. In fact, most of the points in Figure 2 lie below the 45-degree line, indicating a stronger effort response to changes in experimental conditions for men rather than women. The average effect size of changes in experimental conditions for male participants is 0.275; it is 0.223 for female participants.

[9]Men coded on average 59.3 cards in the treatment with a very low piece rate and 48.5 cards in the treatment with high piece rate ($p = 0.001$, two-sided t-test). Women coded 62.1 and 62.5 cards, respectively, for treatments with a very low piece rate and with high piece rates ($p = 0.901$, two-sided t-test).

[10]Men scored on average 1451.3 points (correct types) in the treatment with no piece rate and 1856.9 points in the treatment with a very low piece rate ($p = 0.000$, two-sided t-test). Women averaged 1519.5 and 1872.8 points, respectively, for treatments with no piece rate and with very low piece rates ($p = 0.000$, two-sided t-test).

Figure 5: Scatter Plot of T-values for Test of Mean Effort Across Experimental Conditions for Both Men and Women

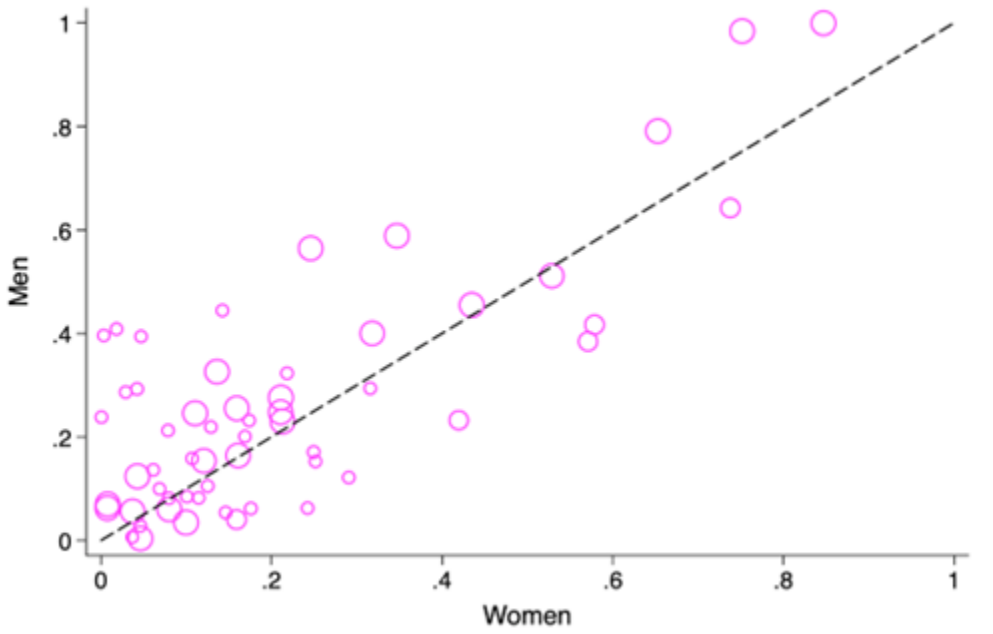


Notes - Notes: Each dot corresponds to a pairwise comparison of mean effort across a change in an experimental condition. The x-axis contains the absolute value of the test statistic for a t-test of the female subjects and the y-axis contains the absolute value of the test statistic for a t-test of the male subjects. For ease of visualization, we top-coded the values at 8 for case in which both men and women's test statistics was greater than 8, which was the case in three pairwise comparisons. See Figure D1 in Online Appendix D for a graph without the top-coding. If women were more sensitive to changes in experimental conditions, we would expect most points to lie below the 45-degree line.

2.5 Discussion and Conclusion

In our final discussion, we turn to the question of how the dissemination of the female-sensitivity hypothesis might have influenced researchers' interpretation and emphasis of gender difference results. We do this by reviewing a second set of papers, which we arrive at in two steps. First, among the papers included in the main literature review, we restrict attention to those that specifically mention the hypothesis as presented in the review paper by Croson and Gneezy [118]. Second, we add studies that do not necessarily investigate gender differences, but that contain literal quotes related to the hypothesis as described in

Figure 6: Scatter Plot of Effect Sizes – Cohen's D – on Effort Across Experimental Conditions for Both Men and Women



Notes - Notes: Each dot corresponds to a pairwise comparison of mean effort across a change in an experimental condition. The x-axis contains the absolute value of the Cohen's-D effect size measure for female subjects and the y-axis contains the absolute value of the of the Cohen's-D effect size measure for male subjects. If women were indeed more sensitive to changes in experimental conditions, we would expect most points to lie below the 45-degree line.

[118]. did not meet the criteria for the main literature review but that contain literal quotes related to the hypothesis.[11] Our goal is to provide a broader view of the impact of the gender-responsiveness hypothesis on researchers' interpretation of results.

For each paper, we classified it with regards to how it portrays the gender-responsiveness result. A study is classified as neutral if it simply acknowledges the existence of the hypothesis, without taking a stand or assuming it as a scientific fact. Alternatively, we classify it as in favor if it either accepts it as a fact or argues that their own evidence concurs with the stated conclusions. Finally, we classify the paper as against if it either presents the hypothesis while casting doubt about its validity or argues that their own results provide

[11] As such, these studies are outside of the scope of our literature review. Indeed, most are non-experimental papers.

Table 8: Are Papers that Directly Mention the Female-sensitivity Hypothesis More Likely to Agree with It?

	Literature Review Sample		Others (Literal Quotes	
	N	% Total	N	% Total
	(1)	(2)	(3)	(4)
In Favor	23	60.5%	11	44%
Neutral	13	34.2%	10	40%
Against 2	5.3%	4	16%	
Total	38	100%	25	100%

Notes - Notes: In Favor: studies which either accept it as a fact or argue that the paper's evidence concurs with the conclusions stated in CG. Neutral: studies which simply acknowledge the existence of the hypothesis, without taking a stand or assuming it as a known fact. Against: studies which present the hypothesis while casting doubt about its validity or argue that the paper provides direct evidence against the gender-responsiveness hypothesis.

direct evidence to the contrary.

We identified 63 studies that directly mention the gender-responsiveness hypothesis. Of those, 38 were included in our original literature search. We summarize our findings in Table 8. As can be seen from Columns (1) and (3) of Table 8, and in sharp contrast with the results from the main literature review, most of the citing papers agree with the female-sensitivity hypothesis, namely that women are more responsive to changes in experimental conditions.

As alluded to before, the gender-differences literature also contains studies that purport to directly test the sensitivity results of CG. [164] test the claim that women experience emotions more strongly than men using online surveys and choices over hypothetical lotteries. Their results support the hypotheses in that women's emotions about outcomes are stronger than men's. In [134] participants are randomly paired and receive information about each other's gender via a pseudonym. Each participant then simultaneously chooses a piece-rate or tournament scheme for a maze-solving task. The results show that women increase their rate of tournament entry compared to baseline, whereas men don't. Based on this result,

the authors conclude that "as reported by Croson and Gneezy [118], women would be more sensitive to the social context than men". The other studies are [67] – men more sensitive –, [277] – women more sensitive –, and [263] – women more sensitive. We should note, however, that no single paper will be able to settle the question, as the stated hypothesis argues that women are more sensitive without qualifications.

Why has the literature, for the most part, relied on those organizing principles when interpreting experimental results? One possibility is that once a result gets traction in the literature, it tends to get mainly positive reinforcement from other studies. If while studying an unrelated question a researcher obtains a result that agrees with a well-known hypothesis, it will get nominally cited and reinforced. On the other hand, if evidence is found that contradicts a well-known claim, it will tend to not be highlighted as such. To negate a well-regarded result or statement, one would need considerably more evidence (for example, like we are doing here) than an ancillary result from an otherwise unrelated work. However, a comparable ancillary result that coincides with the hypothesis might be more easily cited, as it wouldn't be as heavily scrutinized.

We intend this paper to be read as a motivation for more research into this topic. Future work should investigate which domains and types of experiments give rise to gender differences in responsiveness to experimental conditions – and in which direction.

3.0 The Experimenters' Dilemma: Inferential Preferences over Populations (Joint with Luca Rigotti and Alistair Wilson)

We compare five populations commonly used in experiments by economists and other social scientists: undergraduate students at a physical location (lab), undergraduate students in a virtual setting (v-lab), Amazon's Mechanical Turk (MTurk), CloudResearch approved list (Cloud-R) of MTurk workers, and Prolific. The comparison is made along three dimensions: the noise in the data due to inattention, the cost per observation, and the elasticity of response. We draw samples from each population, examining decisions in four one-shot games with varying tensions between the individual and socially efficient choices. When there is no tension, where individual and pro-social incentives coincide, noisy behavior accounts for 55% of the observations on MTurk, 17% on Prolific, 16% on Cloud-R, and 14% and 22% for the lab and v-lab respectively. Taking costs into account, if noisy data is the only concern Prolific dominates the lab from an inferential power point of view, combining relatively low noise with a cost per observation one fifth of the lab's. However, because the lab population is more sensitive to treatment, across our main PD game comparison the lab still outperforms both Prolific and MTurk.

3.1 Introduction

Economic experiments have become a key tool for uncovering facets of economic decision-making that would be veiled in naturally occurring data. Over the previous half-century, the dominant paradigm was the laboratory experiment: a set of typically undergraduate participants are recruited to a fixed time-slot at a physical location, where a tailored set of monetary incentives are then offered to examine and identify the economic hypothesis. But in the last decade,, a number of online populations have emerged for conducting economic experiments. These online populations offer an array of positives: greater convenience, lower barriers to entry, large number of participants and greater representativeness. Moreover, for

researchers with finite budgets they offer another boon: a typically much-lower costs per observation than for the equivalent lab studies. However, not all of these benefits come for free, and one concern that is often raised is of the experimenter having reduced control over the participants on these online populations. As participants recruited from online populations will take part in the study on their own devices, and often at their own pace, their divided attention and an incentive to complete the study as quickly as possible can potentially leading to noisier data. In contrast, in laboratory studies distractions and the the timing of the study can typically be controlled, though normally with the greater time commitments and required focus having higher participant payments.

Prompted by some of these trade-offs across experimental populations, in this paper we try to assess the inferential bang for your buck obtained, running a horse-race across five different populations using a common task. The exercise starts out with a simple motivating idea: an experimenter is attempting to measure a qualitative treatment-effect, to be identified through a difference in means. However, the experimenter has a fixed budget of $\$Y$ to spend. While recruiting from the lab sample might have low noise, the costs per observation will be expensive, and so she will get a smaller sample size. In contrast, while an online population might have noisier responses washing out some of the treatment effects, the lower costs per observation might mean that she can have a much larger sample. By assessing our five populations over the inferential power to detect the treatment effect, and scaling the per participant payments to those representative of the population, we try to examine these tradeoffs directly. Our results can therefore help experimenters identify the merits of each population through a clear inferential lens.

Our first population is a standard laboratory study, with undergraduate students recruited to come to a physical lab. Our second populations also recruits from undergraduate university student however they participate in the experiment online. Our other three populations use online participants: MTurk, likely the most ubiquitous online labor populations; Cloud-R, a subset of MTurk workers selected through various measures of attention and data quality by the platform; and Prolific, an emerging platform with a more curated set of participants.[1]

[1]The exact procedure for selecting the approved list of participants is not disclosed by CloudResearch

Using a set of simple two-player strategic choices we compare the five populations over cost per observation and noise in decision-making - considering noise as the proportion of decision-makers that act independently of incentives. We measure noise in two games where participants lack any effective strategic tension: self-interested behavior is entirely coincident with pro-social behavior. Taking as given that participants have underlying preferences that are increasing in both their own and the total payoffs, decisions against both self-interest *and* efficiency identify inattentive participants.[2]

We also compare populations across two games that embed an hypothesis with extensive prior evidence. We use this comparison to refine our question: given a fixed budget, which population is preferred by a researcher interested in making a qualitative directional inference? Here we use a canonical trade-off between efficiency and own-payoff, with both games having a prisoner's dilemma (PD) structure. Reflecting a literature that emphasizes partial cooperation, engaging cooperatively so long as the trade-offs between own-payoff and social efficiency are not too large, our chosen games vary the relative temptation to defect. We use the prior literature to form expectations over the size of the treatment effect, and thus induce the experimenter's preference over populations via the inferential power in a test of the differences in play across the two games.

Summarizing, we parameterize each experimental population with two properties: (i) the cost per independent observation; and (ii) the size of the attenuation over the treatment effect. We then form the experimenter's preference as if setting up a standard choice problem: drawing both iso-power contours under a fixed budget (analogous to the indirect utility) and the dual iso-budget contours under a fixed power level (analogous to the expenditure function). The attention effect is further decomposed between the effect due to inattention and inelasticity of response across the two games.

Our results confirm that the problem is not trivial. 55 percent of the MTurk participants make decisions that are entirely independent of the induced incentive (either through random choice, or choosing the first-available choice), but their average observation costs

but their claim on the publicly available website claims a strict vetting criteria while keeping their list to be demographically representative of the overall MTurk population.

[2]Moreover, a separate source of variation in our experiments manipulates the first-listed option, thus creating a frame orthogonal to incentives. As such, we can decompose the extent to which participants in each population choose randomly across decisions, or simply choose the first-listed option available to them.

is \$3.01. A the other extreme, the average observation costs in the physical laboratory is \$22.08, but noisy behavior is only 14 percent. Holding constant the experiment's budget, the inferential power from a low-noise sample of 75 laboratory participants must be compared to the inferential power from a high-noise sample of 550 participants on MTurk (or 540 and 380 low-noise participants on Cloud-R and Prolific respectively). Were attenuation in treatment-effects purely driven by the inattentive proportion we find that Cloud-R and Prolific are the clearly dominant populations, followed by MTurk, and then closely together lab and v-lab samples as the least effective.

But the proportion of inattentive/noisy participants is not the entire story. The two online populations on unrestricted MTurk and Prolific also exhibit reduced elasticity of response: a smaller effect-size under the same induced treatment. Factoring in this reduced quantitative size of the response, MTurk and Prolific populations end up having diminished inferential power relative to the laboratory. For example, despite low cost per observation and relatively high attentiveness on Prolific, this sample is just too inelastic with a near negligible response to a shift in the PD tensions. What the literature led us to think of as a moderately sized exogenous treatment for the lab ends up being far too subtle for these online samples. On the other hand, the Cloud-R sample maintains its dominance over the lab samples both terms of noise as well as responsiveness to the treatment where the size of the treatment effect is close to the prediction.

Some of our results are likely specific to social dilemmas, and the potentially higher level of attention required in strategic settings. However, the environments we examine are simple enough that the very noisy data from MTurk points to that population as being dominated by more-curated online populations such as Cloud-R and Prolific. Despite Prolific being almost 50 percent more expensive per observation, the additional signal more than compensates for this. Particularly for very stark economic comparisons, the ability to collect many, many observations favors online populations such as Cloud-R and Prolific which consistently dominate the lab. However, our study also points to a potential benefit of laboratory samples. Despite the expense, for more-nuanced hypotheses or more-complicated environments, lab samples may well be preferable given their ability to consistently retrieve sensible outcomes.

The rest of the paper is organized as follows: Section 3.2 summarizes the related literature

and highlights our contributions, section 3.3 discusses our experiment design, incentives, implementation, and lays out the qualitative comparative hypothesis across all five samples. Section 3.4 presents results for the hypothesis while section 3.5 compares the five populations from the lens of inferential power and 3.6 concludes.

3.2 Related Literature

A number of studies have compared MTurk and laboratory population, primarily focused on whether empirical regularities observe in the lab can be replicated. Our paper novelty is in both adding three more populations to this comparison (Virtual Lab, CloudResearch Approved it, and Prolific) but also in making the focus more explicitly on the effective power on each population, taking into account researchers' financial constraints. In particular, we consider how the possibilities for many more independent observations from cheaper online populations interaction with the potential for noisier data, or a more inelastic response.

One of the earliest works examining the use of MTurk in online experiments is [299], replicating three classical behavioral economics results (the Asian disease, Linda and Physician problems), and finding no significant differences between the populations. Along similar lines, [218] find no significant differences in cooperation between an MTurk sample and the experimental lab literature on one-shot PD games. In [196] an MTurk sample replicates standard decision-making biases.[3] More recently, [347] suggests that strict exclusion criterion for "problematic" participants can reduce statistical noise without introducing sampling bias. In [24] the researchers uncover the same basic behavioral patterns of cooperation and punishment in a repeated public good experiment in both the lab and MTurk, even though dropout can be a challenge for conducting interactive experiments on MTurk.

[335] elicit and compare a battery of behavioral attributes using a survey administered to an entire undergraduate cohort, a self-selected lab sample, and a representative sample of US online participants recruited from MTurk. While they look at many different behav-

[3]MTurk participants exhibit: (i) present bias; (ii) risk-aversion for gains and risk-seeking for losses; (iii) show delay/expedite asymmetries; and (iv) show the certainty effect.

iors, their elicitation does includes two one-shot PD games (though with the same effective incentives). Similar to our findings, they do find significant differences in cooperation levels across populations for the PD game, with the online sample being more cooperative, however they do find comparable comparative statics across populations across many other behaviors. Their other overarching results are that behavioral characteristics are similarly correlated across populations and that noise (as measured by difference in response for duplicate elicitations) is higher for online populations. We confirm this last result when it comes to MTurk but not for Cloud-R or Prolific, though our own measures of noise are based on responses to a more-basic check of rationality and a response to a frame change. Our focus is also switched, where we take as given that the effect, and instead focus on the effective power of the population under a fixed researcher budget.

Our findings match growing concerns over a potential decline in the quality of MTurk data over the past two years. The literature highlights limitations of MTurk including, but not limited to, anticipation of deception by researchers, repeated participation in similar tasks leading to knowledge acquisition and a resultant change in behavior, unmeasurable attrition and programmed bots [205, 108]. [27] identify MTurk workers as being more likely to fail attention checks designed to measure haste and carelessness in responses than college students (though our measures of noise are on a sample that has successfully passed an understanding quiz). Some of our results echo this, and make this more concrete through our focus on inferential power. While we do find low inferential power for data from both MTurk and Prolific, our noise measures point to Prolific as being close to the laboratory levels. Instead, the low-power on Prolific seems to come about through an inelastic response, where the population in general offers much more promise. CloudResearch approved list of participants on the other hand circumvent both these concerns where we find low noise as well as elastic responses to the treatment variation.

3.3 Experiment Design

Our experiment has core $5 \times 4 \times 2$ design over:

Population We use five populations: (i) students recruited from the University of Pittsburgh's undergraduate population (the Lab sample); (ii) again students recruited from the University of Pittsburgh undergraduate population but now in a virtual lab setting where the entire experiment was conducted online (the V-Lab sample); (iii) online workers recruited from Amazon's online marketplace Mechanical Turk (the MTurk sample); (iv) Cloud Research approved online workers recruited from Amazon's Mechanical Turk (the Cloud-R sample); (v) online workers recruited from Prolific (the Prolific sample).[4]

Strategic environment We ask participants to make a binary choice across four symmetric two-player games (with payoffs provided in Table 9. While our experiment uses an A/B action labeling, we use a C(ooperate)/D(efect) labeling in the paper as all four games have joint cooperation as the socially efficient outcome.

Irrelevant Frame The frame variable changes the order in which the actions are presented to subjects, permuting the ordering of the C and D actions in the presented game tables.

3.3.1 Incentives and Implementation

Our design collects data across 10 between-subject treatments, the five populations and the frame-change over the ordering of the cooperate/defect decision. Each participant is asked to submit their choice between the two actions (A or B) in all four games presented to them in a random order. We present games to participants as a table with four rows (one for each possible action profiles) ordered as AA, AB, BA, BB for the self/other action. The re-framing therefore moves the socially efficient CC entry from the top entry in the table (labeled AA in the experiment) to the bottom entry (labeled BB).

For our horse-race between the five populations, our initial plans were for a budget of \$1,500 per population. However, our lab study was run first, and ended up being more expensive at just over \$1,600. We therefore match all other samples to this approximate budget. Within this population budget, we then spend the money across the C-first/D-first frames at a two-to-one ratio (in case pooling the samples was not an option for the lab

[4]CloudResearch was formerly known as TurkPrime, providing tools for online study recruitment. The M-Turk sample is also collected using CloudRearch while targeting the broader M-Turk worker population and not limiting to the "approved list".

Table 9: Experiment Design

Panel A	Payoff π_i on action (a_i, a_j)			
	(C, C)	(C, D)	(D, C)	(D, D)
Game PD1	\$21	\$2	\$28	\$8
Game PD2	\$19	\$8	\$22	\$9
Game Σ-DOM1	\$17	\$12	\$16	\$10
Game Σ-DOM2	\$15	\$16	\$10	\$11

Panel B	Participants & Expenditure				
	Lab	V-Lab	MTurk	Cloud-R	Prolific
Participants:					
C-first frame	50	50	368	374	250
D-first frame	24	24	180	167	135
Total	74	74	548	541	385
Expenditure:					
Total	\$1,634.00	\$1,609.30	\$1,647.32	\$1,746.90	\$1,679.76
Per observation	\$22.08	\$21.75	\$3.01	\$3.23	\$4.36

Notes - Participant numbers exclude those who failed to answer comprehension questions correctly. However, total expenditure includes fixed-payments made to participants who are dismissed on account of over-booking of sessions for the university samples as well as to those dismissed from the online studies for answering the comprehension question incorrectly.

sample).[5]

In our lab sample consists of undergraduate students recruited at the University of Pittsburgh. Participants are offered a \$6 fixed fee, and are randomly paid for one decision over the four games after being matched to an anonymous partner.[6] Payments for the selected game

[5]Focusing purely on the average earnings of the participants (so excluding fees and other costs) divided by the average time taken to complete each study, the effective wage rates are remarkably similar. Across the lab, V-Lab, MTurk, Cloud-R and Prolific the effective wage rates are \$31.66, \$31.13, \$31.81, \$38.27 , and \$40.25 per hour, respectively.

[6]The experiment is programmed in oTree ([106] and conducted at the Pittsburgh Experimental Economics Laboratory (PEEL).

are determined by payoffs shown in Table 9. In total our expenditure for the 74 laboratory observations was \$1,624, where this figure includes \$72 spent on show-up fees for unused participants.[7] This sample was collected right before the pandemic hit, and the sessions were run in person.

The virtual lab experiment follows the in-person lab experiment very closely, where participants are again recruited from University of Pittsburgh undergraduate population. This experiment is however conducted entirely online and follows online protocols mirroring the in-person protocols [126]. The total expenditure to collect 74 observations was \$1,609.30 with the per observation cost being \$21.73.12.[8]

The per-observation cost for the university student samples are about \$22. While we could have offered these incentives to the online participants of MTurk, Clour-R, and Prolific, this would have represented a substantial break from the norm on these platforms. As our aim is to match the effective incentives being offered on each population and to account for this in inference, we scale down the incentives for our online populations substantially.

Participants in the MTurk sample are given a \$0.50 fixed fee and a further \$0.50 if they correctly answer a comprehension question to show that they understand the instructions.[9,10] While payments within each game table exactly matches the lab sample, as given in Table 9, payments are scaled down in the likelihood of payment. Pairs of participant are paid for their decisions from one of the four game tables with a 10 percent probability.[11] In total

[7]Our methodology here is to include all variable costs for a study incurred by the researcher. One possible critique here is that we do not account for the financial costs of setting up and running the PEEL lab, where our approach is to treat these as sunk costs. As such, inferential comparisons across populations are from the point of view of a researcher who has free access to a turn-key lab space.

[8]Total expenditure from virtual lab (V-Lab) includes \$1.30 - the cost of deploying the experiment using Heroku. The cost also includes show-up fee of \$6 paid to all those participants who showed to an already full session of the experiment.

[9]Subjects who failed to answer the comprehension questions are not asked about game decisions and are excluded from out count of N and analysis. However the \$0.50 costs for these subjects as well as the fees for Amazon (20%) and Cloud Research [**?**, 4% of fixed fee]]litman2017turkprime are included in the total expenditure.

[10]The MTurk experiment is coded using Qualtrics and participants are recruited using Cloud Research where participants are restricted to a standard subsample: those located within the US, with a 95% or better approval rate.

[11]Our instructions give participants a clear rule used to conduct randomizations, where all draws are made using public randomization outside of the researchers' control (here public state lottery drawn the evening after the decisions are recorded). Moreover, participants are told that if selected for payment they would be matched to another payment participant, where the final bonus-payment would be determined by the choices of both payment participants' choices. As such, conditional on payment, the externalities and lottery

our MTurk sample contains data from 548 individuals with a total cost of $1,649 ($3 per participant).

Next, our Prolific sample follows an identical process for the marginal incentives and game payment as the MTurk sample. However, rules for the platform requires a larger minimum payment, so we increased the fixed payment to $1.60.[12,13,14] The total expenditure on Prolific was $1,680 for 385 observations. The Cloud-R experiment on the other hand was same as the MTurk experiment in incentives as well as implementation, where the only difference is in terms of the populations. Instead of opening the experiment to the entire pool of MTurk workers we imposed a recruitment restriction of only allowing the approved list of MTurk workers as identified by the Cloud Research platform. The total cost for the Cloud-R sample of 541 participants came out to be $1,746 with per observation cost of $3.23.[15]

Our overall plan embeds an essential question: given the differential costs for each observation, and the potential quality differences in the data collected, which population is superior? MTurk offers the potential for the largest number of observations given a fixed budget. However, it is also potentially the noisiest. On the other extreme, the laboratory is the most expensive per observation. The question is whether this additional expense is warranted through higher quality data or do vetted pools of online participants such as Prolific and the Cloud Research approved list reduce the noise and improve inferential power.

over the four games are identical to our lab study.

[12]Participants failing the comprehension check receive the fixed payment, but are not given the marginal incentive. The total expenditure includes the costs for these participants as well as the 33% fee imposed by the platform.

[13]We conducted a pilot of 20 participants on Prolific to understand the median time taken, where the minimum fixed-fee payment was a function of this time. However, the marginal incentives for this pilot were higher and the fixed fee lower). For the sake of comparability, neither this pilot data, nor its cost are considered in our analysis.

[14]The Prolific experiment was also programmed using Qualtrics and implemented directly on the Prolific platform

[15]While using the approved list feature on Cloud Research is free of cost, the cost of using the platform increased to 10% from the earlier 4% in the MTurk experiment. The cost of Cloud-R sample includes the payment to participants, Amazon fees of 20% levied on all participant payments and the 10% of the fixed fee paid to the participants collected by Cloud Research platform. The total cost here also includes the fixed fee paid to participants who incorrectly answered the comprehension question but excludes these participants from the total N

3.3.2 Hypothesis

We first outline the features of the games subjects play. All four games are dominance solvable in terms of individual payoffs (relabeling the standard notion of strict dominance):

Definition 1 (i-Dominated action)**.** *Action a is i-dominated if there exists another action a' that gives player i higher payoff for any action of the other players.*

The i-dominant action profile (also the Nash action profile) is to defect in games PD1 and PD2 and cooperate in Σ-DOM1 and Σ-DOM2. However, there is a large body of evidence suggesting that many individuals' preferences are other-regarding and sensitive to social efficiency. This evidence shows that many individuals choose i-dominated actions if these choices improve efficiency (as measured by the sum of payoffs). As such, a stronger version of dominance can be based on individual and total payoffs:

Definition 2 (Σ-Dominated action)**.** *Action a is Σ-dominated if there exist another action a' such that a is i-dominated by a', and the sum of the players' payoffs is smaller for a than for a' for any action of the other players.*

Games Σ-DOM1 and Σ-DOM2 are constructed so that the D action is Σ-dominated, and this action choice is thus hard to justify with almost any other-regarding preference.[16] Taking as given that participants from all populations are driven by a preference that is strictly increasing in both the own and social payoff, the only justification for Σ-dominated behavior is therefore that the agent does not fully understand the environment.

These two games therefore provide our first measure of attentiveness in each population, where our null hypothesis is therefore that:

Hypothesis 1 (Dominated-play null)**.** *The five populations have similarly small proportions of Σ-dominated play.*

Our second measure for the quality of choices across the five populations is based on the response to the framing variable. A change in the order in which actions are presented changes nothing with respect to the offered incentives. One plausible bias for inattentive

[16] Game Σ-DOM1 is designed to satisfy an even stronger ordering over unilateral deviations: the Pareto order. However, we do not find that this has any additional predictive content, so we focus purely on Σ-Dominance.

participants is that they move as quickly as possible through the offered choices by selecting the first-available option. A comparison of cooperation rates in the Σ-DOM game pair across the frame change therefore identifies this feature, where we would expect greater cooperation in our main treatments where C is the first listed option.

Hypothesis 2 (Reframing null)**.** *The five populations have the same cooperation rates across the re-framing.*

Our first two hypotheses are about assessing the degree of inattentiveness, using choices where we can say that one option—separate from preferences which may legitimately vary across populations option—is a priori dominated. Our final hypothesis is more nuanced, and instead concerns behavior between games PD1 and PD2, where we change the intensity of the prisoner's dilemma tensions.

To form our hypothesis we make use of a parametric index known as the Rapoport ratio [**?**, cf.]]rapoport1967note, which has been shown to be predictive of cooperation. The Rapoport ratio is given by a function of the PD-game payoffs:

$$\rho = \frac{\pi_i(C,C) - \pi_i(D,D)}{\pi_i(D,C) - \pi_i(C,D)}.$$

The behavioral literature indicates that the frequency of cooperation is increasing with this ratio. In this current experimental setting, games PD1 and PD2 have Rapoport ratio's of 0.50 and 0.71, respectively. As such we would expect cooperation rates to be greater in PD2 than PD1, and the aim of inference will be to identify a significant *directional* effect. Across our five populations we can therefore specify the directional comparative-static that such an experiment might set out to uncover within each population:

Hypothesis 3 (PD comparative static)**.** *Following the Rapoport ratio prediction, each of the five populations will have more cooperation in game PD2 than PD1.*

3.4 Results

We now outline the three core results from the experiment, before presenting them in detail: (i) The laboratory, CloudResearch approved list, and Prolific samples are similar over

the fraction of participants making Σ-dominated choices (∼10 percent), while this proportion is slightly larger for the virtual lab sample ($\sim 16 percent$) it is much larger in the MTurk sample (∼37 percent); (ii) Changing the order of actions has no significant effects in the lab, Cloud-R, and Prolific samples, while the V-Lab and MTurk samples exhibit a 19 and 16 percentage point swing respectively in favor of the first-listed choice; (iii) All five samples detect a significant proportion of participants who choose a fully selfish i-dominant strategy profile in the four games and moreover, the lab and V-Lab samples exhibit a standard response to increasing trade-offs between self and other as a drop in proportion of participants choosing to cooperate going from PD1 to PD2. However, we find this effect to be slightly smaller for the Cloud-R sample, MTurk and Prolific samples are essentially inelastic on this margin, with much stronger other-regarding behavior. We outline these results in detail below.

We highlight the core results in Figure 7 where (A) illustrates the Σ-dominated choice profiles and (B) shows the change in cooperation rates between PD1 and PD2 across the five samples.[17] The arrows in both panels indicate the direction and magnitude of change in participant proportions in the respective illustration when we move from listing C to listing D as the first action.

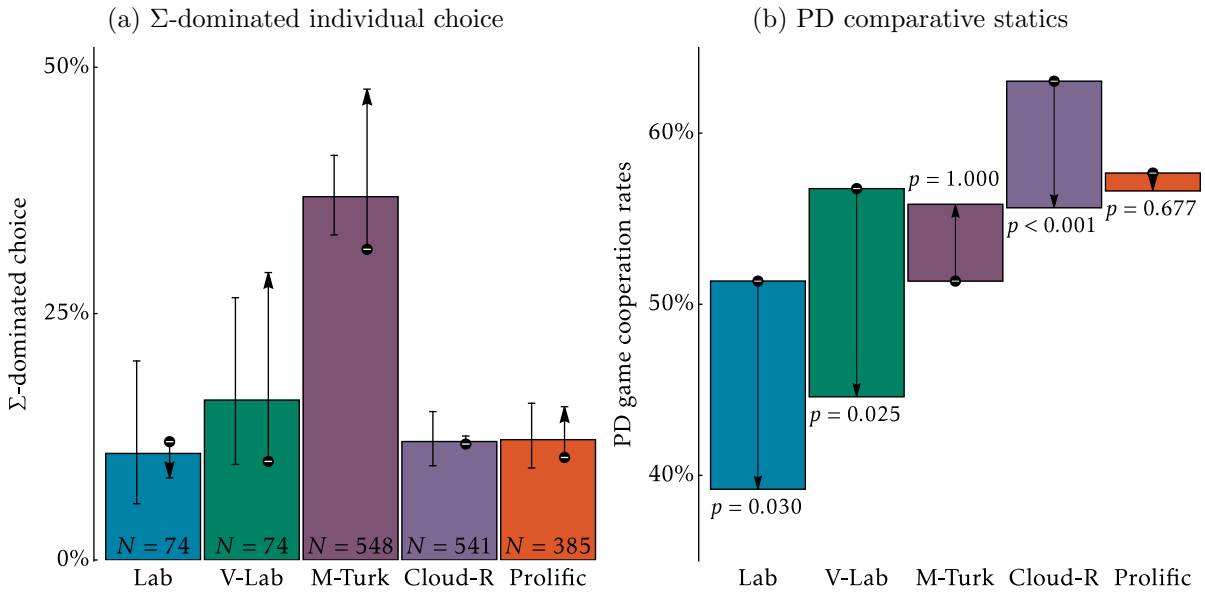
The Figure shows the rate at which individuals in the experiment make an obvious mistake with respect to the offered incentives i.e., choose the Σ-dominated actions. The proportion of participants choosing the σ-dominated actions is not statistically distinguishable between lab, v-lab, Cloud-R, and Prolific samples with approximately 12% of participants make a defect choices in the last two games. In contrast, for the MTurk sample this rate grows to more than one-in-three, significantly different from all other samples.[18] Moreover, as we explain next, even this number is perhaps an underestimate of the fraction of participants making choices orthogonal to the incentives.

The arrows in Figure 7(a) show the change in the participant proportion exhibiting a Σ-dominated choice when we move from listing C to listing D as the first action. The largest effects are in the V-lab and MTurk sample,s where listing the D-action first leads to a 19.2

[17] Table F1 in the Appendix provides detailed results.

[18] $p < 0.001$ for the pair-wise tests of proportions between MTurk and all four populations

Figure 7: Results by population



Notes - Panel (a): Error-bars indicate binomial-exact 95-percent confidence intervals for the proportion, where Arrows indicate the change over the framing variable when the C action is presented after the D action. Panel (b): Bars show the difference between cooperation in PD2 and PD1, with the arrows indicating the change; given p-values for participant-clustered tests for differences in proportion against one-sided alternative.

and 16.3 percentage points increase in the Σ-dominated fraction respectively ($p = 0.37$ and $p < 0.001$ on a test of proportions respectively for v-lab and MTurk). Despite successfully passing the screen questions—where participants must demonstrate their understanding of the game incentives or be kicked out—approximately one half of the MTurk sample make choices that indicate little awareness of the induced games. While approximately a third of this effect can be attributed to participants choosing the D action in games Σ-DOM1 and Σ-DOM2 simply because it is the first-listed option, the result still indicates that just under half of the sample makes choices that are orthogonal to the offered incentives. In contrast, despite similar costs per observation on Cloud-R and Prolific, the rates of such mistakes in these populations seems to be at most 15 percent, and we lack statistical power to say that it is even different from the laboratory.

We summarize these first two results as follows:

Result 1 (Σ-Dominance)**.** *The MTurk sample exhibits significantly more choices that violate any other-regarding preference that seeks to maximize efficiency than the other four populations. The lab, virtual lab, Cloud-R and, Prolific samples are not significantly different based on this data-quality measure.*

Result 2 (Response to frame)**.** *The virtual lab and MTurk samples exhibit significantly more choices that select the first-listed option. While there is also a small effect for the Prolific sample, the effect is only marginally significant. We do not detect any effect in the Cloud-R and lab samples from the re-framing.*

The focus of the above is on measuring the extent to which choices are being driven by mistakes - essentially any non-designed feature of the strategic environment that is orthogonal to the economic payoff variables. We now examine comparisons that may be more indicative of preference differences across the population. We focus on the difference between our two prisoners' dilemma games, PD1 and PD2. The prediction based on the behavioral literature (primarily lab-based studies) is that large proportion of participants will exhibit a form of partial cooperation, defecting in the first game but cooperating in the second.

Across the five samples, the comparative static as change in cooperation rate between PD1 and PD2 is summarized in Figure 7 (b) The arrows in the Figure show the change in the cooperation rate between the two games. Prior literature (see the data in [102]) suggests this change should be about 14 percent. The lab and virtual lab samples yield similar magnitudes, and are similar to each other.

In contrast, the MTurk and Prolific samples show smaller rates, both significantly different from the 14 percent prediction on each. The Cloud-R is closest to the prediction from the literature with 12.9 percent. Comparing the five populations, We find the lab and v-lab samples to be similar which are then followed by Cloud-R with a relatively smaller magnitude of response. The MTurk and Prolific samples on the other hand exhibit inelasticity in responses. Moreover, we reject equality of the proportions when we compare the lab to MTurk and Prolific ($p = 0.011$ for MTurk and $p = 0.045$ for Prolific), but we cannot reject equality between the lab and Cloud-R sample ($p = 0.161$).

Result 3 (Behavior Comparison)**.** *The lab, virtual lab and cloud research approved samples*

replicate the literature finding, with a substantial dropoff in cooperation between games PD2 and PD1, whereas this pattern is not found in either the Prolific or Mturk data.

3.5 Inferential Preferences

While the lab and v-lab samples replicate the standard comparative static over the Rapoport ratio, they do so with a relative lack of power due to the more substantial cost per observation. Cloud-R sample on the other retrieves this standard comparative static at a significantly lower cost per observation. In contrast, the MTurk and Prolific samples show a close-to-zero effect over the PD comparison, despite what should be much greater power even if the effect were substantially attenuated. What then can we conclude?

3.5.1 Framework

We use our PD comparative static hypothesis to generate a horse race between the populations, by examining their inferential power. Given that decisions are binary, tests over the difference across PD1 and PD2 are simply a function of the observed cooperation rates in each game and the sample size N.[19] The T-statistic for a test of a null effect between the two games for two samples of size N with cooperation rates of P_1 and P_2 is given by:

$$T\left(P_1, P_2, N\right) = \frac{\sqrt{N} \cdot \left(P_2 - P_1\right)}{\sqrt{\left(P_1 + P_2\right)\left(1 - \frac{P_1 + P_2}{2}\right)}}$$

For a qualitative alternative hypothesis that there is more cooperation in game PD2, we would therefore want the T-statistic to be greater than approximately 1.64 to attain significance at 95 percent confidence (or 90 percent on the two-sided alternative).

Modeling the number of cooperation decisions within the sample $N \cdot P_1$ and $N \cdot P_2$, as binomial draws with true proportions p_1 and p_2, respectively, it is therefore possible to calculate the likelihood that one would make a type-II error on this T-test when the two

[19] Greater statistical power can be generated if we also use the within-subject nature of the data, however, for simplicity we focus on a more-standard between-subject comparison

populations are in fact different. Using the [102] data, if the true cooperation rates for games PD1 and PD2 are 0.48 and 0.65, respectively, then the power of the test is a direct function of the sample-size N.[20] Given a fixed experimental budget, all else being equal, whichever population has the cheapest observations would yield the greatest power.

The previous conclusion, however, assumes populations are equally noisy while we have shown in the previous section this is not the case. There is a general wariness of online samples, with the thought that reduced control—for example, the ability for participants to multi-task while taking part in the study—leads to a larger proportion of participants being inattentive, or unresponsive to the economic treatment. To model this, we consider a population as having two fundamental properties: a dollar cost per observation c; and a noise/attenuation parameter γ. The population cost per observation is from the point of view of the experimenter. The population attenuation parameter γ affects the population-level expected behavior in PD game j, attenuating it towards a coin flip via $\gamma \cdot \frac{1}{2} + (1-\gamma) \cdot p_j$.[21] Each population identifies a particular bundle of cost per observation and attenuation parameter.

For each of these bundles, we think of the experimenter's problem through the lens of a consumer-choice-like problem, with statistical power in place of utility. Population A is preferred to population B if it provides greater statistical power under a fixed research budget. Populations are characterized by a cost/noise pair (c, γ), leading to a well-defined probability of making a Type-II error on the T-stat in (3.5.1).[22]. Using the idea that the experimenter's preferences are represented by statistical power (the probability of *not* making a type-II error) in Figure 8 we indicate indifference curves in (c, γ)-space for Hypothesis 3. In particular in panel (A) we indicate iso-power contours under a fixed experimental budget ($1,650, the approximate budget in our experiments), analogous to thinking about

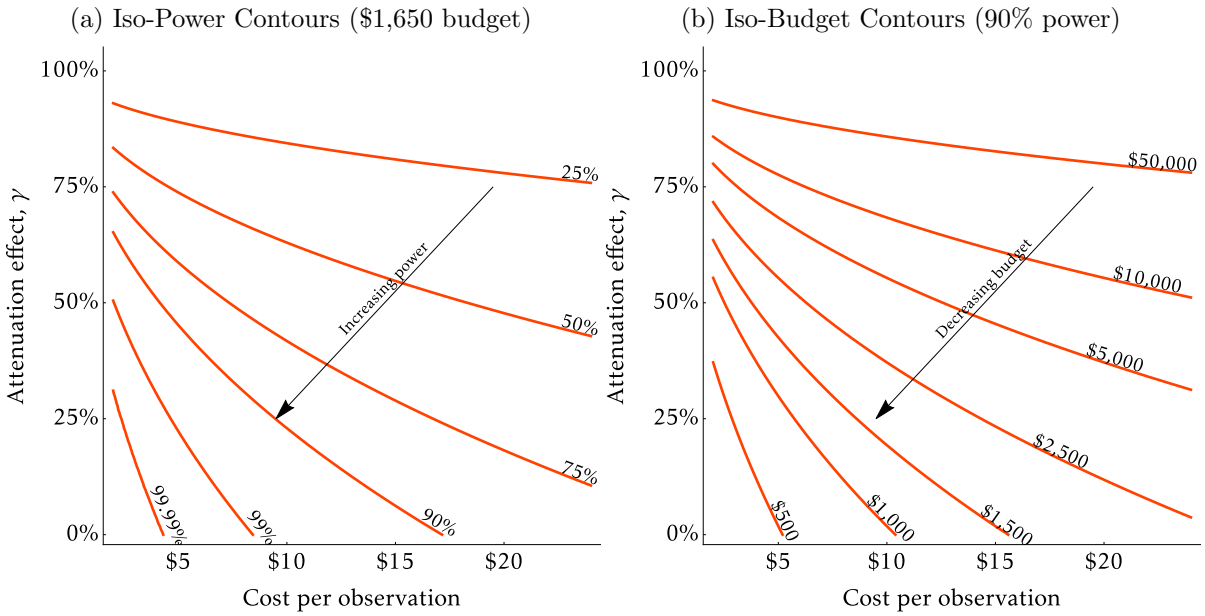
[20] We estimate this from [102] via a logit model with the Rapoport ratio as the sole predictor. The estimated model predicts a cooperation rate given by:

$$\text{Coop}(\rho) = \frac{1}{1 + 5.66 \cdot e^{-3.32\rho}}.$$

[21] The parameter γ here represents any form of attenuation. While one obvious source of attenuation here is inattention to the experimental environment, the parameter can also be interpreted as reduction in the elasticity of response within the population, as this will also reduce the quantitative size of the treatment response.

[22] We calculate the affordable sample size N from the total budget, while the sample proportions P_1 and P_2 are both attenuated towards 12 at the rate γ

Figure 8: experimenter inferential preferences: Noise versus Cost



Notes - Panel (a) shows iso-power contours (where labels indicate the probability of rejecting null) for an experiment with a $1,650 budget, while panel (b) shows iso-budget contours for a test with 90 percent power using a two-sample t-test on the PD cooperation rates with population variables derived from [102].

the indirect utility function in consumer choice. In panel (B) we indicate iso-budget lines under a fixed power level (90 percent), analogous to the expenditure function in the dual consumer choice problem.

Both γ and c are 'bads' from the experimenters point of view, and satisfy local non-satiation. As such, better populations have the smallest possible values for each, and the experimenter's preference is increasing as we move from the top-right to the bottom-left of Figure 8. Rather than fixing the sample size N within each population we have instead parameterized our experiments so that the per participant payments match standard rates on each population, but where we fixed the overall budget. This design choice, alongside the the identification of the population noise rates, allows us to run an inferential horse race between our populations.

Essentially, we ask which populations lies on the researcher-best contour in Figure 8.

For example, a sample with a cost per participant of \$17.50 and no noise has 90 percent power for our comparison of the PD games (the labels on each contour provide the type-II error probability). However, inspecting the figure, an online sample with a \$3 cost per observation is preferable even when half of the sample is pure noise as this sample would have 99 percent power under the same total expenditure. Next we detail the incentives offered to each population, and how budgets were determined.

3.5.2 Results

We first quantify the noise component across populations; this is the proportion of choices that is made randomly and is entirely orthogonal from the offered incentives. We then separate this component from what is potentially inelasticity in response, attenuation in the effect size driven by participants in the online populations simply having different preferences.

To assess the pure noise effects we focus on the behavior in games Σ-DOM1 and Σ-DOM2, where there is no real strategic tension, and suppose that there are three types. (i) An inattentive type that chooses the first-listed option regardless of the offered incentives, with measure γ_F in the population. (ii) an inattentive type that randomly chooses one of the two options in each game regardless of the incentives, measure γ_R in the population. (iii) An attentive type that responds to the incentives and satisfies Σ-dominance, with incidence $\gamma_\Sigma = 1 - \gamma_F - \gamma_R$.

Using a simple mixture model over these three types, we estimate the mass on each of the three types using the population samples.[23] Our model estimates indicate:

Lab For the lab sample we estimate: $\hat{\gamma}_F = 0.000$, $\hat{\gamma}_R = 0.144$ and $\hat{\gamma}_\Sigma = 0.856$.

V-Lab For the v-lab sample we estimate: $\hat{\gamma}_F = 0.000$, $\hat{\gamma}_R = 0.216$ and $\hat{\gamma}_\Sigma = 0.784$.

Mturk For the MTurk sample we estimate: $\hat{\gamma}_F = 0.107$, $\hat{\gamma}_R = 0.447$ and $\hat{\gamma}_\Sigma = 0.445$.

Cloud-R For the Cloud-R sample we estimate: $\hat{\gamma}_F = 0.000$, $\hat{\gamma}_R = 0.160$ and $\hat{\gamma}_\Sigma = 0.840$.

Prolific For the Prolific sample we estimate: $\hat{\gamma}_F = 0.022$, $\hat{\gamma}_R = 0.153$ and $\hat{\gamma}_\Sigma = 0.825$.

[23] The structural model thereby accounts for a one-in-four chance that a random type that would be classified as Σ-dominant.

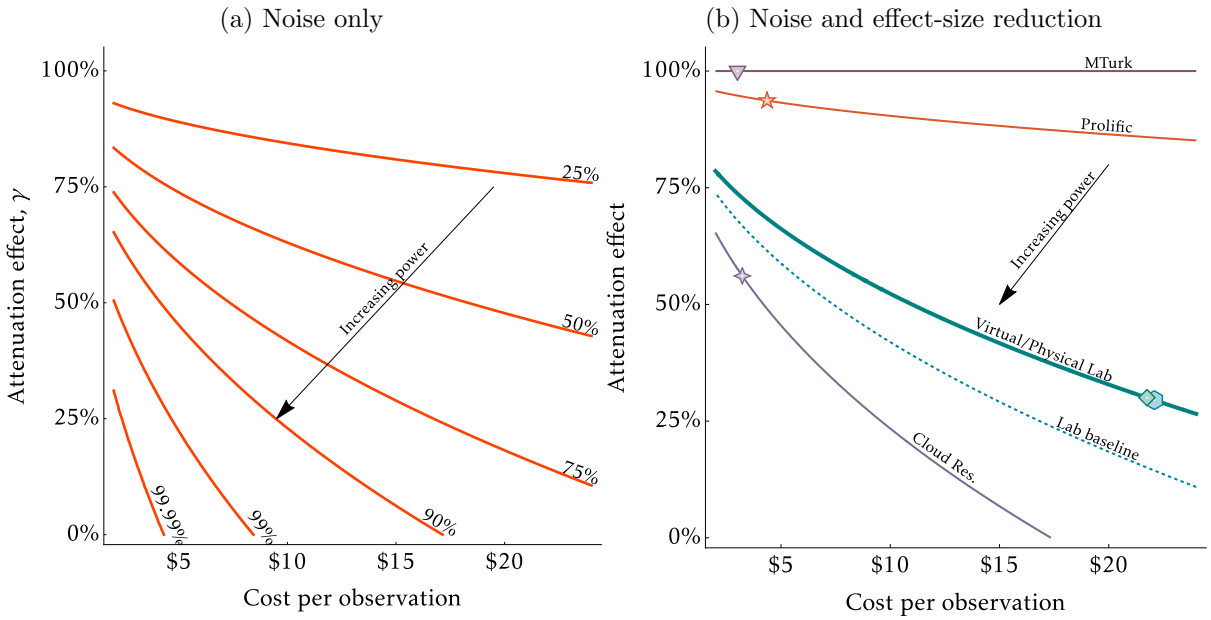
Thinking of the noise in the data simply due to inattentive participants, we find that MTurk data is approximately 55 percent noise, so just 45 percent of the respondents make choices driven by the offered economic incentives. In contrast, the signal component in the data is 83 percent for the Prolific sample, 84 percent for the Cloud-R sample, 78 percent for the v-lab sample and 86 percent for the laboratory sample (in the sense of satisfying Σ-dominance).

While the proportion of random decisions is certainly large in the MTurk sample, each observation is very cheap. At \$22 per lab observation, we can collect seven MTurk observations for every one in the lab. As such, if the minority of the MTurk sample that *is* incentive responsive has a similarly sized response over the two PD games to the prior literature, then the Mturk sample would dominate the lab sample in power terms. At a cost per observation of \$3.01 and an attenuation effect of 55.5 percent the MTurk population should still have 88 percent power for detecting a response when comparing behavior in the PD games, compared to 75.3 percent power for the lab sample (at 14.4 percent noise). Both would be clearly dominated by observations from the curated online Cloud-R and Prolific populations given their low noise and low cost per observation.

The curves in Figure 9(b) represent iso-power contours over populations with (c, γ) where we maintain the prior-literature effect sizes as the true effect size. Each line therefore represents the equivalent population to lab (blue), v-lab (green), MTurk (purple-triangle), Cloud-R (purple-star) and Prolific (orange) samples where the labeled points indicate our estimates for each population. The curve tells us that at 19.5 percent attenuation Prolific would still be preferable to MTurk even if its cost per observation increased to \$12.20. Alternatively, fixing the current MTurk cost the noise would have to shrink to 32.6 percent to match the Prolific sample's power. As such Cloud-R emerges as the winner here due to a noise level comparable to Prolific and a smaller cost per observation of \$3.23. All three online samples would dominate the laboratory if their true effect sizes were similar to the prior lab literature. However, noise due to inattention is not the sole factor to consider. Not only do we want a large fraction of attentive participants that are responsive to offered incentives, we also need that population to show an elastic response across our hypothesis.[24] The true attenuation

[24]For example, see [23] who demonstrate that while a particular real-effort task does show a qualitative

Figure 9: Population power



effects within the populations are an amalgam of both the noisy/inattentive participants, and any reduction in the effect size.

The solid lines in Figure 9(b) represent the iso-power lines when we calculate the total attenuation relative to the lab. Here we calculate the critical attenuation rate γ if the true effect matches the lab response that produces the same power test given the realized average response on each of our online populations. This substantially changes the ranking across our populations. MTurk is entirely unresponsive as the realized levels actually have the opposite sign from the hypothesis, and so full attenuation generates the most powerful test. In contrast Prolific does show a small amount of power, but because the difference in cooperation between games PD1 and PD2 is just 1 percentage point, the laboratory sample has much greater power.[25] When it comes to making inference over a relatively simple self/other strategic tradeoff, while the Prolific sample does exhibit substantial internal consistency, the

response to incentives, the effect sizes are economically too small for it to be an effective tool.

[25]Exacerbating the effect, cooperation rates are closer to 50 percent, causing the maximal standard error for a proportion test.

quantitative response to the Rapoport ratio is tiny. On the other hand, Cloud-R sample comes ahead again because of low noise paired with response to changes in the Rapoport ratio comparable predicted effect size.

Given the observed behavior in two out of our three online samples, one possible conclusion could be that some online samples do not respond to social-dilemma tensions in the same way as laboratory participants do. As such, the Rapoport ratio finding may be a lab-specific phenomenon. To examine this, and check that there is a response in more extreme games, we ran a second robustness study on CloudResearch and Prolific. Recruiting a further set of participants with a budget of about $500 (similar to the re-framed sample), we added two PD games to the previous four games. These additional games ramp up the PD tensions, so that the Rapoport ratios are 0.05 and 0.25 (the precise games are given in Appendix F). Looking to [102] for the prior literature effect size, we estimate that a comparison of the most extreme PD games (Rapport ratios of 0.71 and 0.05) in a lab sample should show a cooperation reduction of 48 percentage points.

While the robustness sample on Cloud-R continued to show responsiveness to changed incentives, we also find a more substantial effect in the second Prolific study. In the more-extreme PD games the cooperation rate in this sample falls to 0.320, with a comparable cooperation rate 0.584 in the least extreme PD game (see Appendix F for full analysis).[26] While the difference in cooperation is now highly significant ($p < 0.001$), the 26 percentage point reduction represent approximately half the effect size we would expect in the lab sample over the same comparison. Modeling the lab as having an attenuation purely driven by noise (0.144) and a cost per observation of $22, a comparison of the two most-extreme PD games with a true cooperation rate difference of 48 percentage points yields a near certain test when the budget is $1,650 ($> 99.99$ percent power). Despite a reduced effect size on Prolific, the cheaper observations yield almost the same effect. A different way to see the results comes from thinking about the dual problem: what is the experimental budget that yields a 95 percent power test in each population on this more-extreme comparison? A budget of $379 on Prolific yields the same 95 percent power test as a $484 budget in the laboratory.

Our results on the extended games suggest that online populations *are* capable of uncov-

[26]We cannot reject the original cooperation levels over PD1 and PD2, despite the increase to 6 choices.

ering the same qualitative patterns as the laboratory. However, two caveats are appropriate here. First, the substantial noise on MTurk suggests that more-curated populations are likely superior (or that greater internal validity checks are required, where sub-sampling the population substantially increases the cost of each valid MTurk observation). Second, the elasticity of response to other-regarding tensions in some online populations can be much-attenuated from previous lab samples. Under more-nuanced parameterizations, the online populations' response is simply too small to have power, where the lab sample shows much greater response. If the aim is purely to uncover qualitative findings, or to gauge the order of magnitude of an effect, the conclusion from our study is to eschew all subtlety. So long as the parameterization can generate a moderate effect size, much greater power can be produced by the smaller cost per observation.

On the flip-side of the coin, our study also points to the usefulness of the laboratory. Lab participants whether physically present for the experiment or participating online consistently show responses. In studies where the aim is to educe more nuanced findings—calibrating a non-linear model say, where estimating curvature requires smaller step-size in the treatments—the lab can play a useful role. Despite the expense per observation, the combination of more-elastic response to shifts in the incentives and a low noise rates make the lab a better tool. While our study has no variation in the level of complexity, the lab offers a conducive environment for testing knottier economic hypotheses. By controlling participants' outside-option activities and removing distraction, lab samples allow for experimenters to induce more-complex artificial settings. While there is certainly a place for online samples, given their low cost and ease of acquisition, a lack of control on attention does seem to be a problem in some online populations.

3.6 Conclusions

We examine five populations for conducting experimental studies. Rather than a validation of comparative statics across the differing populations, we take a different tack. We reflect ecological differences in the price for each observation and the noise in samples from

each by constructing the experimenter's preference over populations via inferential power. That is, to what extent might one want to trade off some level of noise for much cheaper observations, assuming the experimenter has a fixed budget to spend.

Our design measures both the noise in each population (which we attribute to inattention through a weak assumption on preferences) and a more-nuanced hypothesis on the response to social dilemma tensions. We fix the experimental budget on each population and, by varying the scale of the incentives so that they match standard levels in each population, we thereby vary the number of observations collected on each. We then assess the noisiness in behavior and the extent to which each population replicates a standard comparative static results from the literature.

Our findings indicate that even a small number of participants in the lab and the virtual lab samples (with relatively expensive observation costs) replicate standard findings. However, two of the online samples with lower observation costs - MTurk and Prolific - do not. Cloud-R on the other hand wins on this margin where we find that it replicates the standard findings as well as has low cost per observation.

In terms of the proportion of noise in the data, laboratory sample both from the physical lab and virtual lab, Cloud-R and Prolific have relatively low levels. In contrast, at 55 percent our MTurk sample is particularly noisy, despite screens to ensure understanding. However, even at this level of noise, the very cheap observations from MTurk should dominate the laboratory from an inference point of view—though in pure terms of noise versus cost, Cloud-R and Prolific dominates both.

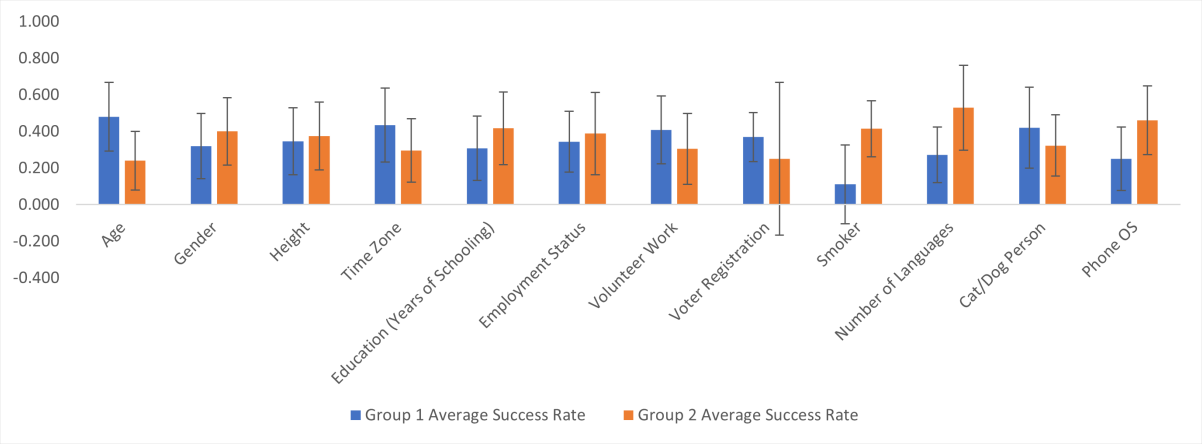
However, beyond just noise, our lab samples consistently exhibit much greater elasticity of response to treatment where the MTurk and Prolific samples are essentially inelastic on that margin. Only one out of three online samples show dominance over the lab sample when attenuation is accounted both in terms of noise as well as elasticity of response, generating an overall doubt over the use of online populations. While these results may be specific to social dilemmas—where our more-generalizable noise estimate clearly outline the power benefits of Prolific samples—they outline that despite the expense per observation, lab samples can offer greater power. However, we go on to show that by making the size of the treatment effect larger Prolific can again dominate the lab in power terms, despite the reduced elasticity of

response.

The very substantial noise in the MTurk sample may be a recent phenomenon (where studies suggest the population has recently declined). However, our analysis suggests that despite being the cheapest of the samples, MTurk offers a false economy. While almost 50 percent more expensive, Prolific observations offers substantially greater inferential power by reducing noise. In fact, even a small increase in cost as posed by the use of CloudResearch approved list of MTurk workers substantially improves inferential power obtained. While there are domains where lab data is preferable, if an online population is to be sampled, our study clearly indicates that more-curated populations such as CloudResearch approved list of MTurk workers and Prolific are superior to general MTurk population.

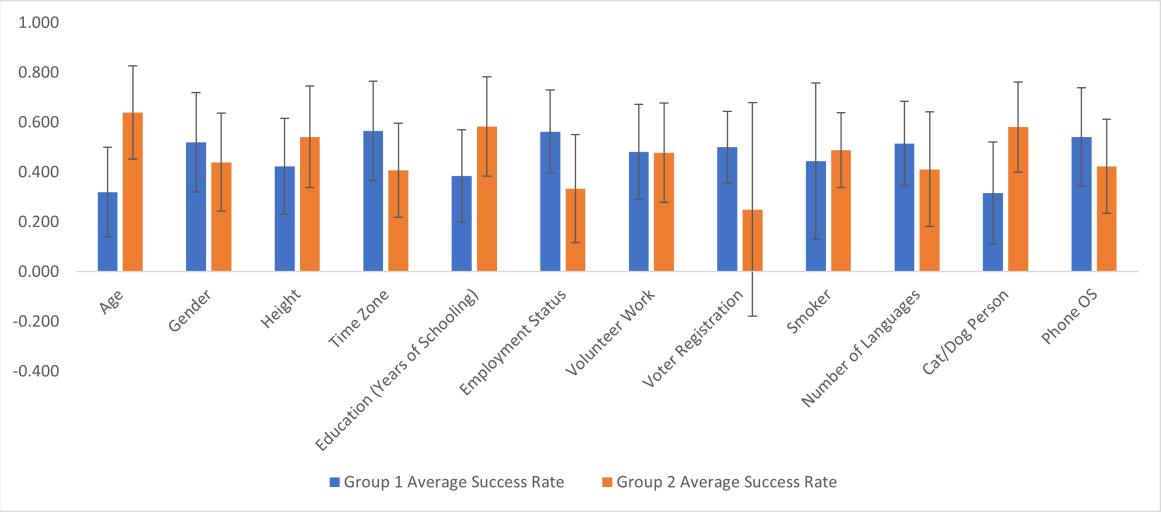
Appendix A Chapter 1: Worker and Evaluator Results

Figure A1: Average rate of success by worker demographic characteristics: Business



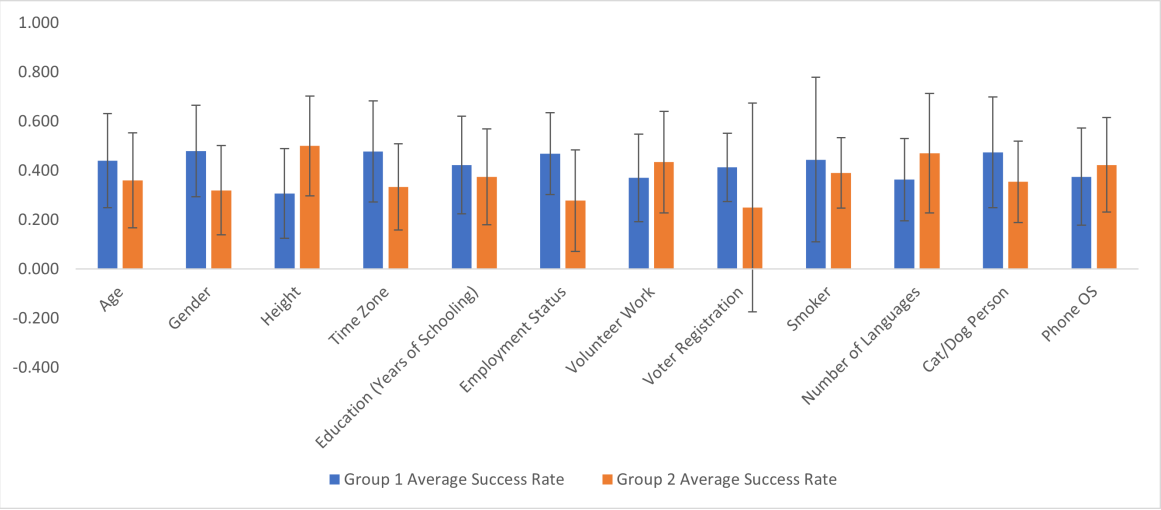
Notes - This figure shows average success rate by worker characteristics. Groups are divided based on demographic characteristics as shown in Table 1. Error bars represent bootstrapped 95% confidence intervals.

Figure A2: Average rate of success by worker demographic characteristics: Sports



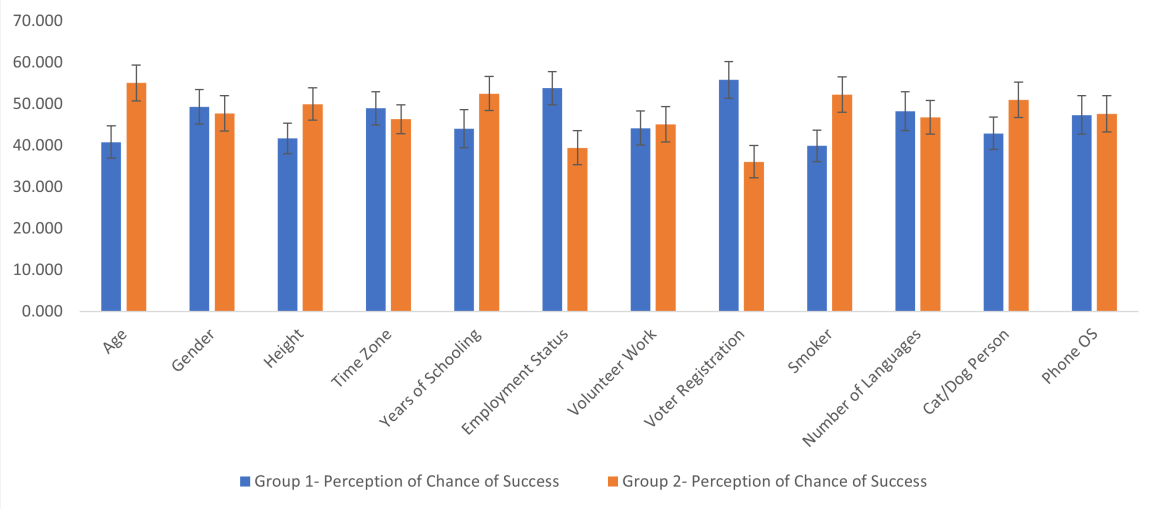
Notes - This figure shows average success rate by worker characteristics. Groups are divided based on demographic characteristics as shown in Table 1. Error bars represent bootstrapped 95% confidence intervals.

Figure A3: Average rate of success by worker demographic characteristics: Video Games



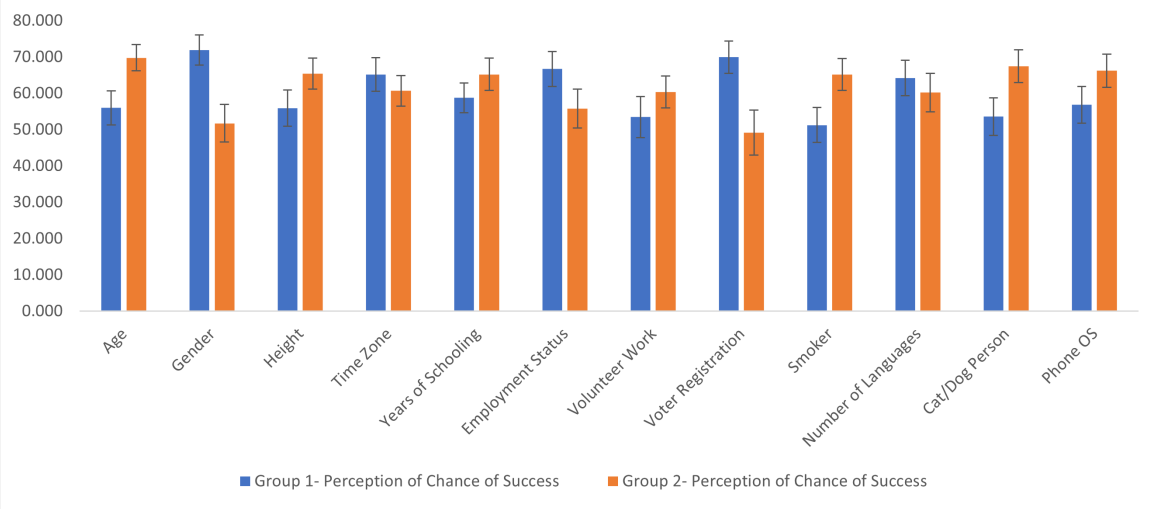
Notes - This figure shows average success rate by worker characteristics. Groups are divided based on demographic characteristics as shown in Table 1. Error bars represent bootstrapped 95% confidence intervals.

Figure A4: Average perceived likelihood of success from experiment 2 split by worker demographic characteristics: Business



Notes - This figure shows average perceived likelihood of success from experiment 2 split by worker characteristics. Groups are divided based on demographic characteristics as shown in Table 1. Error bars represent bootstrapped 95% confidence intervals.

Figure A5: Average perceived likelihood of success from experiment 2 split by worker demographic characteristics: Sports



Notes - This figure shows average perceived likelihood of success from experiment 2 split by worker characteristics. Groups are divided based on demographic characteristics as shown in Table 1. Error bars represent bootstrapped 95% confidence intervals.

Appendix B Chapter 1: Additional Tables - Employer

Table B1: Heterogeneous effects in evaluations with employers disregarding information excluded - OLS with fixed effects

Dependent Variable:	Rounds 4 to 6			
	Evaluation			
	Subgroup - More		Subgroup - Less	
	(1)	(2)	(3)	(4)
I(Female)	-14.27***	-14.95***	-4.045***	-3.853***
	(1.391)	(1.504)	(0.682)	(0.723)
I(TempAA)*I(Female)	5.929***	6.121***	0.260	0.290
	(1.860)	(1.856)	(0.962)	(0.956)
Constant	67.30***	64.93***	68.98***	66.36***
	(0.653)	(1.373)	(0.340)	(0.686)
N	1116	1116	2700	2700
Worker controls		Yes		Yes
Disregard excluded	Yes	Yes	Yes	Yes

Notes - *** p<0.01, ** p<0.05, * p<0.1; This table presents OLS regression results on evolution of beliefs about performance elicited as evaluations on a scale of 0 to 100. The estimation uses a fixed effects model where each group comprises a single round where an employer makes decisions over 4 resumes. Subgroup - More (Less) represents the subgroups of employers within the two experimental treatments who are more (less) than 32% likely to hire both men in round 1. Worker controls include demographic characteristics presented on their resume - employment status, education, number of spoken languages, and time zone of residence. All options within each resume characteristic are aggregated in two groups characterized by the median worker's characteristic from the worker experiment. Employers disregarding information are identified through the final stage of the experiment as those who do not opt in to get information on worker performance for any worker. Standard errors are shown in parentheses.

Table B2: Heterogeneous effects in evaluations with moving cutoff - OLS with fixed effects

Dependent Variable:	Rounds 4 to 6								
	Evaluations								
	> 0	> 0.05	> 0.10	> 0.15	> 0.20	> 0.25	> 0.30	> 0.35	> 0.40
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
I(Female)	-6.739***	-6.747***	-6.837***	-7.755***	-9.254***	-11.64***	-12.68***	-16.78***	-19.43***
	(0.559)	(0.561)	(0.575)	(0.646)	(0.820)	(0.970)	(1.224)	(1.470)	(1.779)
I(TempAA)*I(Female)	0.795	0.763	0.672	1.102	1.400	2.733**	3.255**	7.249***	8.840***
	(0.789)	(0.792)	(0.812)	(0.900)	(1.137)	(1.346)	(1.657)	(1.970)	(2.354)
Constant	68.74***	68.63***	68.79***	68.42***	67.72***	67.64***	67.83***	68.93***	69.75***
	(0.279)	(0.280)	(0.287)	(0.318)	(0.402)	(0.475)	(0.584)	(0.692)	(0.824)
N	5160	5124	4896	4092	2880	2148	1584	1164	840

Notes - *** p<0.01, ** p<0.05, * p<0.1; This table presents OLS regression results on evolution of beliefs about performance elicited as evaluations on a scale of 0 to 100. The estimation uses a fixed effects model where each group comprises a single round where an employer makes decisions over 4 resumes. Moving from columns (1) to (9) the sample restriction based on probability of hiring both men in round 1 increases, for e.g. column (4) shows the results when probability of hiring both men in round 1 is greater than 15%. Standard errors are shown in parentheses.

Table B3: Heterogeneous effects in evaluations with moving cutoff and employers disregarding information excluded - OLS with fixed effects

Dependent Variable:	Rounds 4 to 6								
	Evaluations								
	> 0	> 0.05	> 0.10	> 0.15	> 0.20	> 0.25	> 0.30	> 0.35	> 0.40
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
I(Female)	-6.785***	-6.785***	-6.803***	-7.924***	-9.431***	-11.73***	-12.55***	-15.90***	-17.05***
	(0.633)	(0.633)	(0.644)	(0.727)	(0.929)	(1.088)	(1.344)	(1.620)	(1.997)
I(TempAA)*I(Female)	1.564*	1.544*	1.339	2.027**	2.357*	3.660**	4.441**	7.078***	6.582**
	(0.879)	(0.880)	(0.896)	(0.999)	(1.268)	(1.489)	(1.804)	(2.140)	(2.594)
Constant	68.49***	68.44***	68.66***	68.22***	67.55***	67.04***	67.14***	67.92***	68.94***
	(0.311)	(0.311)	(0.317)	(0.353)	(0.447)	(0.525)	(0.634)	(0.748)	(0.901)
N	3816	3804	3648	3060	2124	1596	1212	900	648
Disregard excluded	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes - *** p<0.01, ** p<0.05, * p<0.1; This table presents OLS regression results on evolution of beliefs about performance elicited as evaluations on a scale of 0 to 100. The estimation uses a fixed effects model where each group comprises a single round where an employer makes decisions over 4 resumes. Moving from columns (1) to (9) the sample restriction based on probability of hiring both men in round 1 increases, for e.g. column (4) shows the results when probability of hiring both men in round 1 is greater than 15%. Employers disregarding information are identified through the final stage of the experiment as those who do not opt in to get information on worker performance for any worker. Standard errors are shown in parentheses.

Table B4: Heterogeneous effects in evaluations as mechanism with employers disregarding information excluded - Conditional Fixed Effects Logit Regression

Dependent Variable:	Rounds 4 to 6			
	I(Hired)			
	Subgroup - More		Subgroup - Less	
	(1)	(2)	(3)	(4)
I(Female)	-0.586***	0.219	-0.429***	-0.248**
	(0.156)	(0.186)	(0.0945)	(0.101)
I(TempAA)*I(Female)	0.355*	0.0415	0.393***	0.413***
	(0.209)	(0.236)	(0.133)	(0.142)
Evaluation		0.0654***		0.0576***
		(0.00577)		(0.00412)
N	1116	1116	2700	2700
Disregard excluded	Yes	Yes	Yes	Yes

Notes - *** p<0.01, ** p<0.05, * p<0.1; This table presents conditional fixed logit regression results from hiring decisions where a group comprises a single round where an employer makes decisions over 4 resumes. Subgroup - More (Less) represents the subgroups of employers within the two experimental treatments who are more (less) than 32% likely to hire both men in round 1. Dependent variable is an indicator variable =1 when a worker is hired. Standard errors are shown in parentheses.

Table B5: 1st and 2nd hiring results- Conditional Fixed Effects Logit Regression

	Rounds 4 to 6			
Dependent Variable:	I(1st Hired)		I(2nd Hired)	
	(1)	(2)	(3)	(4)
I(Female)	-0.522***	-0.513***	-0.315***	-0.219**
	(0.0816)	(0.0882)	(0.0833)	(0.0906)
I(TempAA)*I(Female)	0.299***	0.308***	0.204*	0.195
	(0.114)	(0.116)	(0.118)	(0.119)
N	5160	5160	3870	3870
Worker controls		Yes		Yes

Notes - *** p<0.01, ** p<0.05, * p<0.1; This table presents conditional fixed logit regression results from hiring decisions where a group comprises a single round where an employer makes decisions over 4 resumes. Dependent variable for specifications 1 and 2 is an indicator variable =1 when a worker is hired 1st and, for specifications 3 and 4 it is an indicator variable =1 when worker is hired 2nd after excluding those hired 1st. Worker controls include demographic characteristics presented on their resume - employment status, education, number of spoken languages, and time zone of residence. All options within each resume characteristic are aggregated in two groups characterized by the median worker's characteristic from the worker experiment. Standard errors are shown in parentheses.

Table B6: Heterogeneous effects in evaluations as mechanism for 1st and 2nd hiring decisions - Conditional Fixed Effects Logit Regression

	Rounds 4 to 6							
Dependent Variable:	I(1st Hired)		I(1st Hired)		I(2nd Hired)		I(2nd Hired)	
	Subgroup - More		Subgroup - Less		Subgroup - More		Subgroup - Less	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
I(Female)	-0.665***	-0.0903	-0.475***	-0.284***	-0.457***	0.124	-0.267***	-0.0768
	(0.166)	(0.187)	(0.0939)	(0.101)	(0.166)	(0.189)	(0.0964)	(0.102)
I(TempAA)*I(Female)	0.303	0.204	0.312**	0.352**	0.266	0.00159	0.191	0.241*
	(0.221)	(0.240)	(0.133)	(0.141)	(0.225)	(0.244)	(0.139)	(0.146)
Evaluation		0.0550***		0.0581***		0.0493***		0.0566***
		(0.00567)		(0.00404)		(0.00560)		(0.00450)
N	1428	1428	3732	3732	1071	1071	2799	2799

Notes - *** p<0.01, ** p<0.05, * p<0.1; his table presents conditional fixed logit regression results from hiring decisions where a group comprises a single round where an employer makes decisions over 4 resumes. Subgroup - More (Less) represents the subgroups of employers within the two experimental treatments who are more (less) than 32% likely to hire both men in round 1. Dependent variable for specifications 1 to 4 is an indicator variable =1 when a worker is hired 1st and, for specifications 5 to 8 it is an indicator variable =1 when worker is hired 2nd after excluding those hired 1st. Standard errors are shown in parentheses.

Appendix C Chapter 2: Full Table

Table C1: Full Classification of Papers According to Female-Sensitivity Hypothesis

Paper	**Topic**	**Classification**
[2]	Risk preferences	Men $\sim$ Women
[3]	NA	NA
[4]	Other-regarding preferences	NA
[6]	Competition - Entry	NA
[8]	Risk preferences	NA
[9]	Negotiation - Performance	Women > Men
[11]	Competition - Entry	Men $\sim$ Women
[10]	Other-regarding preferences	Men > Women
[12]	Combination	Men > Women
[14]	Other-regarding preferences	Men > Women
[15]	Other-regarding preferences	Men > Women
[13]	Risk preferences	NA
[16]	Competition - Performance	Men > Women
[19]	NA	Insufficient Information
[21]	Competition - Performance and Entry	Men > Women
[22]	Competition - Entry	NA
[20]	Competition - Performance and Entry	Men $\sim$ Women
[25]	NA	Men $\sim$ Women
[28]	NA	Men $\sim$ Women
[30]	NA	Women > Men
[29]	Competition - Performance	Men $\sim$ Women
[31]	Non-promotable tasks	Men $\sim$ Women

Paper	Topic	Classification
[32]	Competition - Performance and Entry	Men > Women
[33]	Competition - Entry	NA
[35]	Competition - Performance and Entry	Men > Women
[34]	Competition - Performance and Entry	NA
[36]	NA	Men ∼ Women
[37]	Competition - Performance and Entry	Men > Women
[38]	Other-regarding preferences	Men ∼ Women
[40]	Other-regarding preferences	Insufficient Information
[349]	Other-regarding preferences	Women > Men
[42]	Other-regarding preferences	Men > Women
[43]	Competition - Performance and Entry	Men > Women
[44]	NA	Men > Women
[45]	Other-regarding preferences	Men ∼ Women
[48]	Other-regarding preferences	Women > Men
[50]	Other-regarding preferences	Insufficient Information
[51]	NA	Women > Men
[53]	NA	NA
[52]	Discrimination	Men ∼ Women
[54]	Discrimination	NA
[55]	Other-regarding preferences	Men ∼ Women
[56]	Combination	Men ∼ Women
[58]	Discrimination	Insufficient Information
[59]	Competition - Performance and Entry	Women > Men
[60]	Risk preferences	NA
[57]	Risk preferences	NA
[62]	NA	Men ∼ Women
[63]	Risk preferences	NA
[64]	NA	Men ∼ Women
[65]	Other-regarding preferences	Men ∼ Women

Paper	Topic	Classification
[66]	Combination	Men $\sim$ Women
[67]	Other-regarding preferences	Men $>$ Women
[69]	Negotiation - Performance	Women $>$ Men
[71]	Other-regarding preferences	Men $\sim$ Women
[72]	Other-regarding preferences	Men $\sim$ Women
[73]	NA	Men $\sim$ Women
[76]	Competition - Entry	Insufficient Information
[75]	Competition - Entry	Men $>$ Women
[74]	Competition - Performance and Entry	Men $>$ Women
[77]	Other-regarding preferences	Men $\sim$ Women
[78]	Other-regarding preferences	Women $>$ Men
[80]	Competition - Performance and Entry	Men $\sim$ Women
[81]	Speaking out	NA
[82]	Competition - Entry	NA
[87]	Competition - Entry	Women $>$ Men
[84]	Competition - Entry	Women $>$ Men
[83]	Competition - Entry	Women $>$ Men
[85]	Competition - Entry	NA
[86]	Competition - Performance and Entry	Men $\sim$ Women
[88]	Discrimination	NA
[90]	Other-regarding preferences	Women $>$ Men
[91]	Competition - Performance and Entry	Men $\sim$ Women
[89]	Other-regarding preferences	Women $>$ Men
[258]	NA	NA
[92]	Competition - Entry	Men $>$ Women
[94]	Risk preferences	Insufficient Information
[93]	Competition - Performance and Entry	NA
[95]	Competition - Entry	Women $>$ Men
[98]	Other-regarding preferences	Men $>$ Women

Paper	Topic	Classification
[97]	Discrimination	NA
[96]	Risk preferences	Women > Men
[101]	Risk preferences	NA
[103]	Other-regarding preferences	Men ∼ Women
[100]	Risk preferences	Men > Women
[99]	Discrimination	Women > Men
[104]	Other-regarding preferences	Men ∼ Women
[105]	Other-regarding preferences	NA
[259]	NA	NA
[107]	Combination	NA
[107]	NA	Women > Men
[109]	Other-regarding preferences	Women > Men
[110]	Other-regarding preferences	NA
[99]	Other-regarding preferences	Insufficient Information
[114]	Competition - Entry	Men ∼ Women
[115]	Competition - Performance	Men ∼ Women
[116]	Other-regarding preferences	Women > Men
[117]	Other-regarding preferences	NA
[120]	Other-regarding preferences	Men > Women
[119]	Other-regarding preferences	Men ∼ Women
[122]	Risk preferences	Men ∼ Women
[121]	NA	NA
[123]	Competition - Entry	Women > Men
[152]	Negotiation - Performance	Insufficient Information
[124]	Other-regarding preferences	Men ∼ Women
[125]	NA	Women > Men
[128]	Competition - Performance and Entry	Men > Women
[129]	Competition - Performance and Entry	Men > Women
[130]	Competition - Performance and Entry	NA

Paper	Topic	Classification
[131]	Competition - Entry	NA
[132]	Competition - Performance	Men > Women
[133]	NA	Men ∼ Women
[298]	Speaking out	Women > Men
[351]	Non-promotable tasks	Women > Men
[137]	Competition - Performance	NA
[138]	Other-regarding preferences	Women > Men
[141]	Other-regarding preferences	Men > Women
[144]	Time preferences	NA
[143]	Negotiation - Performance	Men ∼ Women
[145]	Competition - Performance and Entry	NA
[146]	Risk preferences	NA
[148]	NA	Men > Women
[22]	Competition - Entry	Men ∼ Women
[147]	Combination	Women > Men
[83]	Competition - Entry	Men > Women
[149]	Risk preferences	Women > Men
[151]	Other-regarding preferences	Men ∼ Women
[150]	Other-regarding preferences	NA
[156]	NA	Men > Women
[157]	Other-regarding preferences	Women > Men
[158]	Other-regarding preferences	NA
[159]	Other-regarding preferences	Men ∼ Women
[160]	Risk preferences	NA
[154]	Risk preferences	Men > Women
[161]	Risk preferences	NA
[155]	Negotiation - Performance	Men ∼ Women
[153]	Negotiation - Performance	Men ∼ Women
[199]	NA	Insufficient Information

Paper	Topic	Classification
[162]	Other-regarding preferences	Women > Men
[163]	NA	NA
[164]	Risk preferences	Women > Men
[166]	Risk preferences	Men > Women
[167]	Competition - Performance	Men ∼ Women
[165]	Other-regarding preferences	NA
[168]	NA	Insufficient Information
[169]	Negotiation - Entry	Men ∼ Women
[170]	NA	Men ∼ Women
[171]	Risk preferences	NA
[173]	Risk preferences	Men ∼ Women
[174]	Risk preferences	Men > Women
[175]	NA	NA
[176]	Competition - Performance and Entry	NA
[177]	Competition - Entry	NA
[301]	NA	Men ∼ Women
[178]	NA	Women > Men
[179]	Risk preferences	NA
[180]	Other-regarding preferences	Men > Women
[182]	NA	NA
[183]	NA	Men ∼ Women
[184]	Other-regarding preferences	Men ∼ Women
[185]	NA	Men > Women
[186]	Competition - Performance	Men ∼ Women
[188]	NA	Men ∼ Women
[191]	Competition - Performance	NA
[189]	Competition - Entry	NA
[190]	Competition - Performance	NA
[192]	Other-regarding preferences	Women > Men

Paper	Topic	Classification
[195]	Competition - Entry	Men $\sim$ Women
[194]	Other-regarding preferences	Men $>$ Women
[197]	Risk preferences	Men $\sim$ Women
[198]	NA	Women $>$ Men
[200]	Competition - Performance	Men $\sim$ Women
[134]	Competition - Entry	Men $>$ Women
[202]	Negotiation - Performance	Insufficient Information
[203]	Competition - Performance and Entry	NA
[204]	NA	NA
[207]	Risk preferences	Men $\sim$ Women
[206]	Risk preferences	Men $\sim$ Women
[208]	Competition - Performance and Entry	Men $\sim$ Women
[209]	Other-regarding preferences	Women $>$ Men
[210]	Competition - Performance and Entry	Men $>$ Women
[211]	Competition - Entry	Men $>$ Women
[212]	Negotiation - Performance	NA
[214]	Combination	NA
[215]	NA	Women $>$ Men
[216]	Competition - Performance and Entry	Women $>$ Men
[237]	Other-regarding preferences	Men $\sim$ Women
[246]	Other-regarding preferences	NA
[217]	NA	NA
[258]	Other-regarding preferences	NA
[355]	NA	Insufficient Information
[220]	Competition - Entry	Men $>$ Women
[221]	Risk preferences	Men $\sim$ Women
[222]	NA	NA
[348]	NA	NA
[7]	NA	NA

Paper	Topic	Classification
[223]	Competition - Performance	Men $\sim$ Women
[224]	NA	Women > Men
[226]	NA	NA
[227]	Discrimination	NA
[228]	Competition - Entry	Men > Women
[229]	Risk preferences	NA
[231]	NA	NA
[230]	Risk preferences	NA
[233]	Competition - Entry	Men > Women
[232]	Other-regarding preferences	Men $\sim$ Women
[234]	Other-regarding preferences	Men $\sim$ Women
[235]	Other-regarding preferences	NA
[236]	Other-regarding preferences	Men > Women
[238]	NA	Women > Men
[239]	Speaking out	NA
[240]	Other-regarding preferences	Men $\sim$ Women
[241]	Risk preferences	NA
[242]	NA	Men $\sim$ Women
[243]	Other-regarding preferences	NA
[244]	NA	NA
[245]	Other-regarding preferences	Women > Men
[247]	Discrimination	Men $\sim$ Women
[248]	Other-regarding preferences	Men $\sim$ Women
[249]	NA	Men $\sim$ Women
[250]	NA	Women > Men
[287]	Other-regarding preferences	Men $\sim$ Women
[251]	Competition - Performance	Men $\sim$ Women
[252]	Risk preferences	NA
[253]	Negotiation - Entry	Men > Women

Paper	Topic	Classification
[254]	Competition - Performance and Entry	Women > Men
[255]	Other-regarding preferences	Men > Women
[260]	Other-regarding preferences	NA
[262]	NA	Men ∼ Women
[263]	Other-regarding preferences	Women > Men
[264]	Competition - Entry	Men > Women
[265]	Competition - Performance and Entry	Women > Men
[266]	Competition - Entry	Women > Men
[267]	Competition - Performance	Women > Men
[268]	NA	Men > Women
[270]	Competition - Performance and Entry	Men > Women
[272]	Risk preferences	NA
[273]	Other-regarding preferences	Women > Men
[274]	Other-regarding preferences	Men > Women
[275]	NA	NA
[276]	NA	Men ∼ Women
[277]	Other-regarding preferences	Women > Men
[279]	Other-regarding preferences	Insufficient Information
[280]	Risk preferences	Insufficient Information
[282]	Risk preferences	NA
[281]	Competition - Performance and Entry	NA
[283]	NA	Men ∼ Women
[284]	NA	Men ∼ Women
[285]	Other-regarding preferences	Men ∼ Women
[286]	Competition - Performance and Entry	Men > Women
[290]	Competition - Performance and Entry	Men ∼ Women
[291]	NA	Men ∼ Women
[289]	Competition - Entry	NA
[292]	NA	Women > Men

Paper	Topic	Classification
[293]	Other-regarding preferences	NA
[294]	NA	Men > Women
[295]	Competition - Performance	NA
[296]	NA	Men ∼ Women
[297]	NA	Men ∼ Women
[135]	Competition - Performance	NA
[300]	Other-regarding preferences	Insufficient Information
[302]	NA	NA
[303]	Other-regarding preferences	Women > Men
[304]	Discrimination	NA
[305]	NA	NA
[307]	NA	Women > Men
[308]	Competition - Performance and Entry	Men > Women
[344]	Other-regarding preferences	Women > Men
[309]	Risk preferences	NA
[312]	NA	NA
[313]	Competition - Performance and Entry	NA
[310]	Discrimination	NA
[311]	Discrimination	NA
[315]	Discrimination	NA
[70]	Negotiation - Performance	Women > Men
[68]	Negotiation - Entry	Men ∼ Women
[316]	NA	NA
[317]	NA	NA
[319]	Competition - Performance and Entry	Men > Women
[320]	Competition - Performance and Entry	Men ∼ Women
[321]	Competition - Entry	Women > Men
[323]	NA	NA
[324]	Competition - Entry	Men ∼ Women

Paper	Topic	Classification
[325]	Risk preferences	NA
[326]	Risk preferences	NA
[327]	Discrimination	NA
[328]	Other-regarding preferences	NA
[329]	Competition - Performance	Men $\sim$ Women
[330]	Other-regarding preferences	Men $\sim$ Women
[332]	Combination	Men $\sim$ Women
[331]	Competition - Performance and Entry	Men $\sim$ Women
[333]	Discrimination	NA
[334]	Negotiation - Entry	Men $\sim$ Women
[336]	Other-regarding preferences	Men $\sim$ Women
[337]	Other-regarding preferences	Men $>$ Women
[338]	NA	Women $>$ Men
[225]	Competition - Performance	Men $>$ Women
[340]	NA	Men $>$ Women
[342]	Competition - Entry	Insufficient Information
[341]	NA	Men $\sim$ Women
[343]	Combination	NA
[346]	Other-regarding preferences	Women $>$ Men
[350]	Competition - Performance and Entry	Men $>$ Women
[352]	Other-regarding preferences	NA
[353]	NA	Women $>$ Men
[354]	Other-regarding preferences	Men $>$ Women
[356]	Competition - Performance and Entry	Men $>$ Women
[357]	NA	Men $>$ Women
[359]	Competition - Entry	Men $>$ Women
[361]	Competition - Performance and Entry	Men $\sim$ Women
[362]	Risk preferences	NA

Paper	Topic	Classification
[363]	Competition - Performance and Entry	Men $\sim$ Women

Appendix D Chapter 2: Paper Review Process

We started with a few seminal papers in experimental economics which either focus directly on gender or have tangential but important results about gender. We started with a forward literature review on these papers - [41], [146], [229], [104], [15], and [158].

We also did a thorough search of papers which cite CG (2009). Our search was focused on papers published after 2009, but we also included earlier papers that either (1) were highly cited and relevant to our question or (2) directly cited earlier versions of CG [118]. We did not do a back search of papers from these seminal papers since we wanted to focus more on the effect the results from CG have on the narrative of gender differences. We complemented our literature review with Google Scholar searches using the following keywords and combinations thereof: "gender", "experiment", "economics", "gender difference", "female", "context", "social cues".

The focus was on published papers, but we also included relevant working papers released after 2009. The stopping rule here was when we started finding the same papers repeatedly. The search of papers was done independently by two of the authors. We ultimately compiled our results on a spreadsheet and removed duplicates to create a single pool of relevant papers.

D.1 Review Process

For the review, we grouped the papers based on the journals in which they were published (top 5 vs others) and the number of citations (100+, 50+, others). We began the review process with papers published in the top 5 journals and moved down based on the number of citations (100+, then 50+, then all others). We reviewed all papers we compiled in the pool and did not exclude any papers at any stage. For each paper we identified the key question, summarized the design and listed out the main result. Next, we classified papers based on the topic they were focusing on based on categories we created at the beginning of the review process:

- Competition - Entry
- Competition - Performance
- Competition - Performance and Entry
- Negotiation - Entry
- Negotiation - Performance
- Non-promotable tasks
- Speaking out
- Other-regarding preferences
- Risk preferences
- Time preferences
- Discrimination
- NA
- Combination

We delved deeper into each paper from the lenses of responsiveness to social cues or experimental context or some combination. For this part we first determined if any element of the question, or design, or analysis of the data can speak responsiveness by gender – a paper was characterized as "NA" if we didn't find any responsiveness element. We then moved to classify the rest of the papers based on whether we find either men or women or both to be responsive changes in experimental context, social cues or both. A paper was classified as say Men $>$ Women when either there was a single result in the paper pointing in this direction or all results were pointing in the same direction. On the other hand, if there were multiple results in the paper pointing in different directions, the paper was classified as Men $\sim$ Women. A final categorization here was "Insufficient Information" which was used when the data presented in the paper was insufficient to determine the direction of responsiveness.

The final step of the review process was an easy determination of whether the paper cites the CG results on gender difference in sensitivity to context and if so do they side with the result or argue against it. And finally, we identified any paper that directly tests this CG result. We compiled all this information in a spreadsheet and two co-authors of this paper pre-decided these aspects and each reviewed about 50% of the pooled papers.

The first pool of papers was compiled over the summer of 2018 and the review process was completed over the fall semester of the same year. The pool of papers was appended in the summer of 2021 to ensure that the pool is up to date and the review for any new papers was completed simultaneously. While reviewing the new papers we randomly checked review for some of the existing papers to make sure there is no gap in our understanding and no revisions were required.

Appendix E Chapter 2: Analysis of Data from DellaVigna and Pope (2022)

This appendix provides details about the behavioral treatments in [139] that were used to create Figures 5 and 6 in the main text. Table C1 describes the treatments that were used in our pairwise comparisons. The information comes from Table 1 in the original paper. The wording on each one of the treatments differs from the original table to reflect the fact that we do not use data from two of the conditions (no consent form and effort after the first 20 minutes) due to the absence of gender information. As such, the wording for each treatment does not include the variations implemented on those two conditions. The main text of the treatments refers to sessions of the typing task, while the text in square brackets refers to sessions with the WWII coding task.

We conducted pairwise treatment comparisons within each of the five categories (piece rates, social preferences, discounting, probability weighting, and psychological manipulations). Additionally, we also conduct pairwise tests between a few treatments in different categories. For example, we compare mean effort between treatments 2 (1-cent piece rate now), 8 (1-cent piece rate in 2 weeks), and 9 (1-cent piece rate in 4 weeks). The complete list of all pairwise comparisons is:

- Pairwise comparisons of treatments 1 to 5 for each of the two real-effort tasks, for a total of 20 mean-difference tests for both men and women.
- Comparison of treatments 6 and 7 for each of the two real-effort tasks, for a total of 2 mean-difference tests for both men and women.
- Pairwise comparisons of treatments 2, 8, and 9, for each of the two real-effort tasks, for a total of 6 mean-difference tests for both men and women.
- Comparison of treatments 10 and 11 for each of the two real-effort tasks, for a total of 2 mean-difference tests for both men and women.
- Pairwise comparisons of treatments 12 to 16 for each of the two real-effort tasks, for a total of 20 mean-difference tests for both men and women.
- Comparison of treatments 2 and 16 for each of the two real-effort tasks, for a total of 2 mean-difference tests for both men and women.

- Comparison of treatments 2 and 6 for each of the two real-effort tasks, for a total of 2 mean-difference tests for both men and women.
- Comparison of treatments 4 and 7 for each of the two real-effort tasks, for a total of 2 mean-difference tests for both men and women.

We thus have a total of 56 treatment comparisons for each gender. Since we are interested in comparing men and women's responsiveness to changes in experiment conditions.

Table E1: Treatment variations in [139]

Category			**Treatments**
	Description	Code	Wording
Piece rate	No piece rate	1	"Your score [The number of cards you complete] will not affect your payment in any way."
	1-cent piece rate	2	"As a bonus, you will be paid an extra 1 cent for every 100 points that you score [2 cards that you complete]"
	4-cent piece rate	3	"As a bonus, you will be paid an extra 4 cents for every 100 points that you score [2 cents for every card that you complete]."
	10-cent piece rate	4	"As a bonus, you will be paid an extra 10 cents for every 100 points that you score [5 cents for every card that you complete].
	0.1-cent piece rate	5	"As a bonus, you will be paid an extra 1 cent for every 1,000 points that you score [20 cards you complete]."

Category			**Treatments**
	Description	Code	Wording
Social preferences	1-cent piece rate for Red Cross	6	"As a bonus, the Red Cross charitable fund will be given 1 cent for every 100 points that you score [2 cards you complete]."
	10-cent piece rate for Red Cross	7	"As a bonus, the Red Cross charitable fund will be given 10 cents for every 100 points that you score [5 cents for every card you complete]."
Discounting	1-cent piece-rate in two weeks	8	"As a bonus, you will be paid an extra 1 cent for every 100 points that you score [every 2 cards you complete]. This bonus will be paid to your account two weeks from today."
	1-cent piece-rate in four weeks	9	"As a bonus, you will be paid an extra 1 cent for every 100 points that you score [every 2 cards you complete]. This bonus will be paid to your account four weeks from today."
Probability weighting	1% chance of 1-dollar piece rate	10	"As a bonus, you will have a 1 percent chance of being paid an extra $1 for every 100 points that you score [extra 50 cents for every card you complete]."
	50% chance of 2-cents piece rate	11	"As a bonus, you will have a 50 percent chance of being paid an extra 2 cents for every 100 points that you score [extra 1 cent for every card you complete]."

Category			**Treatments**
	Description	Code	Wording
Psychological treatments	Gift exchange	12	"In appreciation to you for performing this task, you will be paid a bonus of 40 cents. Your score will not affect your payment in any way [The number of cards you complete will not affect your payment in any way]."
	Social comparison	13	"Your score [The number of cards you complete] will not affect your payment in any way. In a previous version of this task, many participants were able to score more than 2,000 points [completed more than 70 cards]."
	Ranking	14	"Your score [The number of cards you complete] will not affect your payment in any way. After you play, we will show you how well you did [how many cards you completed] relative to other participants who have previously done this task."
	Task significance	15	"Your score [The number of cards you complete] will not affect your payment in any way [, but your work is very valuable for us, and we would really appreciate your help]. We are interested in how fast people choose to press digits and we would like you to do your very best. So please try as hard [do as many] as you can."

Category		**Treatments**	
	Description	Code	Wording
	Task significance and 1-cent piece rate	16	"We are interested in how fast people choose to press digits and we would like you to do your very best [Your work is very valuable for us, and we would really appreciate your help]. So please try as hard [do as many cards] as you can. As a bonus, you will be paid an extra 1 cent for every 100 points that you score [2 cards you complete]."

Appendix F Chapter 3: Results Formerly in Main Text

Table F1: Results Summary

	Lab	**V-Lab**	**MTurk**	**Cloud-R**	**Prolific**
Panel A:	Avg	Avg	Avg	Avg	Avg
Σ-Dominated:	0.108 (0.036)	0.162 (0.043)	0.369 (0.021)	0.120 (0.014)	0.122 (0.017)
i-Dominant ($DDCC$):	0.324 (0.054)	0.270 (0.052)	0.159 (0.016)	0.240 (0.018)	0.260 (0.022)
Rapoport identifier ($DCCC$):	0.189 (0.046)	0.176 (0.045)	0.093 (0.012)	0.129 (0.014)	0.106 (0.016)
Full Cooperator ($CCCC$):	0.284 (0.052)	0.338 (0.055)	0.297 (0.020)	0.447 (0.021)	0.416 (0.025)
Σ-Dominant:	0.892 (0.036)	0.838 (0.043)	0.631 (0.021)	0.880 (0.014)	0.878 (0.017)
Rapoport ordered:	0.905 (0.034)	0.919 (0.032)	0.828 (0.016)	0.917 (0.012)	0.886 (0.016)
Both:	0.797 (0.047)	0.784 (0.048)	0.549 (0.021)	0.817 (0.017)	0.782 (0.021)
Panel B:	Δ_{Frame}	Δ_{Frame}	Δ_{Frame}	Δ_{Frame}	Δ_{Frame}
Σ-Dominated:	-0.037 (0.073)	0.192 (0.090)	0.163 (0.044)	0.008 (0.030)	0.052 (0.037)
i-Dominant ($DDCC$):	0.075 (0.118)	-0.092 (0.111)	0.028 (0.034)	0.042 (0.040)	0.056 (0.048)
Rapoport identifier ($DCCC$):	-0.033 (0.095)	0.048 (0.096)	-0.048 (0.024)	0.012 (0.031)	0.056 (0.047)
Full Cooperator ($CCCC$):	0.012 (0.112)	-0.192 (0.117)	-0.046 (0.041)	-0.023 (0.046)	-0.070 (0.052)
Σ-Dominant:	0.037 (0.073)	-0.192 (0.090)	-0.163 (0.044)	-0.008 (0.030)	-0.052 (0.037)
Rapoport ordered:	0.017 (0.071)	-0.065 (0.068)	0.049 (0.032)	0.025 (0.026)	-0.018 (0.035)
Both:	0.053 (0.096)	-0.235 (0.100)	-0.066 (0.045)	0.031 (0.036)	-0.040 (0.045)

Notes - Standard errors for proportions in parentheses.

Table F1 provides average outcomes across the five samples with standard errors derived from simple tests of proportion. In Panel A we first outline the proportion of individuals

with particular focal behaviors over the four games (pooling data across the frame), then outline the relative effects across the re-framing in Panel B.

The first row in Panel A of Table F1 shows the rate at which individuals in the experiment make an obvious mistake with respect to the offered incentives i.e., choose the Σ-dominated actions. The proportion of participants choosing the σ-dominated actions is statistically inseparable between the lab, v-lab, Cloud-R, and Prolific samples with approximately 12% of participants make a defect choices in the last two games.[1] In contrast, for the MTurk sample this rate grows to more than one-in-three, significantly different from all other samples.[2] Moreover, as we explain next, even this number is perhaps an underestimate of the fraction of participants making choices orthogonal to the incentives.

Where panel A in Table F1 provides the overall average results by population sample (pooling across both the C-first and D-first frames), Panel B indicates the change in the proportion across the re-frame. The first row of Panel B shows the change in the participant proportion exhibiting a Σ-dominated choice when we move from listing C to listing D as the first action. Our results across the re-frame show that the lab sample moves in the opposite direction from a first-option bias with a slight decrease in Σ-dominance when the D action is listed first (though this is not significant, $p = 0.640$). The first-option bias is the smallest for the Cloud-R sample (0.8 percentage points with $p = 7894$). The Prolific sample does show a movement 5.2 percentage point movement, where 15.6 percent of choices in the D-first sample are Σ-dominated choices. Though this difference is not significant ($p = 0.160$) but if we allowed for a one-sided test there is marginal evidence for a small first-action bias on Prolific. The largest effects though are in the V-lab and MTurk sample,s where listing the D-action first leads to a 19.2 and 16.3 percentage points increase in the Σ-dominated fraction respectively ($p = 0.37$ and $p < 0.001$ on a test of proportions respectively for v-lab and MTurk).[3]

In the worst-case D-first treatment 47.8 percent of the MTurk choices are Σ-dominated.

[1]The pairwise p-values for the test of proportions for {lab vs. v-lab, lab vs. Cloud-R, lab vs. Prolific, v-lab vs. Cloud-R, v-lab vs. Prolific, Cloud-R vs. Prolific} are {0.3395, 0.7644, 0.735, 0.3065, 0.3465, 0.9294} respectively.

[2]$p < 0.001$ for the pair-wise tests of proportions between MTurk and all four populations

[3]The bottom section of Panel B in Table F1 indicates that the re-framing has a consistent effect in increasing the selection of D in the v-lab and MTurk samples when this action is listed first.

Despite successfully passing the screen questions—where participants must demonstrate their understanding of the game incentives or be kicked out—approximately one half of the MTurk sample then make choices that indicate little awareness of the induced games. While approximately a third of this effect can be attributed to participants choosing the D action in games Σ-DOM1 and Σ-DOM2 simply because it is the first-listed option, the result still indicates that just under half of the sample are making choices that are orthogonal to the offered incentives. In contrast, despite similar costs per observation on Cloud-R and Prolific, the rates of such mistakes in these populations seems to be at most 15 percent, and we lack statistical power to say that it is even different from the laboratory.

Appendix G Chapter 3: CloudResearch Approved List (Cloud-R) Robustness Sessions: Extended Response

Table G1: Experimental Games: Robustness Sample

PD1 game:

	C	D
C	21,21	2,28
D	28,2	8,8

PD2 game:

	C	D
C	19,19	8,22
D	22,8	9,9

PD3 game:

	C	D
C	14,14	5,25
D	25,5	13,13

PD4 game:

	C	D
C	18,18	3,27
D	27,3	12,12

Σ-DOM1 game:

	C	D
C	17,17	12,16
D	16,12	10,10

Σ-DOM2 game

	C	D
C	15,15	16,10
D	10,16	11,11

Table G2: Cloud-R Participants per treatment

	Cloud-R	Cloud-R-Robustness
Main	374	165
Re-frame	167	

Notes - Excludes participants who did not answer the comprehension question correctly.

Table G3: Behavior Across Cloud-R Samples: Cooperation

Game	Robustness Sample	p-value	Original Sample
PD1	0.485 (0.039)	0.107	0.556 (0.021)
PD2	0.600 (0.038)	0.483	0.630 (0.021)
Σ-DOM1	0.964 (0.015)	0.174	0.935 (0.011)
Σ-DOM2	0.939 (0.019)	0.560	0.926 (0.011)
PD3	0.339 (0.037)		
PD4	0.406 (0.038)		

Notes - Standard error for the proportion in parentheses. All p values are for two-sided tests of equality between the samples.

Table G4: Subject Types Across Cloud-R Samples: Pooled Data

Type	Robustness Sample	p-value	Original Sample
Choice Profiles in Original 4 Games:			
Nash ($DDCC$)	0.297 (0.036)	0.143	0.24 (0.018)
Uncond Coop ($CCCC$)	0.400 (0.038)	0.284	0.447 (0.021)
Cond Coop ($DCCC$ & $CDCC$)	0.230 (0.033)	0.286	0.192 (0.017)
Σ-dominated	0.073 (0.02)	0.087	0.12 (0.014)

Notes - Standard error for the proportion in parentheses. All p values are for two-sided tests of equality between the populations. Choice profiles are given in order of the Rapoport ratio in the PD games (so PD1, PD2-Σ-DOM1, Σ-DOM2 in this table).

Table G5: Additional Subject Types in Cloud-R Robustness Sample

Type	Robustness Sample
Σ-dominant	0.927 (0.020)
Rapoport ordered	0.812 (0.031)
Both	0.770 (0.033)
Σ-dominant profiles:	
Nash, $DDDDCC$	0.261 (0.034)
$DDDCCC$	0.121 (0.025)
$DDCCCC$	0.073 (0.020)
$DCCCCC$	0.061 (0.019)
Uncond Coop, $CCCCCC$	0.255 (0.034)
Non-Rapoport ordered (11 profiles)	0.158 (0.028)

Notes - Standard error for the proportion in parentheses. Choice profiles ar given in order of the Rapoport ratio in the PD games (so PD3, PD4, PD1, PD2-Σ-DOM1, Σ-DOM2 in this table).

Appendix H Chapter 3: Prolific Robustness Sessions: Extended Response

Table H1: Prolific Participants per treatment

	Prolific	Prolific-Robustness
Main	250	125
Re-frame	135	

Notes - Excludes participants who did not answer the comprehension question correctly.

Table H2: Behavior Across Prolific Samples: Cooperation

Game	Robustness Sample	p-value	Original Sample
PD1	0.488 (0.045)	0.127	0.566 (0.025)
PD2	0.584 (0.044)	0.885	0.577 (0.025)
Σ-DOM1	0.904 (0.026)	0.865	0.909 (0.015)
Σ-DOM2	0.952 (0.019)	0.623	0.94 (0.012)
PD3	0.320 (0.042)		
PD4	0.328 (0.042)		

Notes - Standard error for the proportion in parentheses. All p values are for two-sided tests of equality between the samples.

Table H3: Subject Types Across Prolific Samples: Pooled Data

Type	Robustness Sample	p-value	Original Sample
Choice Profiles in Original 4 Games:			
Nash ($DDCC$)	0.312 (0.042)	0.255	0.260 (0.022)
Uncond Coop ($CCCC$)	0.352 (0.043)	0.208	0.416 (0.025)
Cond Coop ($DCCC$ & $CDCC$)	0.208 (0.036)	0.897	0.203 (0.021)
Σ-dominated	0.128 (0.030)	0.862	0.122 (0.017)

Notes - Standard error for the proportion in parentheses. All p values are for two-sided tests of equality between the populations. Choice profiles are given in order of the Rapoport ratio in the PD games (so PD1, PD2-Σ-DOM1, Σ-DOM2 in this table).

Table H4: Additional Subject Types in Prolific Robustness Sample

Type	Robustness Sample
Σ-dominant	0.872 (0.030)
Rapoport ordered	0.808 (0.035)
Both	0.728 (0.040)
Σ-dominant profiles:	
Nash, $DDDDCC$	0.272 (0.040)
$DDDCCC$	0.120 (0.029)
$DDCCCC$	0.088 (0.025)
$DCCCCC$	0.072 (0.023)
Uncond Coop, $CCCCCC$	0.176 (0.034)
Non-Rapoport ordered (11 profiles)	0.144 (0.032)

Notes - Standard error for the proportion in parentheses. Choice profiles ar given in order of the Rapoport ratio in the PD games (so PD3, PD4, PD1, PD2-Σ-DOM1, Σ-DOM2 in this table).

Appendix I Chapter 3: Experiments Instructions and Screenshots

I.1 Instructions for Main Lab Treatment

Your earnings in today's experiment will depend on your decisions, the decisions of others in the room, and on chance. Any money you make will be paid privately and in cash at the end of the experiment. We will start with a brief description of your task today. If you have any questions, please raise your hand and we will come to answer you in private.

Explanation of your task

There are four rounds in today's study, each consisting of a decision table. Your task will be to choose one option from two alternatives for each decision table. A round will end when all participants submit their choices.

At the end of the fourth round, the computer will randomly and anonymously pair you with another participant in the room. Next, the computer will randomly select one of your four rounds. You will be paid for that round based on you and the matched participant's choices in that round. Your final earnings will then consist of payoff from this one round and a participation fee of $6.

Every round is equally likely to be selected for payment, so you should treat each round as if it determines your final payment. Also, there are only four decisions in this study, so you should consider them carefully.

Description of a Decision Table

Below is an example decision table: Both you and the matched participant make choices between Option A and Option B. The decision table indicates the payout for you and the other participant for each possible combination of choices.

Suppose this decision table was selected for payment, then in addition to the participation fee:

(1) if both participants choose A, they each receive $18;

(2) if you choose A and the matched participant chooses B, then you receive $6, and they receive $15;

Figure I1: Screenshot of decision table

Your Decision	Other's Decision	Your Payoff	Other's Payoff
A	A	$18	$18
A	B	$6	$15
B	A	$15	$6
B	B	$10	$10

(3) Vice versa if you choose B and the matched participant chooses A, then you receive $15, and they receive $6.

(4) if both participants choose B, they each receive $10;

We will begin the study with a few questions about your understanding of the decision table and then proceed to the first round.

I.2 Instructions for Re-framed Lab Treatment

[Introductory instructions and section with "Explanation of your task" were identical to I]

Description of a Decision Table

Below is an example decision table:

Figure I2: Screenshot of decision table

Your Decision	Other's Decision	Your Payoff	Other's Payoff
A	A	$10	$10
A	B	$15	$6
B	A	$6	$15
B	B	$18	$18

Both you and the matched participant make choices between Option A and Option B. The decision table indicates the payout for you and the other participant for each possible combination of choices.

Suppose this decision table was selected for payment, then in addition to the participation fee:

(1) if both participants choose A, they each receive $10;

(2) if you choose A and the matched participant chooses B, then you receive $15, and they receive $6;

(3) Vice versa if you choose B and the matched participant chooses A, then you receive $6, and they receive $15.

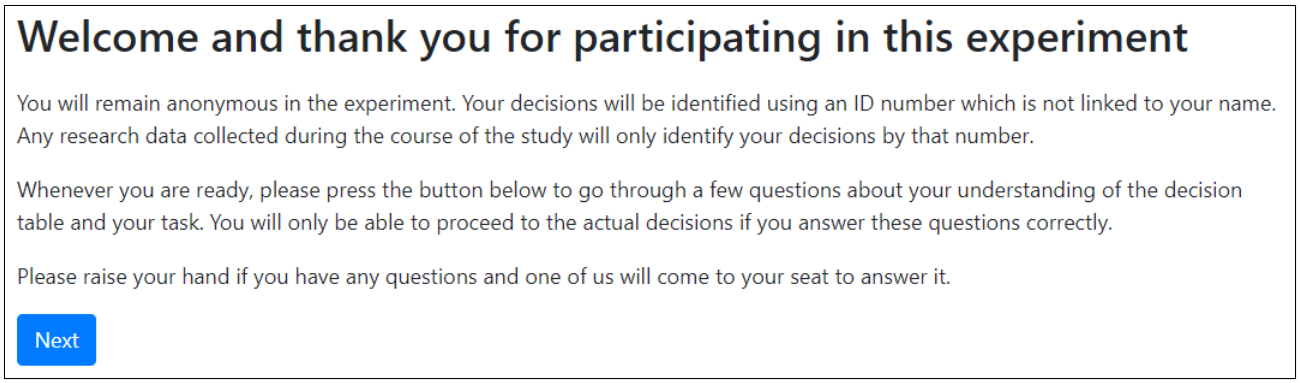
(4) if both participants choose B, they each receive $18;

We will begin the study with a few questions about your understanding of the decision table and then proceed to the first round.

I.3 Screenshots of the Laboratory Experiment

Following are the screenshots of the lab experiment for the main sample. The screens for the re-framed sample were identical except that the labels of options on the decision table reversed.

Figure I3: Screenshot of lab experiment - welcome screen



[For the re-framed sample option A corresponded to D and option B corresponded to C. The answers to the comprehension questions changed accordingly. Participants couldn't move forward without answering these questions correctly.]

[Rounds 2, 3 and 4 screens were the same as round 1 with different decision tables. For the re-framed sample option A corresponded to D and option B corresponded to C, the screens were otherwise the same as the main sample. The four decision tables were presented to the participants in random order.]

Figure I4: Screenshot of lab experiment - comprehension check

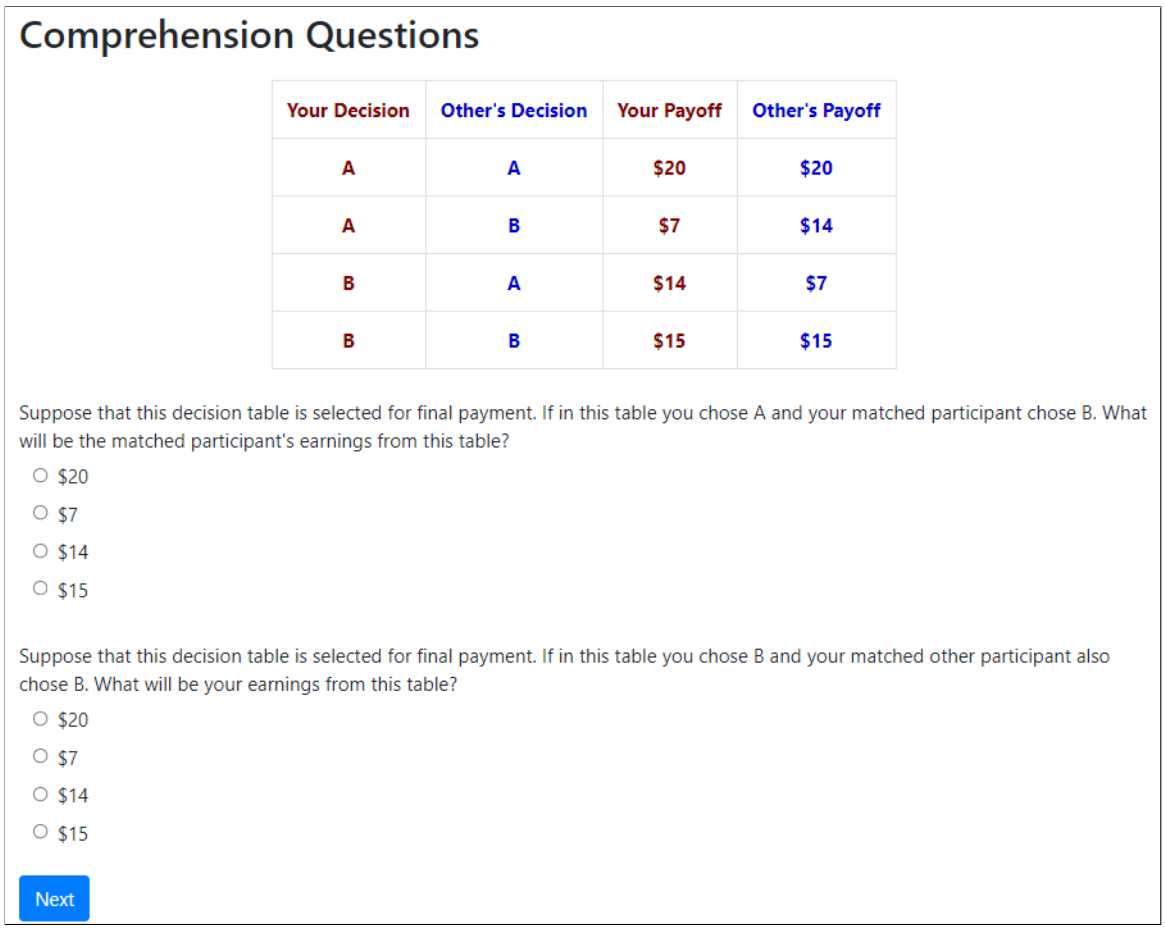


Figure I5: Screenshot of lab experiment - decision screen

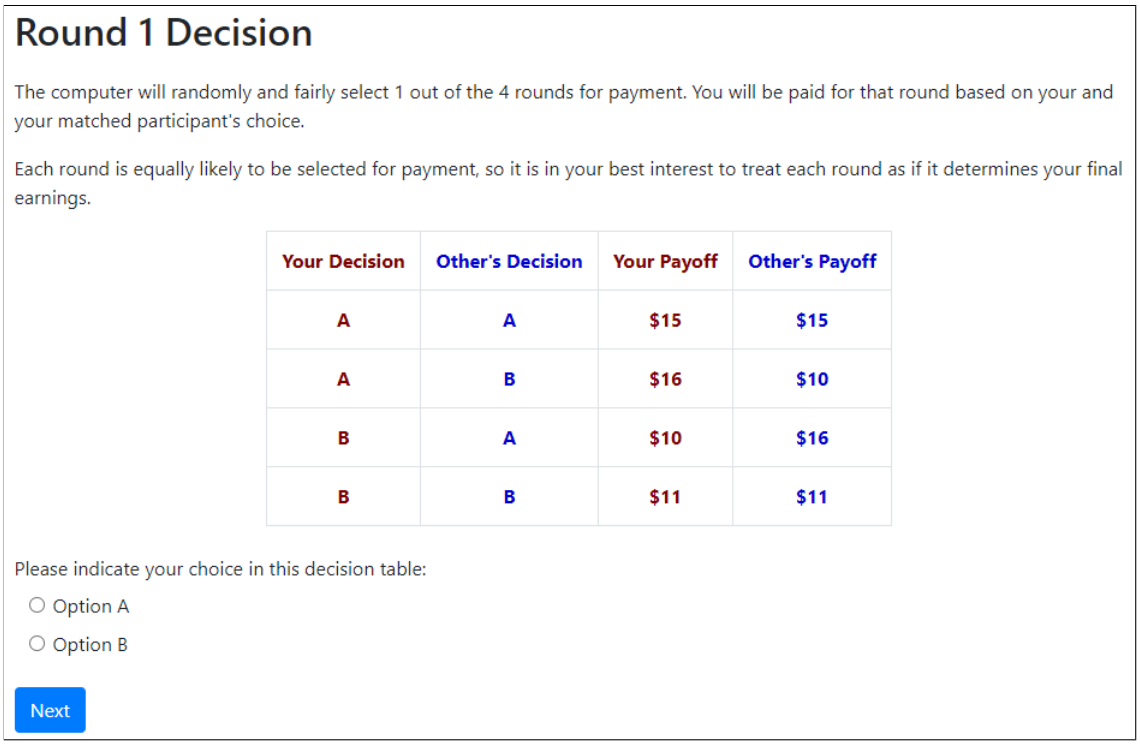


Figure I6: Screenshot of lab experiment - exit survey

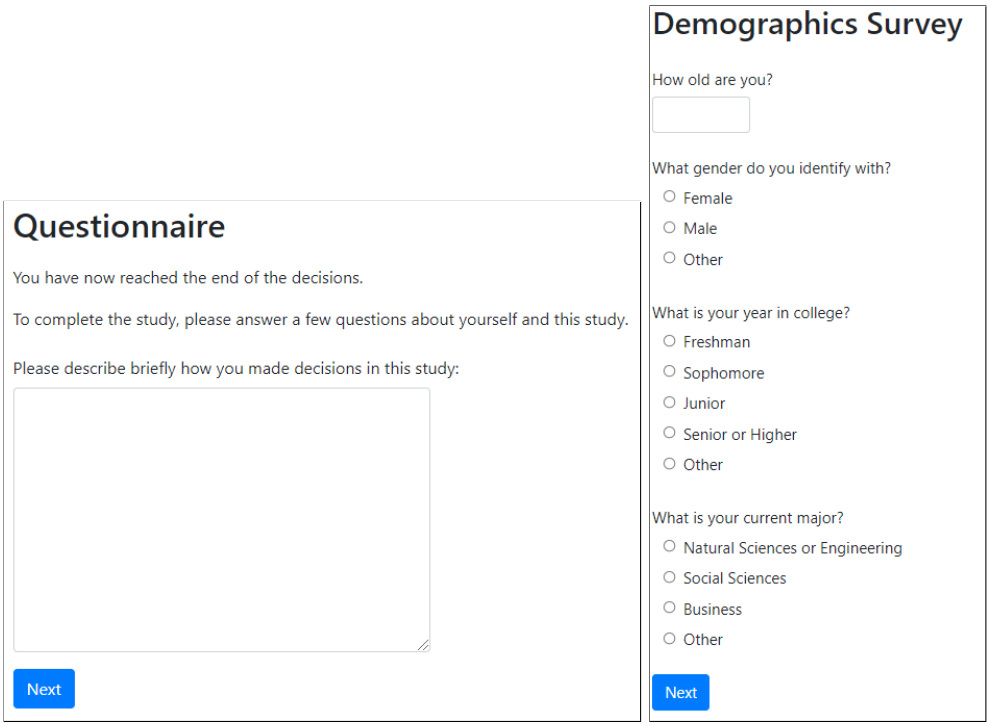


Figure I7: Screenshot of lab experiment - payment instructions

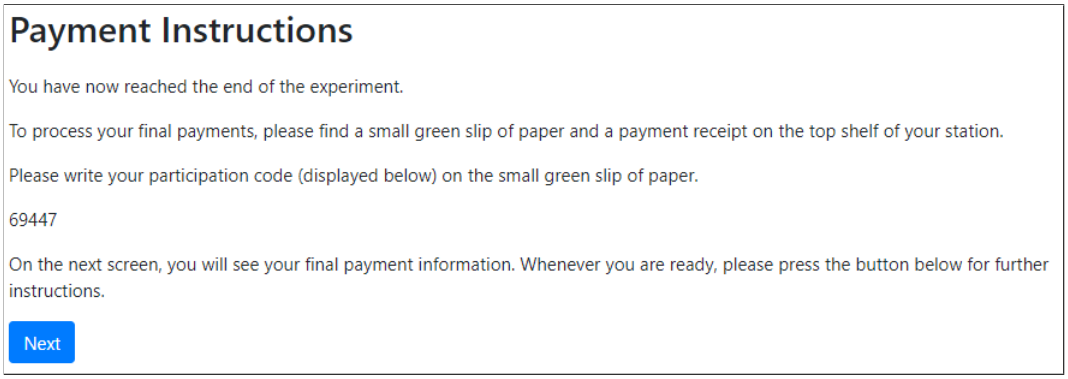
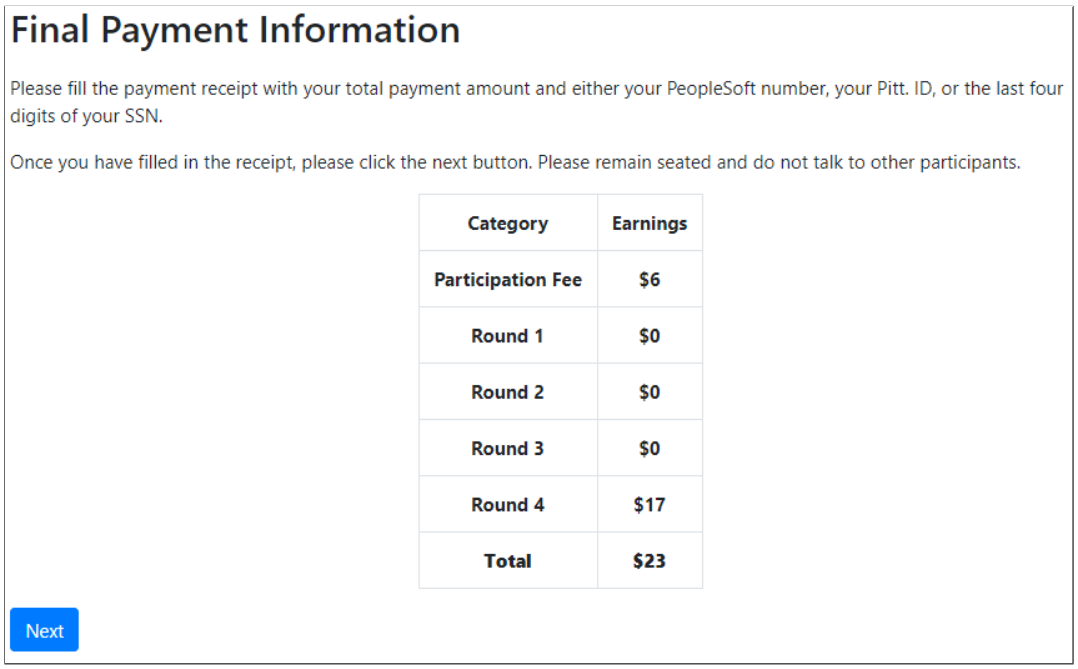


Figure I8: Screenshot of lab experiment - final payment screen

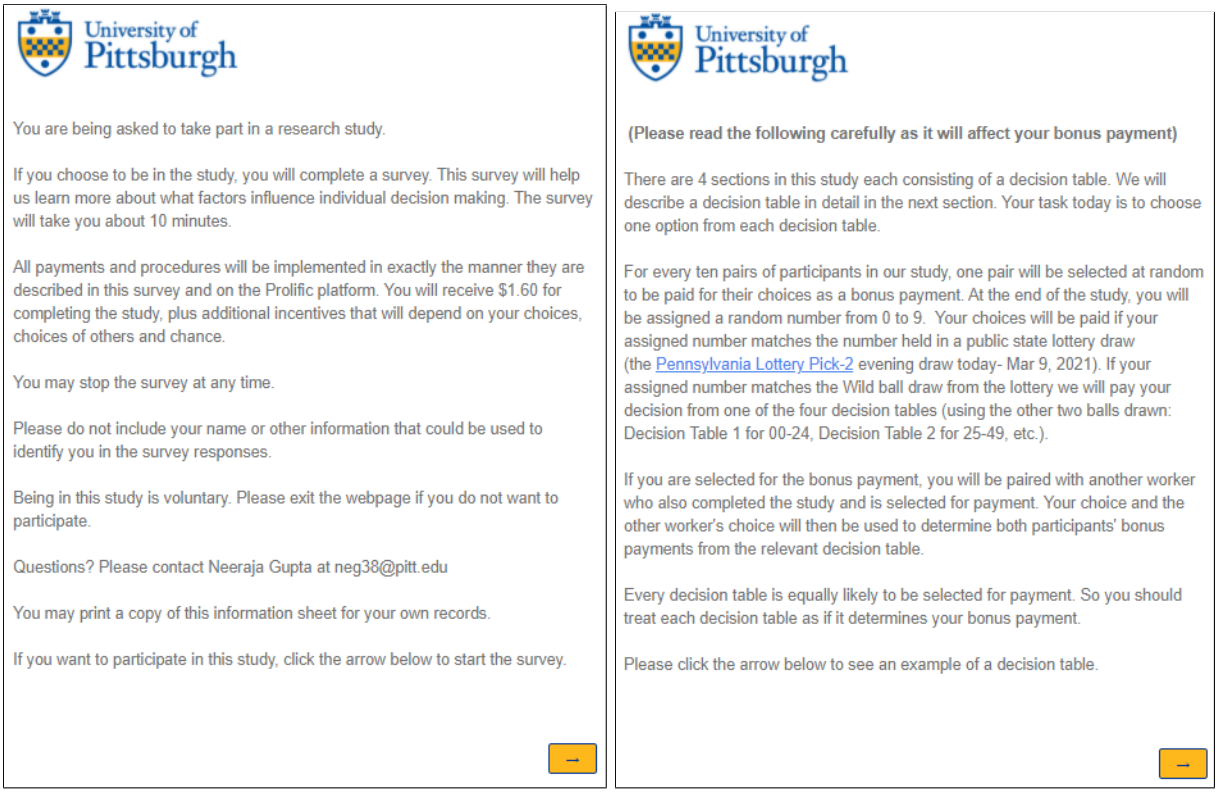


[Participants were then invited to the payment room one by one and paid in cash in private.]

I.4 Instructions for Online Experiment

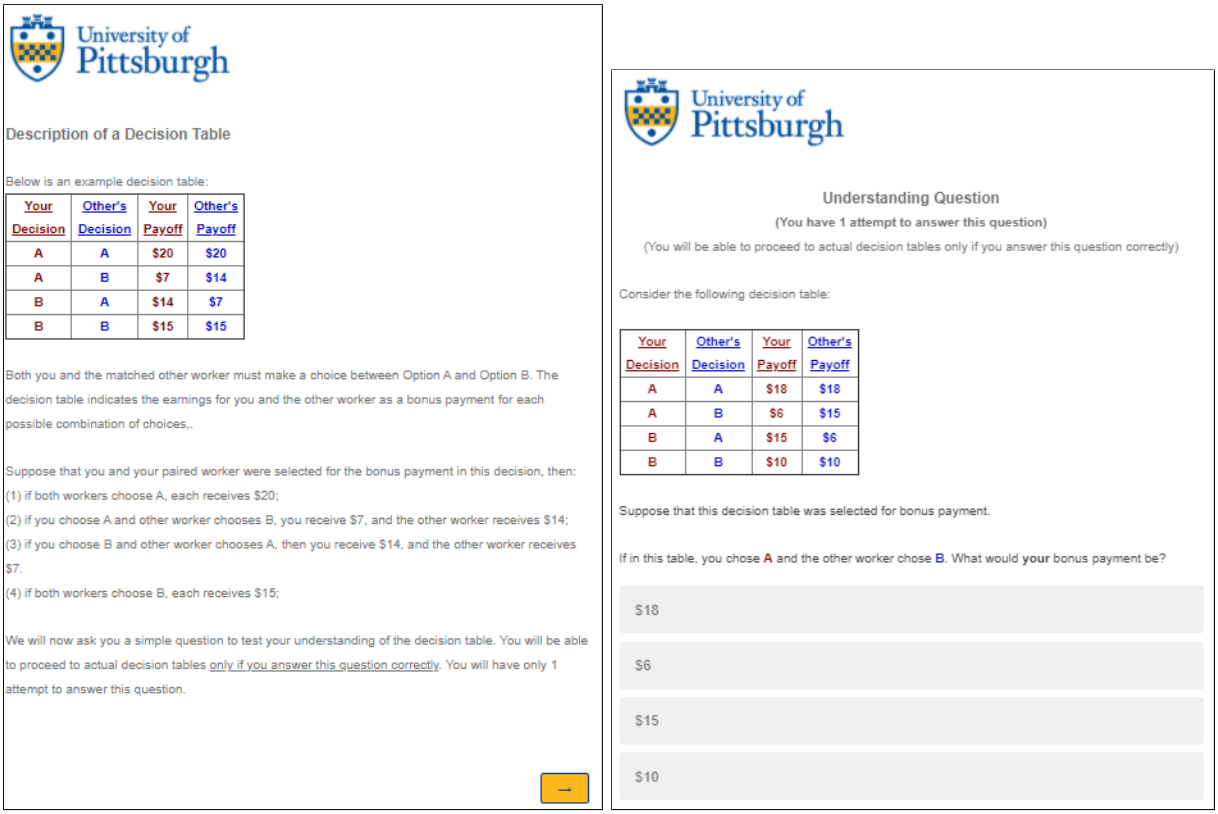
Following are the screenshots of the online experiment for the main Prolific sample. The screens for the MechTurk sample were the same as the Prolific sample.

Figure I9: Screenshot of online experiment - welcome and instructions



[For the re-framed sample option A corresponded to D and option B corresponded to C. The answers to the comprehension questions changed accordingly. Participants were dismissed with the show-up of \$1.60 for answering the comprehension question incorrectly on Prolific (\$0.50 on MechTurk). On MechTurk, participants who answered the comprehension question correctly were offered additional \$0.50.]

Figure I10: Screenshot of online experiment - decision table



[Next, the four decision tables were presented to the participants in random order. For the re-framed sample option A corresponded to D and option B corresponded to C, the screens were otherwise the same as the main sample.]

[Fixed fees were credited to the participants immediately upon approval of the submission and the bonus payments were made within 24 hours of completion.]

Figure I11: Screenshot of online experiment - decision screen

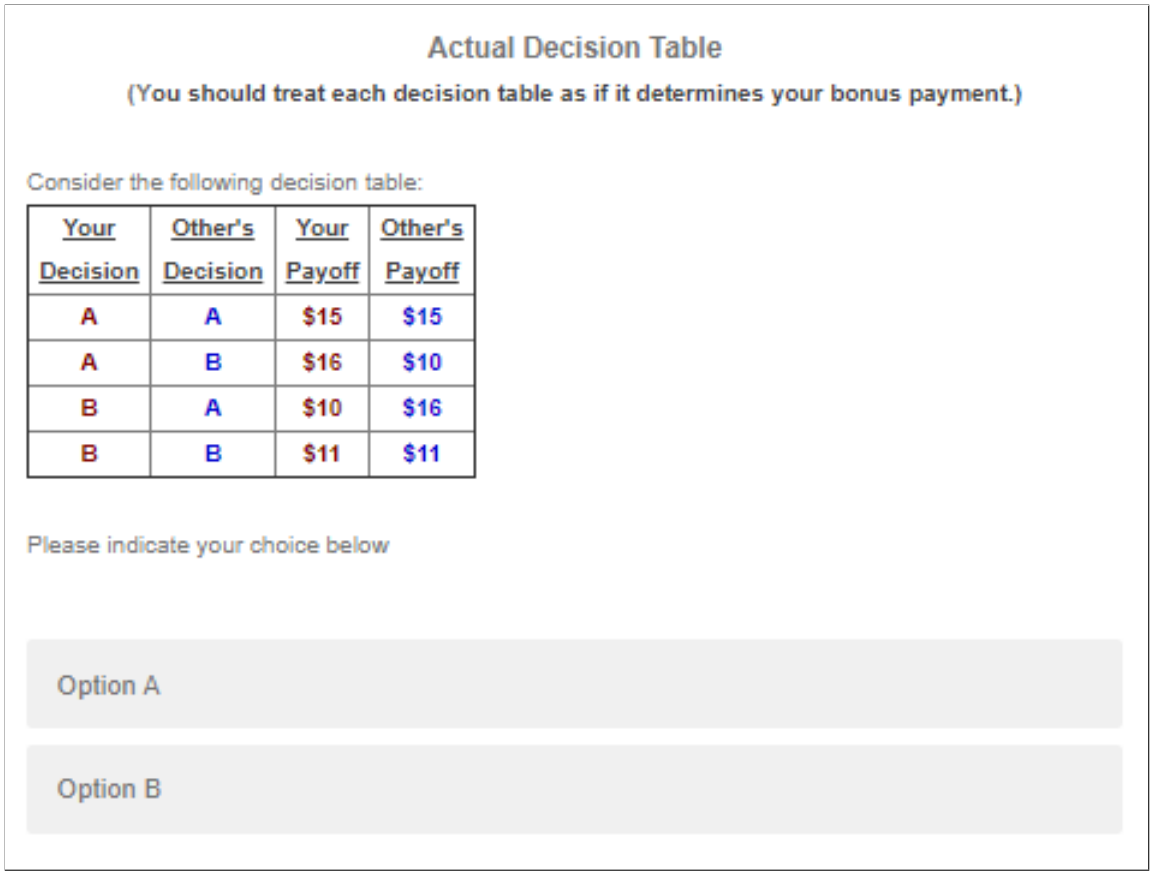
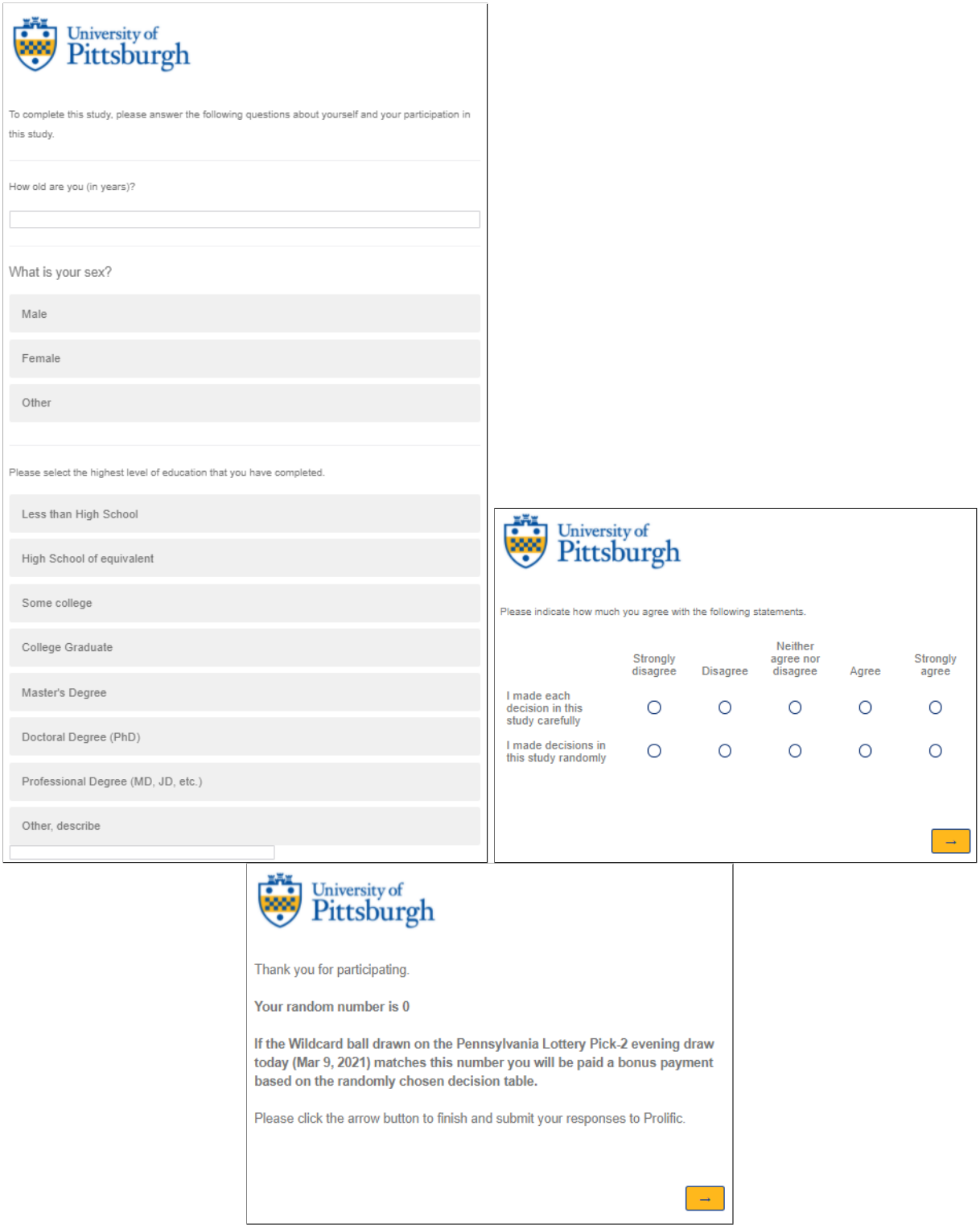


Figure I12: Screenshot of online experiment - exit survey and instructions



Bibliography

[1] Alberto Abadie, Susan Athey, Guido W Imbens, and Jeffrey M Wooldridge. When should you adjust standard errors for clustering? *The Quarterly Journal of Economics*, 138(1):1–35, 2022.

[2] Julie R Agnew, Lisa R Anderson, Jeffrey R Gerlach, and Lisa R Szykman. Who chooses annuities? an experimental investigation of the role of gender, framing, and defaults. *American Economic Review*, 98:418–422, 2008.

[3] Anjali Agrawal, Ellen P Green, and Lisa Lavergne. Gender effects in the credence goods market: An experimental study. *Economics Letters*, 174:195–199, 2019.

[4] Fernando Aguiar, Pablo Brañas-Garza, Ramón Cobo-Reyes, Natalia Jimenez, and Luis M Miller. Are women expected to be more generous? *Experimental Economics*, 12:93–98, 2009.

[5] Dennis J Aigner and Glen G Cain. Statistical theories of discrimination in labor markets. *Ilr Review*, 30:175–187, 1977.

[6] Sule Alan and Seda Ertac. Mitigating the gender gap in the willingness to compete: Evidence from a randomized field experiment. *Journal of the European Economic Association*, 17:1147–1185, 2019.

[7] Ingvild Almås, Alexander W Cappelen, Kjell G Salvanes, Erik Ø Sørensen, and Bertil Tungodden. What explains the gender gap in college track dropout? experimental and administrative evidence. *American Economic Review*, 106(5):296–302, 2016.

[8] Johan Almenberg and Anna Dreber. Gender, stock market participation and financial literacy. *Economics Letters*, 137:140–142, 2015.

[9] Emily T Amanatullah and Michael W Morris. Negotiating gender roles: Gender differences in assertive negotiating are mediated by women's fear of backlash and attenuated when negotiating on behalf of others. *Journal of personality and social psychology*, 98:256, 2010.

[10] Steffen Andersen, Erwin Bulte, Uri Gneezy, and John A List. Do women supply more public goods than men? preliminary experimental evidence from matrilineal and patriarchal societies. *American Economic Review*, 98:376–381, 2008.

[11] Steffen Andersen, Seda Ertac, Uri Gneezy, John A List, and Sandra Maximiano. Gender, competitiveness, and socialization at a young age: Evidence from a matrilineal and a patriarchal society. *Review of Economics and Statistics*, 95:1438–1443, 2013.

[12] Per A Andersson, Arvid Erlandsson, Daniel Västfjäll, and Gustav Tinghög. Prosocial and moral behavior under decision reveal in a public environment. *Journal of Behavioral and Experimental Economics*, 87:101561, 2020.

[13] James Andreoni, Amalia Di Girolamo, John A List, Claire Mackevicius, and Anya Samek. Risk preferences of children and adolescents in relation to gender, cognitive skills, soft skills, and executive functions. *Journal of economic behavior organization*, 179:729–742, 2020.

[14] James Andreoni and Ragan Petrie. Beauty, gender and stereotypes: Evidence from laboratory experiments. *Journal of Economic Psychology*, 29:73–93, 2008.

[15] James Andreoni and Lise Vesterlund. Which is the fair sex? gender differences in altruism. *The Quarterly Journal of Economics*, 116:293–312, 2001.

[16] Kate Antonovics, Peter Arcidiacono, and Randall Walsh. The effects of gender interactions in the lab and in the field. *The Review of Economics and Statistics*, 91:152–162, 2009.

[17] Kate Antonovics and Ben Backes. The effect of banning affirmative action on college admissions policies and student quality. *Journal of Human Resources*, 49:295–322, 2014.

[18] Kate L Antonovics and Richard H Sander. Affirmative action bans and the "chilling effect". *American law and economics review*, 15:252–299, 2013.

[19] Jose Apesteguia, Ghazala Azmat, and Nagore Iriberri. The impact of gender composition on team performance and decision making: Evidence from the field. *Management Science*, 58:78–93, 2012.

[20] Coren L Apicella, Elif E Demiral, and Johanna Mollerstrom. No gender difference in willingness to compete when competing against self. *American Economic Review*, 107:136–140, 2017.

[21] Coren L Apicella and Anna Dreber. Sex differences in competitiveness: Hunter-gatherer women and girls compete less in gender-neutral and male-centric tasks. *Adaptive Human Behavior and Physiology*, 1:247–269, 2015.

[22] Coren L Apicella, Anna Dreber, Peter B Gray, Moshe Hoffman, Anthony C Little, and Benjamin C Campbell. Androgens and competitiveness in men. *Journal of Neuroscience, Psychology, and Economics*, 4:54, 2011.

[23] Felipe A Araujo, Erin Carbone, Lynn Conell-Price, Marli W Dunietz, Ania Jaroszewicz, Rachel Landsman, Diego Lamé, Lise Vesterlund, Stephanie W Wang, and Alistair J Wilson. The slider task: an example of restricted inference on incentive effects. *Journal of the Economic Science Association*, 2(1):1–12, 2016.

[24] Antonio A Arechar, Simon Gächter, and Lucas Molleman. Conducting interactive experiments online. *Experimental economics*, 21(1):99–131, 2018.

[25] Dan Ariely, Uri Gneezy, George Loewenstein, and Nina Mazar. Large stakes and big mistakes. *The Review of Economic Studies*, 76:451–469, 2009.

[26] Kenneth Arrow. The theory of discrimination, 10 1971.

[27] Mara S Aruguete, Ho Huynh, Blaine L Browne, Bethany Jurs, Emilia Flint, and Lynn E McCutcheon. How serious is the 'carelessness' problem on mechanical turk? *International Journal of Social Research Methodology*, 22(5):441–449, 2019.

[28] Ghazala Azmat, Caterina Calsamiglia, and Nagore Iriberri. Gender differences in response to big stakes. *Journal of the European Economic Association*, 14:1372–1400, 2016.

[29] Ghazala Azmat, Caterina Calsamiglia, and Nagore Iriberri. Gender differences in response to big stakes. *Journal of the European Economic Association*, 14(6):1372–1400, 2016.

[30] Ghazala Azmat and Barbara Petrongolo. Gender and the labor market: What have we learned from field and lab experiments? *Labour Economics*, 30:32–40, 2014.

[31] Linda Babcock, Maria P Recalde, Lise Vesterlund, and Laurie Weingart. Gender differences in accepting and receiving requests for tasks with low promotability. *American Economic Review*, 107:714–747, 2017.

[32] Alexandra Baier, Brent J Davis, Tarek Jaber-Lopez, and Michael Seidl. Gender, competition and the effect of feedback and task: An experiment, 2018.

[33] Loukas Balafoutas, Rudolf Kerschbamer, and Matthias Sutter. Distributional preferences and competitive behavior. *Journal of economic behavior organization*, 83:125–135, 2012.

[34] Loukas Balafoutas and Matthias Sutter. Affirmative action policies promote women and do not harm efficiency in the laboratory. *science*, 335(6068):579–582, 2012.

[35] Loukas Balafoutas and Matthias Sutter. How uncertainty and ambiguity in tournaments affect gender differences in competitive behavior. *European Economic Review*, 118:1–13, 2019.

[36] Katherine Baldiga. Gender differences in willingness to guess. *Management Science*, 60:434–448, 2014.

[37] Nancy R Baldiga and Katherine B Coffman. Laboratory evidence on the effects of sponsorship on the competitive preferences of men and women. *Management Science*, 64:888–901, 2018.

[38] Daniel Balliet, Norman P Li, Shane J Macfarlan, and Mark Van Vugt. Sex differences in cooperation: a meta-analytic review of social dilemmas. *Psychological bulletin*, 137:881, 2011.

[39] Ritwik Banerjee, Nabanita Datta Gupta, and Marie Claire Villeval. The spillover effects of affirmative action on competitiveness and unethical behavior. *European Economic Review*, 101:567–604, 2018.

[40] Sheheryar Banuri and Philip Keefer. Pro-social motivation, effort and the call to public service. *European Economic Review*, 83:139–164, 2016.

[41] Brad M Barber and Terrance Odean. Boys will be boys: Gender, overconfidence, and common stock investment. *The quarterly journal of economics*, 116(1):261–292, 2001.

[42] Florian Baumann, Volker Benndorf, and Maria Friese. Loss-induced emotions and criminal behavior: an experimental analysis. *Journal of Economic Behavior Organization*, 159:134–145, 2019.

[43] Lori Beaman, Raghabendra Chattopadhyay, Esther Duflo, Rohini Pande, and Petia Topalova. Powerful women: does exposure reduce bias? *The Quarterly journal of economics*, 124:1497–1540, 2009.

[44] Lori Beaman, Niall Keleher, and Jeremy Magruder. Do job networks disadvantage women? evidence from a recruitment experiment in malawi. *Journal of Labor Economics*, 36:121–157, 2018.

[45] Leonardo Becchetti, Francesco Salustri, Vittorio Pelligra, and Alejandra Vásquez. Gender differences in socially responsible consumption. an experimental investigation. *Applied Economics*, 50:3630–3643, 2018.

[46] Gary Stanley Becker. *The economics of discrimination: an economic view of racial discrimination*. University of Chicago, 1957.

[47] David N Beede, Tiffany A Julian, David Langdon, George McKittrick, Beethika Khan, and Mark E Doms. Women in stem: A gender gap to innovation. *Economics and Statistics Administration Issue Brief*, 2011.

[48] Avner Ben-Ner, Fanmin Kong, and Louis Putterman. Share and share alike? gender-pairing, personality, and cognitive ability as determinants of giving. *Journal of Economic Psychology*, 25(5):581–589, 2004.

[49] Rikhil R Bhavnani. Do the effects of temporary ethnic group quotas persist? evidence from india. *American Economic Journal: Applied Economics*, 9:105–123, 2017.

[50] Andrea Blasco, Olivia S Jung, Karim R Lakhani, and Michael Menietti. Incentives for public goods inside organizations: Field experimental evidence. *Journal of Economic Behavior Organization*, 160:214–229, 2019.

[51] Cecilia Boggio, Flavia Coda Moscarola, and Andrea Gallice. What is good for the goose is good for the gander?: How gender-specific conceptual frames affect financial participation and decision-making. *Economics of Education Review*, 75:101952, 2020.

[52] Iris Bohnet, Alexandra Van Geen, and Max Bazerman. When performance trumps gender bias: Joint vs. separate evaluation. *Management Science*, 62:1225–1234, 2016.

[53] Iris Bohnet and Richard Zeckhauser. Trust, risk and betrayal. *Journal of Economic Behavior Organization*, 55:467–484, 2004.

[54] J Aislinn Bohren, Kareem Haggag, Alex Imas, and Devin G Pope. Inaccurate statistical discrimination: An identification problem, 2019.

[55] Gary E Bolton and Elena Katok. An experimental test for gender differences in beneficent behavior. *Economics Letters*, 48:287–292, 1995.

[56] Evelina Bonnier, Anna Dreber, Karin Hederos, and Anna Sandberg. Exposure to half-dressed women and economic behavior. *Journal of Economic Behavior Organization*, 168:393–418, 2019.

[57] Alison Booth, Lina Cardona-Sosa, and Patrick Nolen. Gender differences in risk aversion: do single-sex environments affect their development? *Journal of economic behavior organization*, 99:126–154, 2014.

[58] Alison Booth and Andrew Leigh. Do employers discriminate by gender? a field experiment in female-dominated occupations. *Economics Letters*, 107:236–238, 2010.

[59] Alison Booth and Patrick Nolen. Choosing to compete: How different are girls and boys? *Journal of Economic Behavior & Organization*, 81(2):542–555, 2012.

[60] Alison L Booth and Patrick Nolen. Gender differences in risk behaviour: does nurture matter? *The economic journal*, 122:F56–F78, 2012.

[61] Pedro Bordalo, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer. Stereotypes. *The Quarterly Journal of Economics*, 131:1753–1794, 2016.

[62] Pedro Bordalo, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer. Beliefs about gender. *American Economic Review*, 109:739–773, 2019.

[63] Lex Borghans, James J Heckman, Bart H H Golsteyn, and Huub Meijers. Gender differences in risk aversion and ambiguity aversion. *Journal of the European Economic Association*, 7:649–658, 2009.

[64] Andreas Born, Eva Ranehill, and Anna Sandberg. Gender and willingness to lead: Does the gender composition of teams matter? *The Review of Economics and Statistics*, pages 1–46, 2020.

[65] Anne Boschini, Anna Dreber, Emma von Essen, Astri Muren, and Eva Ranehill. Gender and altruism in a random sample. *Journal of behavioral and experimental economics*, 77:72–77, 2018.

[66] Anne Boschini, Anna Dreber, Emma von Essen, Astri Muren, and Eva Ranehill. Gender, risk preferences and willingness to compete in a random sample of the swedish population. *Journal of Behavioral and Experimental Economics*, 83:101467, 2019.

[67] Anne Boschini, Astri Muren, and Mats Persson. Constructing gender differences in the economics lab. *Journal of Economic Behavior Organization*, 84:741–752, 2012.

[68] Hannah Riley Bowles, Linda Babcock, and Lei Lai. Social incentives for gender differences in the propensity to initiate negotiations: Sometimes it does hurt to ask. *Organizational Behavior and human decision Processes*, 103:84–103, 2007.

[69] Hannah Riley Bowles, Linda Babcock, and Kathleen L McGinn. Constraints and triggers: situational mechanics of gender in negotiation. *Journal of personality and social psychology*, 89:951, 2005.

[70] Hannah Riley Bowles and Francis Flynn. Gender and persistence in negotiation: A dyadic perspective. *Academy of Management Journal*, 53:769–787, 2010.

[71] Pablo Brañas-Garza, Marisa Bucheli, and Maria Paz Espinosa. Altruism and information. *Journal of Economic Psychology*, 81:102332, 2020.

[72] Pablo Brañas-Garza, Valerio Capraro, and Ericka Rascon-Ramirez. Gender differences in altruism on mechanical turk: Expectations and actual behaviour. *Economics Letters*, 170:19–23, 2018.

[73] Jordi Brandts and Orsola Garofalo. Gender pairings and accountability effects. *Journal of Economic Behavior Organization*, 83:31–41, 2012.

[74] Jordi Brandts, Valeska Groenert, and Christina Rott. The impact of advice on women's and men's selection into competition. *Management Science*, 61:1018–1035, 2015.

[75] Jordi Brandts, Klarita Gërxhani, and Arthur Schram. Are there gender differences in status-ranking aversion? *Journal of Behavioral and Experimental Economics*, 84:101485, 2 2020.

[76] Jordi Brandts and Christina Rott. Advice from women and men and selection into competition. *Journal of Economic Psychology*, 82:102333, 2021.

[77] Lisa Bruttel and Florian Stolley. Getting a yes. an experiment on the power of asking. *Journal of Behavioral and Experimental Economics*, 86:101550, 2020.

[78] Nancy R Buchan, Rachel T A Croson, and Sara Solnick. Trust and gender: An examination of behavior and beliefs in the investment game. *Journal of Economic Behavior Organization*, 68:466–476, 2008.

[79] Kasey Buckles. Fixing the leaky pipeline: Strategies for making economics work for women at every stage. *Journal of economic perspectives*, 33:43–60, 2019.

[80] Norma Burow, Miriam Beblo, Denis Beninger, and Melanie Schröder. Why do women favor same-gender competition? evidence from a choice experiment. 2017.

[81] Leonardo Bursztyn, Thomas Fujiwara, and Amanda Pallais. 'acting wife': Marriage market incentives and labor market investments. *American Economic Review*, 107:3288–3319, 2017.

[82] Thomas Buser. The impact of the menstrual cycle and hormonal contraceptives on competitiveness. *Journal of Economic Behavior Organization*, 83:1–10, 2012.

[83] Thomas Buser, Anna Dreber, and Johanna Mollerstrom. Stress reactions cannot explain the gender gap in willingness to compete. 2015.

[84] Thomas Buser, Anna Dreber, and Johanna Mollerstrom. The impact of stress on tournament entry. *Experimental Economics*, 20:506–530, 2017.

[85] Thomas Buser, Muriel Niederle, and Hessel Oosterbeek. Gender, competitiveness, and career choices. *The quarterly journal of economics*, 129:1409–1447, 2014.

[86] Thomas Buser, Eva Ranehill, and Roel van Veldhuizen. Gender differences in willingness to compete: The role of public observability. *Journal of Economic Psychology*, 83:102366, 2021.

[87] Thomas Buser and Huaiping Yuan. Do women give up competing more easily? evidence from the lab and the dutch math olympiad. *American Economic Journal: Applied Economics*, 11:225–252, 2019.

[88] Meghan R Busse, Ayelet Israeli, and Florian Zettelmeyer. Repairing the damage: The effect of gender and price knowledge on auto-repair price quotes. *Journal of Marketing Research*, 2016.

[89] C Bram Cadsby and Elizabeth Maynes. Gender and free riding in a threshold public goods game: Experimental evidence. *Journal of economic behavior organization*, 34:603–620, 1998.

[90] C Bram Cadsby, Maroš Servátka, and Fei Song. Gender and generosity: does degree of anonymity or group gender composition matter? *Experimental economics*, 13:299–308, 2010.

[91] C Bram Cadsby, Maroš Servátka, and Fei Song. How competitive are female professionals? a tale of identity conflict. *Journal of Economic Behavior Organization*, 92:284–303, 2013.

[92] David Card, Ana Rute Cardoso, and Patrick Kline. Bargaining, sorting, and the gender wage gap: Quantifying the impact of firms on the relative pay of women. *The Quarterly Journal of Economics*, 131:633–686, 2016.

[93] Fredrik Carlsson, Elina Lampi, Peter Martinsson, and Xiaojun Yang. Replication: Do women shy away from competition? experimental evidence from china. *Journal of Economic Psychology*, 81:102312, 2020.

[94] Fredrik Carlsson, Peter Martinsson, Ping Qin, and Matthias Sutter. The influence of spouses on household decision making under risk: an experiment in rural china. *Experimental Economics*, 16:383–401, 2013.

[95] Jeffrey Carpenter, Rachel Frank, and Emiliano Huet-Vaughn. Gender differences in interpersonal and intrapersonal competitive behavior. *Journal of behavioral and experimental economics*, 77:170–176, 2018.

[96] Marco Castillo, Greg Leo, and Ragan Petrie. Room composition effects on risk taking by gender. *Experimental Economics*, 23:895–911, 2020.

[97] Marco Castillo, Ragan Petrie, Maximo Torero, and Lise Vesterlund. Gender differences in bargaining outcomes: A field experiment on discrimination. *Journal of Public Economics*, 99:35–48, 2013.

[98] Marco E Castillo and Philip J Cross. Of mice and men: Within gender variation in strategic behavior. *Games and Economic Behavior*, 64:421–432, 2008.

[99] Gary Charness, Ramón Cobo-Reyes, Simone Meraglia, and Ángela Sánchez. Anticipated discrimination, choices, and performance: Experimental evidence. *European Economic Review*, 127:103473, 2020.

[100] Gary Charness, Catherine Eckel, Uri Gneezy, and Agne Kajackaite. Complexity in risk elicitation may affect the conclusions: A demonstration using gender differences. *Journal of Risk and Uncertainty*, 56:1–17, 2018.

[101] Gary Charness and Uri Gneezy. Strong evidence for gender differences in risk taking. *Journal of Economic Behavior Organization*, 83:50–58, 2012.

[102] Gary Charness, Luca Rigotti, and Aldo Rustichini. Social surplus determines cooperation rates in the one-shot prisoner's dilemma. *Games and Economic Behavior*, 100:113–124, 2016.

[103] Gary Charness and Aldo Rustichini. Gender differences in cooperation with group membership. *Games and Economic Behavior*, 72:77–85, 2011.

[104] Raghabendra Chattopadhyay and Esther Duflo. Women as policy makers: Evidence from a randomized policy experiment in india. *Econometrica*, 72(5):1409–1443, 2004.

[105] Ananish Chaudhuri and Erwann Sbai. Gender differences in trust and reciprocity in repeated gift exchange games. *New Zealand Economic Papers*, 45:81–95, 2011.

[106] Daniel L Chen, Martin Schonger, and Chris Wickens. otree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9:88–97, 2016.

[107] Zhuoqiong Charlie Chen, David Ong, and Roman M Sheremeta. The gender difference in the value of winning. *Economics Letters*, 137:226–229, 2015.

[108] Michael Chmielewski and Sarah C Kucker. An mturk crisis? shifts in data quality and the impact on study results. *Social Psychological and Personality Science*, 11(4):464–473, 2020.

[109] Subhasish M Chowdhury, Philip J Grossman, and Joo Young Jeon. Gender differences in giving and the anticipation regarding giving in dictator games. *Oxford Economic Papers*, 72:772–779, 2020.

[110] Leonardo Christov-Moore, Elizabeth A Simpson, Gino Coudé, Kristina Grigaityte, Marco Iacoboni, and Pier Francesco Ferrari. Empathy: Gender effects in brain and behavior. *Neuroscience & biobehavioral reviews*, 46:604–627, 2014.

[111] Stephen Coate and Glenn C Loury. Will affirmative-action policies eliminate negative stereotypes? *The American Economic Review*, pages 1220–1240, 1993.

[112] Katherine Coffman, Manuela Collis, and Leena Kulkarni. *Stereotypes and belief updating*. Harvard Business School, 2019.

[113] Katherine Baldiga Coffman. Evidence on self-stereotyping and the contribution of ideas. *The Quarterly Journal of Economics*, 129:1625–1660, 2014.

[114] Francesca Cornaglia, Michalis Drouvelis, and Paolo Masella. Competition and the role of group identity. *Journal of Economic Behavior Organization*, 162:136–145, 2019.

[115] Christopher Cotton, Frank McIntyre, and Joseph Price. Gender differences in repeated competition: Evidence from school math contests. *Journal of Economic Behavior Organization*, 86:52–66, 2013.

[116] James C Cox and Cary A Deck. When are women more generous than men? *Economic Inquiry*, 44:587–598, 2006.

[117] Rachel Croson and Nancy Buchan. Gender and culture: International experimental evidence from trust games. *American Economic Review*, 89:386–391, 1999.

[118] Rachel Croson and Uri Gneezy. Gender differences in preferences. *Journal of Economic literature*, 47:448–474, 2009.

[119] Rachel Croson, Melanie Marks, and Jessica Snyder. Groups work for women: Gender and group identity in social dilemmas. *Negotiation Journal*, 24:411–427, 2008.

[120] Rachel T A Croson, Femida Handy, and Jen Shang. Gendered giving: the influence of social norms on the donation behavior of men and women. *International Journal of Nonprofit and Voluntary Sector Marketing*, 15:199–213, 2010.

[121] Carlos Cueva, Iñigo Iturbe-Ormaetxe, Giovanni Ponti, and Josefa Tomás. Boys will still be boys: Gender differences in trading activity are not due to differences in (over) confidence. *Journal of Economic Behavior Organization*, 160:100–120, 2019.

[122] Carlos Cueva and Aldo Rustichini. Is financial instability male-driven? gender and cognitive skills in experimental asset markets. *Journal of Economic Behavior Organization*, 119:330–344, 2015.

[123] Eszter Czibor and Silvia Dominguez Martinez. Never too late: Gender quotas in the final round of a multistage tournament. *The Journal of Law, Economics, and Organization*, 35:319–363, 2019.

[124] Ernesto Dal Bó and Pedro Dal Bó. "do the right thing:" the effects of moral suasion on cooperation. *Journal of Public Economics*, 117:28–38, 2014.

[125] Katarína Danková and Maroš Servátka. Gender robustness of overconfidence and excess entry. *Journal of Economic Psychology*, 72:179–199, 2019.

[126] David Danz, Neeraja Gupta, Marissa Lepper, Lise Vesterlund, and K Pun Winichakul. Going virtual: A step-by-step guide to taking the in-person experimental lab online. *Available at SSRN 3931028*, 2021.

[127] David Danz, Lise Vesterlund, and Alistair J Wilson. Belief elicitation and behavioral incentive compatibility. *American Economic Review*, 2022.

[128] Marie-Pierre Dargnies. Men too sometimes shy away from competition: The case of team competition. *Management Science*, 58:1982–2000, 2012.

[129] Aurelie Dariel, Curtis Kephart, Nikos Nikiforakis, and Christina Zenker. Emirati women do not shy away from competition: Evidence from a patriarchal society in transition. *Journal of the Economic Science Association*, 3:121–136, 2017.

[130] Aurélie Dariel, Nikos Nikiforakis, Jan Stoop, et al. Does selection bias cause us to overestimate gender differences in competitiveness?, 2020.

[131] Utteeyo Dasgupta, Subha Mani, Smriti Sharma, and Saurabh Singhal. Can gender differences in distributional preferences explain gender gaps in competition? *Journal of Economic Psychology*, 70:1–11, 2019.

[132] Simon Dato and Petra Nieken. Gender differences in competition and sabotage. *Journal of Economic Behavior Organization*, 100:64–80, 2014.

[133] Simon Dato and Petra Nieken. Gender differences in sabotage: the role of uncertainty and beliefs. *Experimental Economics*, 23:353–391, 2020.

[134] Nabanita Datta Gupta, Anders Poulsen, and Marie Claire Villeval. Gender matching and competitiveness: Experimental evidence. *Economic Inquiry*, 51(1):816–835, 2013.

[135] Maria De Paola, Francesca Gioia, and Vincenzo Scoppa. Are females scared of competing with males? results from a field experiment. *Economics of Education Review*, 48:117–128, 2015.

[136] Klaus Deininger, Songqing Jin, Hari K Nagarajan, and Fang Xia. Does female reservation affect long-term political outcomes? evidence from rural india. *The Journal of Development Studies*, 51:32–49, 2015.

[137] Josse Delfgaauw, Robert Dur, Joeri Sol, and Willem Verbeke. Tournament incentives in the field: Gender differences in the workplace. *Journal of Labor Economics*, 31:305–326, 2013.

[138] Stefano DellaVigna, John A List, Ulrike Malmendier, and Gautam Rao. The importance of being marginal: Gender differences in generosity. *American Economic Review*, 103:586–590, 2013.

[139] Stefano DellaVigna and Devin Pope. Stability of experimental results: Forecasts and evidence. *American Economic Journal: Microeconomics*, 14:889–925, 8 2022.

[140] Ahrash Dianat, Federico Echenique, and Leeat Yariv. Statistical discrimination and affirmative action in the lab. *Games and Economic Behavior*, 132:41–58, 2022.

[141] David L Dickinson and Jill Tiefenthaler. What is fair? experimental evidence. *Southern Economic Journal*, pages 414–428, 2002.

[142] Lisa M Dickson. Does ending affirmative action in college admissions lower the percent of minority students applying to college? *Economics of Education Review*, 25:109–119, 2006.

[143] Marcus Dittrich, Andreas Knabe, and Kristina Leipold. Gender differences in experimental wage negotiations. *Economic Inquiry*, 52(2):862–873, 2014.

[144] Marcus Dittrich and Kristina Leipold. Gender differences in time preferences. *Economics Letters*, 122:413–415, 2014.

[145] Thomas Dohmen and Armin Falk. Performance pay and multidimensional sorting: Productivity, preferences, and gender. *American economic review*, 101:556–590, 2011.

[146] Thomas Dohmen, Armin Falk, David Huffman, Uwe Sunde, Jürgen Schupp, and Gert G Wagner. Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the european economic association*, 9(3):522–550, 2011.

[147] Anna Dreber, Emma Von Essen, and Eva Ranehill. Gender and competition in adolescence: task matters. *Experimental Economics*, 17:154–172, 2014.

[148] Anna Dreber and Magnus Johannesson. Gender differences in deception. *Economics Letters*, 99:197–199, 2008.

[149] Moritz A Drupp, Menusch Khadjavi, Marie-Catherine Riekhof, and Rudi Voss. Professional identity and the gender gap in risk-taking. evidence from field experiments with scientists. *Journal of Economic Behavior Organization*, 170:418–432, 2020.

[150] Martin Dufwenberg and Astri Muren. Gender composition in teams. *Journal of Economic Behavior Organization*, 61:50–54, 2006.

[151] Martin Dufwenberg and Astri Muren. Generosity, anonymity, gender. *Journal of Economic Behavior & Organization*, 61(1):42–49, 2006.

[152] Ben D'Exelle, Christine Gutekunst, and Arno Riedl. The effect of gender and gender pairing on bargaining: Evidence from an artefactual field experiment. *Journal of Economic Behavior & Organization*, 205:237–269, 2023.

[153] Catherine Eckel, Angela CM De Oliveira, and Philip J Grossman. Gender and negotiation in the small: are women (perceived to be) more cooperative than men? *Negotiation Journal*, 24(4):429–445, 2008.

[154] Catherine Eckel, Angela C M De Oliveira, and Philip J Grossman. Gender and negotiation in the small: are women (perceived to be) more cooperative than men? *Negotiation Journal*, 24:429–445, 2008.

[155] Catherine Eckel and Rick K Wilson. Whom to trust? choice of partner in a trust game. *Department of Economics, Virginia Tech http://www. ruf. rice. edu/rkw/RKW$_F$OLDER/EckelWilsonWhomToTrust.pdf*, 2004.

[156] Catherine C Eckel and Sascha C Füllbrunn. Thar she blows? gender, competition, and bubbles in experimental asset markets. *American Economic Review*, 105:906–920, 2015.

[157] Catherine C Eckel and Philip J Grossman. The relative price of fairness: Gender differences in a punishment game. *Journal of Economic Behavior Organization*, 30:143–158, 1996.

[158] Catherine C Eckel and Philip J Grossman. Are women less selfish than men?: Evidence from dictator experiments. *The economic journal*, 108(448):726–735, 1998.

[159] Catherine C Eckel and Philip J Grossman. Chivalry and solidarity in ultimatum games. *Economic inquiry*, 39:171–188, 2001.

[160] Catherine C Eckel and Philip J Grossman. Sex differences and statistical stereotyping in attitudes toward financial risk. *Evolution and human behavior*, 23:281–295, 2002.

[161] Catherine C Eckel and Rick K Wilson. Is trust a risky decision? *Journal of Economic Behavior & Organization*, 55(4):447–465, 2004.

[162] Tore Ellingsen, Magnus Johannesson, Johanna Mollerstrom, and Sara Munkhammar. Gender differences in social framing effects. *Economics Letters*, 118:470–472, 2013.

[163] Nicole M Else-Quest, Ashley Higgins, Carlie Allison, and Lindsay C Morton. Gender differences in self-conscious emotional experience: a meta-analysis. *Psychological bulletin*, 138(5):947, 2012.

[164] Kimmo Eriksson and Brent Simpson. Emotional reactions to losing explain gender differences in entering a risky lottery. *Judgment and Decision making*, 5:159, 2010.

[165] Seda Ertac, Mert Gumren, and Mehmet Y Gurdal. Demand for decision autonomy and the desire to avoid responsibility in risky environments: Experimental evidence. *Journal of Economic Psychology*, 77:102200, 2020.

[166] Seda Ertac and Mehmet Y Gurdal. Deciding to decide: Gender, leadership and risk-taking in groups. *Journal of Economic Behavior Organization*, 83:24–30, 2012.

[167] Seda Ertac and Balazs Szentes. The effect of performance feedback on gender differences in competitiveness: experimental evidence. *Work. Pap., Koc Univ., Turkey*, 2010.

[168] Christine L Exley and Judd B Kessler. The gender gap in self-promotion, 2019.

[169] Christine L Exley, Muriel Niederle, and Lise Vesterlund. Knowing when to ask: The cost of leaning in. *Journal of Political Economy*, 128:816–854, 2020.

[170] Lara Ezquerra, Gueorgui I Kolev, and Ismael Rodriguez-Lara. Gender differences in cheating: Loss vs. gain framing. *Economics Letters*, 163:46–49, 2018.

[171] Gerlinde Fellner and Boris Maciejovsky. Risk attitude and market behavior: Evidence from experimental asset markets. *Journal of Economic Psychology*, 28:338–350, 2007.

[172] Chaim Fershtman and Uri Gneezy. Discrimination in a segmented society: An experimental approach. *The Quarterly Journal of Economics*, 116:351–377, 2001.

[173] Antonio Filippin and Paolo Crosetto. A reconsideration of gender differences in risk attitudes. *Management Science*, 62:3138–3160, 2016.

[174] Antonio Filippin and Francesca Gioia. Competition and subsequent risk-taking behaviour: Heterogeneity across gender and outcomes. *Journal of Behavioral and Experimental Economics*, 75:84–94, 2018.

[175] Raymond Fisman, Sheena S Iyengar, Emir Kamenica, and Itamar Simonson. Gender differences in mate selection: Evidence from a speed dating experiment. *The Quarterly Journal of Economics*, 121:673–697, 2006.

[176] Jeffrey A Flory, Uri Gneezy, Kenneth L Leonard, and John A List. Gender, age, and competition: A disappearing gap? *Journal of Economic Behavior Organization*, 150:256–276, 2018.

[177] Jeffrey A Flory, Andreas Leibbrandt, and John A List. Do competitive workplaces deter female workers? a large-scale natural field experiment on job entry decisions. *The Review of Economic Studies*, 82:122–155, 2015.

[178] Guido Friebel, Marie Lalanne, Bernard Richter, Peter Schwardmann, and Paul Seabright. Gender differences in social interactions. *Journal of Economic Behavior Organization*, 186:33–45, 2021.

[179] Andreas Friedl, Andreas Pondorfer, and Ulrich Schmidt. Gender differences in social risk taking. *Journal of Economic Psychology*, 77:102182, 2020.

[180] Hiroaki Fujimoto and Eun-Soo Park. Framing effects and gender differences in voluntary public goods provision experiments. *The Journal of Socio-Economics*, 39:455–457, 2010.

[181] Frank Fujita, Ed Diener, and Ed Sandvik. Gender differences in negative affect and well-being: the case for emotional intensity. *Journal of personality and social psychology*, 61:427, 1991.

[182] Nadja C Furtner, Martin G Kocher, Peter Martinsson, Dominik Matzat, and Conny Wollbrant. Gender and cooperative preferences. *Journal of Economic Behavior Organization*, 181:39–48, 2021.

[183] Ellen Garbarino and Robert Slonim. The robustness of trust and reciprocity across a heterogeneous us population. *Journal of Economic Behavior Organization*, 69:226–240, 2009.

[184] Aurora García-Gallego, Nikolaos Georgantzís, and Ainhoa Jaramillo-Gutiérrez. Gender differences in ultimatum games: Despite rather than due to risk attitudes. *Journal of Economic Behavior & Organization*, 83(1):42–49, 2012.

[185] Leonie Gerhards and Michael Kosfeld. I (don't) like you! but who cares? gender differences in same-sex and mixed-sex teams. *The Economic Journal*, 130:716–739, 2020.

[186] David Gill and Victoria Prowse. Gender differences and dynamics in competition: The role of luck. *Quantitative Economics*, 5:351–376, 2014.

[187] Carol Gilligan. *In a different voice: Psychological theory and women's development*. Harvard University Press, 1993.

[188] William Gilje Gjedrem and Ola Kvaløy. Relative performance feedback to teams. *Labour Economics*, 66:101865, 2020.

[189] Uri Gneezy, Kenneth L Leonard, and John A List. Gender differences in competition: Evidence from a matrilineal and a patriarchal society. *Econometrica*, 77:1637–1664, 2009.

[190] Uri Gneezy, Muriel Niederle, and Aldo Rustichini. Performance in competitive environments: Gender differences. *The quarterly journal of economics*, 118:1049–1074, 2003.

[191] Uri Gneezy and Aldo Rustichini. Gender and competition at a young age. *American Economic Review*, 94:377–381, 2004.

[192] Timo Goeschl, Sara Elisa Kettner, Johannes Lohse, and Christiane Schwieren. From social information to social norms: Evidence from two experiments on donation behaviour. *Games*, 9:91, 2018.

[193] Claudia Goldin. Gender and the undergraduate economics major: Notes on the undergraduate economics major at a highly selective liberal arts college. *manuscript, April*, 12, 2015.

[194] Binglin Gong, Huibin Yan, and Chun-Lei Yang. Gender differences in the dictator experiment: evidence from the matrilineal mosuo and the patriarchal yi. *Experimental economics*, 18:302–313, 2015.

[195] Binglin Gong and Chun-Lei Yang. Gender differences in risk attitudes: Field experiments on the matrilineal mosuo and the patriarchal yi. *Journal of economic behavior organization*, 83:59–65, 2012.

[196] Joseph K Goodman, Cynthia E Cryder, and Amar Cheema. Data collection in a flat world: The strengths and weaknesses of mechanical turk samples. *Journal of Behavioral Decision Making*, 26(3):213–224, 2013.

[197] Stefan Grimm. Effects of choice observability on risk taking: The role of norms. *Journal of Behavioral and Experimental Economics*, 80:34–46, 2019.

[198] Kerstin Grosch, Stephan Müller, Holger A Rau, and Lilia Zhurakhovska. Selection into leadership and dishonest behavior of leaders: A gender experiment. 2020.

[199] Philip J Grossman, Catherine Eckel, Mana Komai, and Wei Zhan. It pays to be a man: Rewards for leaders in a coordination game. *Journal of Economic Behavior Organization*, 161:197–215, 2019.

[200] Christina Günther, Neslihan Arslan Ekinci, Christiane Schwieren, and Martin Strobel. Women can't jump?—an experiment on competitive attitudes and stereotype threat. *Journal of Economic Behavior & Organization*, 75(3):395–401, 2010.

[201] Neeraja Gupta, Luca Rigott, and Alistair Wilson. The experimenters' dilemma: Inferential preferences over populations. *arXiv preprint arXiv:2107.05064*, 2021.

[202] Werner Güth, Carsten Schmidt, and Matthias Sutter. Bargaining outside the lab–a newspaper experiment of a three-person ultimatum game. *The Economic Journal*, 117(518):449–469, 2007.

[203] Marja-Liisa Halko and Lauri Sääksvuori. Competitive behavior, stress, and gender. *Journal of Economic Behavior Organization*, 141:96–109, 2017.

[204] John C Ham and John H Kagel. Gender effects in private value auctions. *Economics Letters*, 92:375–382, 2006.

[205] David Hauser, Gabriele Paolacci, and Jesse Chandler. Common concerns with mturk as a participant pool: Evidence and solutions. In *Handbook of research methods in consumer psychology*, pages 319–337. Routledge, 2019.

[206] Haoran He, Peter Martinsson, and Matthias Sutter. Group decision making under risk: An experiment with student couples. *Economics Letters*, 117:691–693, 2012.

[207] Xin He, J Jeffrey Inman, and Vikas Mittal. Gender jeopardy in financial risk taking. *Journal of Marketing Research*, 45:414–424, 2008.

[208] Andrew Healy and Jennifer Pate. Can teams help to close the gender competition gap? *The Economic Journal*, 121:1192–1204, 2011.

[209] Matthias Heinz, Steffen Juranek, and Holger A Rau. Do women behave more reciprocally than men? gender differences in real effort dictator games. *Journal of Economic Behavior Organization*, 83:105–110, 2012.

[210] Matthias Heinz, Hans-Theo Normann, and Holger A Rau. How competitiveness may cause a gender wage gap: Experimental evidence. *European Economic Review*, 90:336–349, 2016.

[211] Iñigo Hernandez-Arenaz. Stereotypes and tournament self-selection: A theoretical and experimental approach. *European Economic Review*, 126:103448, 2020.

[212] Inigo Hernández-Arenaz and Nagore Iriberri. Women ask for less (only from men): Evidence from alternating-offer bargaining in the field. 2016.

[213] Andrew J Hill. State affirmative action bans and stem degree completions. *Economics of Education Review*, 57:31–40, 2017.

[214] Amy Hinsley, William J Sutherland, and Alison Johnston. Men ask more questions than women at a scientific conference. *PloS one*, 12:e0185534, 2017.

[215] Moshe Hoffman, Uri Gneezy, and John A List. Nurture affects gender differences in spatial abilities. *Proceedings of the National Academy of Sciences*, 108:14786–14788, 2011.

[216] Robin M Hogarth, Natalia Karelaia, and Carlos Andrés Trujillo. When should i quit? gender differences in exiting competitions. *Journal of Economic Behavior Organization*, 83:136–150, 2012.

[217] Sander Hoogendoorn, Hessel Oosterbeek, and Mirjam Van Praag. The impact of gender diversity on the performance of business teams: Evidence from a field experiment. *Management Science*, 59:1514–1528, 2013.

[218] John J Horton, David G Rand, and Richard J Zeckhauser. The online laboratory: Conducting experiments in a real labor market. *Experimental economics*, 14(3):399–425, 2011.

[219] Tanjim Hossain and Ryo Okui. The binarized scoring rule. *Review of Economic Studies*, 80:984–1001, 2013.

[220] Britta Hoyer, Thomas van Huizen, Linda Keijzer, Sarah Rezaei, Stephanie Rosenkranz, and Bastian Westbrock. Gender, competitiveness, and task difficulty: Evidence from the field. *Labour Economics*, 64:101815, 2020.

[221] Sabine Hügelschäfer and Anja Achtziger. On confident men and rational women: It's all on your mind (set). *Journal of Economic Psychology*, 41:31–44, 2014.

[222] Steven J Humphrey and Stefan Mondorf. Testing the causes of betrayal aversion. *Economics Letters*, 198:109663, 2021.

[223] Nagore Iriberri and Pedro Rey-Biel. *Let's (not) talk about sex: Gender awareness and stereotype-threat on performance under competition*. Citeseer, 2012.

[224] Nagore Iriberri and Pedro Rey-Biel. Brave boys and play-it-safe girls: Gender differences in willingness to guess in a large scale natural field experiment. *European Economic Review*, 131:103603, 2021.

[225] Radosveta Ivanova-Stenzel and Dorothea Kübler. Gender differences in team work and team competition. *Journal of Economic Psychology*, 32:797–808, 2011.

[226] Ben Jacobsen, John B Lee, Wessel Marquering, and Cherry Y Zhang. Gender differences in optimism and asset allocation. *Journal of Economic Behavior Organization*, 107:630–651, 2014.

[227] Lars Bo Jeppesen and Karim R Lakhani. Marginality and problem-solving effectiveness in broadcast search. *Organization science*, 21:1016–1033, 2010.

[228] Sabrina Jeworrek. Gender stereotypes still in mind: Information on relative performance and competition entry. *Journal of Behavioral and Experimental Economics*, 82:101448, 2019.

[229] Nancy Ammon Jianakoplos and Alexandra Bernasek. Are women more risk averse? *Economic inquiry*, 36:620–630, 1998.

[230] SeEun Jung, Chung Choe, and Ronald L Oaxaca. Gender wage gaps and risky vs. secure employment: An experimental analysis. *Labour Economics*, 52:112–121, 2018.

[231] SeEun Jung and Radu Vranceanu. Competitive compensation and subjective well-being: The effect of culture and gender. *Journal of Economic Psychology*, 70:90–108, 2019.

[232] Linda Kamas and Anne Preston. Gender and social preferences in the us: an experimental study. *Feminist Economics*, 18:135–160, 2012.

[233] Linda Kamas and Anne Preston. The importance of being confident; gender, career choice, and willingness to compete. *Journal of Economic Behavior Organization*, 83:82–97, 2012.

[234] Linda Kamas and Anne Preston. Can social preferences explain gender differences in economic behavior? *Journal of Economic Behavior Organization*, 116:525–539, 2015.

[235] Linda Kamas and Anne Preston. Empathy, gender, and prosocial behavior. *Journal of Behavioral and Experimental Economics*, 92:101654, 2021.

[236] Linda Kamas, Anne Preston, and Sandy Baum. Altruism in individual and joint-giving decisions: What's gender got to do with it? *Feminist Economics*, 14:23–50, 2008.

[237] Hå kan Holm and Peter Engseld. Choosing bargaining partners—an experimental study on the impact of information about income, status and gender. *Experimental Economics*, 8:183–216, 2005.

[238] Kristin Kanthak and Jonathan Woon. Women don't run? election aversion and candidate entry. *American Journal of Political Science*, 59:595–612, 2015.

[239] Christopher F Karpowitz, Tali Mendelberg, and Lee Shaker. Gender inequality in deliberative participation. *American Political Science Review*, 106:533–547, 2012.

[240] Sara Elisa Kettner and Smarandita Ceccato. Framing matters in gender-paired dictator games, 2014.

[241] Ling Yee Khor, Orkhan Sariyev, and Tim Loos. Gender differences in risk behavior and the link to household effects and individual wealth. *Journal of Economic Psychology*, 80:102266, 2020.

[242] Janina Kleinknecht. A man of his word? an experiment on gender differences in promise keeping. *Journal of Economic Behavior Organization*, 168:251–268, 2019.

[243] David Klinowski. Gender differences in giving in the dictator game: the role of reluctant altruism. *Journal of the Economic Science Association*, 4:110–122, 2018.

[244] David Klinowski. Selection into self-improvement and competition pay: Gender, stereotypes, and earnings volatility. *Journal of Economic Behavior Organization*, 158:128–146, 2019.

[245] Mikael Knutsson, Peter Martinsson, Emil Persson, and Conny Wollbrant. Gender differences in altruism: Evidence from a natural field experiment on matched donations. *Economics Letters*, 176:47–50, 2019.

[246] James Konow, Tatsuyoshi Saijo, and Kenju Akai. Equity versus equality: Spectators, stakeholders and groups. *Journal of Economic Psychology*, 77:102171, 2020.

[247] Michał Krawczyk and Magdalena Smyk. Author s gender affects rating of academic articles: Evidence from an incentivized, deception-free laboratory experiment. *European Economic Review*, 90:326–335, 2016.

[248] Eryk Krysowski and James Tremewan. Why does anonymity make us misbehave: Different norms or less compliance? *Economic Inquiry*, 59:776–789, 2021.

[249] Peter Kuhn and Marie Claire Villeval. Are women more attracted to co-operation than men? *The Economic Journal*, 125:115–140, 2015.

[250] Jacob Ladenburg and Søren Bøye Olsen. Gender-specific starting point bias in choice experiments: Evidence from an empirical study. *Journal of Environmental Economics and Management*, 56:275–285, 2008.

[251] Victor Lavy. Gender differences in market competitiveness in a real workplace: Evidence from performance-based pay tournaments among teachers. *The Economic Journal*, 123:540–573, 2013.

[252] Anh T Le, Paul W Miller, Wendy S Slutske, and Nicholas G Martin. Attitudes towards economic risk and the gender pay gap. *Labour economics*, 18:555–561, 2011.

[253] Andreas Leibbrandt and John A List. Do women avoid salary negotiations? evidence from a large-scale natural field experiment. *Management Science*, 61:2016–2024, 2015.

[254] Andreas Leibbrandt, Liang Choon Wang, and Cordelia Foo. Gender quotas, competitions, and peer review: Experimental evidence on the backlash against women. *Management Science*, 64:3501–3516, 2018.

[255] Stephen Leider, Tanya Rosenblat, Markus M Möbius, and Quoc-Anh Do. What do we expect from our friends? *Journal of the European Economic Association*, 8:120–138, 2010.

[256] Louis Pierre Lepage. Endogenous learning, persistent employer biases, and discrimination. *Persistent Employer Biases, and Discrimination (March 2, 2021)*, 2021.

[257] Sarah-Jane Leslie, Andrei Cimpian, Meredith Meyer, and Edward Freeland. Expectations of brilliance underlie gender distributions across academic disciplines. *Science*, 347:262–265, 2015.

[258] Shuwen Li, Xiangdong Qin, and Daniel Houser. Revisiting gender differences in ultimatum bargaining: experimental evidence from the us and china. *Journal of the Economic Science Association*, 4:180–190, 2018.

[259] Yaxiong Li, Erwann Sbai, and Ananish Chaudhuri. An experimental study of gender differences in agency relationships. *Journal of Behavioral and Experimental Economics*, 90:101650, 2021.

[260] John A List. Young, selfish and male: Field evidence of social preferences. *The Economic Journal*, 114:121–149, 2004.

[261] George F Loewenstein, Elke U Weber, Christopher K Hsee, and Ned Welch. Risk as feelings. *Psychological bulletin*, 127:267, 2001.

[262] Tim Lohse and Salmai Qari. Gender differences in face-to-face deceptive behavior. *Journal of Economic Behavior Organization*, 187:1–15, 2021.

[263] Sebastian Lotz. Is women's behavior more context-dependent than men's? gender differences in reluctant altruism. *Gender Differences in Reluctant Altruism (December 18, 2014)*, 2014.

[264] Sara Lowes. Kinship structure, stress, and the gender gap in competition, 2018.

[265] Sandra Ludwig, Gerlinde Fellner-Röhling, and Carmen Thoma. Do women have more shame than men? an experiment on self-assessment and the shame of overestimating oneself. *European Economic Review*, 92:31–46, 2017.

[266] Valeria Maggian, Natalia Montinari, and Antonio Nicolò. Do quotas help women to climb the career ladder? a laboratory experiment. *European economic review*, 123:103390, 2020.

[267] Shakun D Mago and Laura Razzolini. Best-of-five contest: An experiment on gender differences. *Journal of Economic Behavior Organization*, 162:164–187, 2019.

[268] Samreen Malik, Benedikt Mihm, Maximilian Mihm, and Florian Timme. Can gender stereotypes mitigate gender differences? an experiment on bargaining with asymmetric information. 2018.

[269] Allison Mann and Thomas A DiPrete. Trends in gender segregation in the choice of science and engineering majors. *Social science research*, 42:1519–1541, 2013.

[270] David Masclet, Emmanuel Peterle, and Sophie Larribeau. Gender differences in tournament and flat-wage schemes: An experimental study. *Journal of Economic Psychology*, 47:103–115, 2015.

[271] David A Matsa and Amalia R Miller. Chipping away at the glass ceiling: Gender spillovers in corporate leadership. *American Economic Review*, 101:635–639, 2011.

[272] Sylvia Maxfield, Mary Shapiro, Vipin Gupta, and Susan Hass. Gender and risk: women, risk taking and risk aversion. *Gender in Management: An International Journal*, 2010.

[273] Bryan C McCannon and Amir B Ferreira Neto. Charitable giving for cultural goods: Asymmetric gender responses to votes on tax increases. *Economics Letters*, 204:109879, 2021.

[274] Stephan Meier. Do women behave less or more prosocially than men? evidence from two field experiments. *Public Finance Review*, 35:215–232, 2007.

[275] Friederike Mengel. Gender differences in networking. *The Economic Journal*, 130:1842–1873, 2020.

[276] Vanessa Mertins and Christian Walter. In absence of money: a field experiment on volunteer work motivation. *Experimental Economics*, 24:952–984, 2021.

[277] Amalia R Miller and Carmit Segal. Does temporary affirmative action produce persistent effects? a study of black and female employment in law enforcement. *Review of Economics and Statistics*, 94:1107–1125, 2012.

[278] Conrad Miller. The persistent effect of temporary affirmative action. *American Economic Journal: Applied Economics*, 9:152–190, 2017.

[279] J Alberto Molina, J Ignacio Giménez-Nadal, José A Cuesta, Carlos Gracia-Lazaro, Yamir Moreno, and Angel Sanchez. Gender differences in cooperation: experimental evidence on high school students. *PloS one*, 8:e83700, 2013.

[280] Natalia Montinari and Michela Rancan. A friend is a treasure: On the interplay of social distance and monetary incentives when risk is taken on behalf of others. *Journal of Behavioral and Experimental Economics*, 86:101544, 2020.

[281] John Morgan, Henrik Orzen, and Martin Sefton. Endogenous entry in contests. *Economic Theory*, 51:435–463, 2012.

[282] John Morgan, Henrik Orzen, Martin Sefton, and Dana Sisak. Strategic and natural risk in entrepreneurship: An experimental study. *Journal of Economics Management Strategy*, 25:420–454, 2016.

[283] Gerd Muehlheusser, Andreas Roider, and Niklas Wallmeier. Gender differences in honesty: Groups versus individuals. *Economics Letters*, 128:25–29, 2015.

[284] Shagata Mukherjee. What drives gender differences in trust and trustworthiness? *Public Finance Review*, 48:778–805, 2020.

[285] Daniel Müller. The anatomy of distributional preferences with group identity. *Journal of Economic Behavior & Organization*, 166:785–807, 2019.

[286] Julia Müller and Christiane Schwieren. Can personality explain what is underlying women's unwillingness to compete? *Journal of Economic Psychology*, 33(3):448–460, 2012.

[287] Kene Boun My, Nicolas Lampach, Mathieu Lefebvre, and Jacopo Magnani. Effects of gain-loss frames on advantageous inequality aversion. *Journal of the Economic Science Association*, 4:99–109, 2018.

[288] Muriel Niederle, Carmit Segal, and Lise Vesterlund. How costly is diversity? affirmative action in light of gender differences in competitiveness. *Management Science*, 59(1):1–16, 2013.

[289] Muriel Niederle, Carmit Segal, and Lise Vesterlund. How costly is diversity? affirmative action in light of gender differences in competitiveness. *Management Science*, 59:1–16, 2013.

[290] Muriel Niederle and Lise Vesterlund. Do women shy away from competition? do men compete too much? *The quarterly journal of economics*, 122:1067–1101, 2007.

[291] Muriel Niederle and Lise Vesterlund. Gender differences in competition. *Negotiation Journal*, 24, 10 2008.

[292] Clifford Nowell and Sarah Tinkler. The influence of gender on the provision of a public good. *Journal of Economic Behavior Organization*, 25:25–36, 1994.

[293] Elizabeth A Olson, Isabelle M Rosso, Lauren A Demers, Shreya Divatia, and William D S Killgore. Sex differences in psychological factors associated with social discounting. *Journal of Behavioral Decision Making*, 29:60–66, 2016.

[294] Hessel Oosterbeek and Reyn Van Ewijk. Gender peer effects in university: Evidence from a randomized experiment. *Economics of Education Review*, 38:51–63, 2014.

[295] Evren Ors, Frédéric Palomino, and Eloic Peyrache. Performance gender gap: does competition matter? *Journal of Labor Economics*, 31:443–499, 2013.

[296] Andreas Ortmann and Lisa K Tichy. Gender differences in the laboratory: evidence from prisoner's dilemma games. *Journal of Economic Behavior Organization*, 39:327–339, 1999.

[297] Kai Ou and Xiaofei Pan. The effect of task choice and task assignment on the gender earnings gap: An experimental study. *European Economic Review*, 136:103753, 2021.

[298] Maria De Paola, Vincenzo Scoppa, and Rosetta Lombardo. Can gender quotas break down negative stereotypes? evidence from changes in electoral rules. *Journal of Public Economics*, 94:344–353, 2010.

[299] Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis. Running experiments on amazon mechanical turk. *Judgment and Decision making*, 5(5):411–419, 2010.

[300] Matt Parrett. An analysis of the determinants of tipping behavior: A laboratory experiment and evidence from restaurant tipping. *Southern Economic Journal*, pages 489–514, 2006.

[301] Jennifer Pate and Richard Fox. Getting past the gender gap in political ambition. *Journal of Economic Behavior Organization*, 156:166–183, 2018.

[302] Kathryn Pearson and Eric McGhee. Should women win more often than men? the roots of electoral success and gender bias in us house elections. *The Roots of Electoral Success and Gender Bias in US House Elections (April 29, 2013)*, 2013.

[303] Marco Perugini, Jonathan H W Tan, and Daniel John Zizzo. Which is the more predictable gender? public good contribution and personality. *Public Good Contribution and Personality (March 1, 2005)*, 2005.

[304] Pascale Petit. The effects of age and family constraints on gender hiring discrimination: A field experiment in the french financial sector. *Labour Economics*, 14:371–391, 2007.

[305] Lea M Petters and Marina Schröder. Negative side effects of affirmative action: How quotas lead to distortions in performance evaluation. *European Economic Review*, 130:103500, 2020.

[306] Edmund S Phelps. The statistical theory of racism and sexism. *The american economic review*, 62:659–661, 1972.

[307] Catherine Porter and Danila Serra. Gender differences in the choice of major: The importance of female role models. *American Economic Journal: Applied Economics*, 12:226–254, 2020.

[308] Curtis R Price. Gender, competition, and managerial decisions. *Management Science*, 58:114–122, 2012.

[309] Holger A Rau. The disposition effect and loss aversion: Do gender differences matter? *Economics Letters*, 123:33–36, 2014.

[310] Ernesto Reuben, Pedro Rey-Biel, Paola Sapienza, and Luigi Zingales. The emergence of male leadership in competitive environments. *Journal of Economic Behavior Organization*, 83:111–117, 2012.

[311] Ernesto Reuben, Paola Sapienza, and Luigi Zingales. How stereotypes impair women's careers in science. *Proceedings of the National Academy of Sciences*, 111:4403–4408, 2014.

[312] Ernesto Reuben and Krisztina Timko. On the effectiveness of elected male and female leaders and team coordination. *Journal of the Economic Science Association*, 4:123–135, 2018.

[313] Ernesto Reuben, Matthew Wiswall, and Basit Zafar. Preferences and biases in educational choices and labour market expectations: Shrinking the black box of gender. *The Economic Journal*, 127:2153–2186, 2017.

[314] Deborah L Rhode. *Women and leadership*. Oxford University Press, 2017.

[315] Peter A Riach and Judith Rich. An experimental investigation of sexual discrimination in hiring in the english labor market. *Advances in Economic Analysis Policy*, 5, 2006.

[316] Patrick Ring, Levent Neyse, Tamas David-Barett, and Ulrich Schmidt. Gender differences in performance predictions: Evidence from the cognitive reflection test. *Frontiers in psychology*, 7:1680, 2016.

[317] M Fernanda Rivas. An experiment on corruption and gender. *Bulletin of Economic Research*, 65:10–42, 2013.

[318] Dan-Olof Rooth. Automatic associations and discrimination in hiring: Real world evidence. *Labour Economics*, 17:523–534, 2010.

[319] Silvia Saccardo, Aniela Pietrasz, and Uri Gneezy. On the size of the gender difference in competitiveness. *Management Science*, 64:1541–1554, 2018.

[320] Anya C Samak. Is there a gender gap in preschoolers' competitiveness? an experiment in the us. *Journal of Economic Behavior Organization*, 92:22–31, 2013.

[321] Anya Samek. Gender differences in job entry decisions: A university-wide field experiment. *Management Science*, 65:3272–3281, 2019.

[322] Heather Savigny. Women, know your limits: Cultural sexism in academia. *Gender and education*, 26:794–809, 2014.

[323] Burkhard C Schipper. Sex hormones and competitive bidding. *Management Science*, 61:249–266, 2015.

[324] Arthur Schram, Jordi Brandts, and Klarita Gërxhani. Social-status ranking: a hidden channel to gender inequality under competition. *Experimental economics*, 22:396–418, 2019.

[325] Renate Schubert, Martin Brown, Matthias Gysler, and Hans Wolfgang Brachinger. Financial decision-making: are women really more risk-averse? *American economic review*, 89:381–385, 1999.

[326] Renate Schubert, Matthias Gysler, Martin Brown, and Hans-Wolfgang Brachinger. Gender specific attitudes towards risk and ambiguity: an experimental investigation, 2000.

[327] Christiane Schwieren. The gender wage gap in experimental labor markets. *Economics Letters*, 117:592–595, 2012.

[328] Christiane Schwieren and Matthias Sutter. Trust in cooperation or ability? an experimental study on gender differences. *Economics Letters*, 99:494–497, 2008.

[329] Christiane Schwieren and Doris Weichselbaumer. Does competition enhance performance or cheating? a laboratory experiment. *Journal of Economic Psychology*, 31:241–253, 2010.

[330] Smriti Sharma. Gender and distributional preferences: Experimental evidence from india. *Journal of Economic Psychology*, 50:113–123, 2015.

[331] Olga Shurchkov. Under pressure: gender differences in output quality and quantity under competition and time constraints. *Journal of the European Economic Association*, 10:1189–1213, 2012.

[332] Olga Shurchkov and Alexandra V M van Geen. Why female decision-makers shy away from promoting competition. *Kyklos*, 72:297–331, 2019.

[333] Robert Slonim and Pablo Guillen. Gender selection discrimination: Evidence from a trust game. *Journal of Economic Behavior Organization*, 76:385–405, 2010.

[334] Deborah A Small, Michele Gelfand, Linda Babcock, and Hilary Gettman. Who goes to the bargaining table? the influence of gender and framing on the initiation of negotiation. *Journal of personality and social psychology*, 93:600, 2007.

[335] Erik Snowberg and Leeat Yariv. Testing the waters: Behavior across participant pools. Technical report, National Bureau of Economic Research, 2018.

[336] Sara J Solnick. Gender differences in the ultimatum game. *Economic Inquiry*, 39:189–200, 2001.

[337] John L Solow and Nicole Kirkwood. Group identity and gender in public goods experiments. *Journal of Economic Behavior Organization*, 48:403–412, 2002.

[338] Stuart Soroka, Elisabeth Gidengil, Patrick Fournier, and Lilach Nir. Do women and men respond differently to negative news? *Politics Gender*, 12:344–368, 2016.

[339] Thomas Sowell. *Affirmative action around the world*. Yale University Press, 2008.

[340] Sigrid Suetens and Jean-Robert Tyran. The gambler's fallacy and gender. *Journal of Economic Behavior Organization*, 83:118–124, 2012.

[341] Matthias Sutter, Ronald Bosman, Martin G Kocher, and Frans van Winden. Gender pairing and bargaining—beware the same sex! *Experimental Economics*, 12:318–331, 2009.

[342] Matthias Sutter and Daniela Glätzle-Rützler. Gender differences in the willingness to compete emerge early in life and persist. *Management Science*, 61:2339–2354, 2015.

[343] Kurtis J Swope, John Cadigan, Pamela M Schmitt, and Robert Shupp. Personality preferences in laboratory economics experiments. *The Journal of Socio-Economics*, 37:998–1009, 2008.

[344] Ruth Cadaoas Tacneng and Klarizze Anne Martin Puzon. Gender priming in solidarity games. 2021.

[345] Dawn Langan Teele, Joshua Kalla, and Frances Rosenbluth. The ties that double bind: social roles and women's underrepresentation in politics. *American Political Science Review*, 112:525–541, 2018.

[346] Metin Tetik. Investigating factors affecting cooperative and non-cooperative behavior: An experimental game in the classroom. *Theoretical and Applied Economics*, 27:205–214, 2020.

[347] Kyle A Thomas and Scott Clifford. Validity and mechanical turk: An assessment of exclusion methods and interactive experiments. *Computers in Human Behavior*, 77:184–197, 2017.

[348] Peter Van der Windt, Macartan Humphreys, and Raul Sanchez de la Sierra. Gender quotas in development programming: Null results from a field experiment in congo. *Journal of Development Economics*, 133:326–345, 2018.

[349] Jordan van Rijn, Esteban J Quiñones, and Bradford L Barham. Empathic concern for children and the gender-donations gap. *Journal of behavioral and experimental economics*, 82:101462, 2019.

[350] Donald Vandegrift and Abdullah Yavas. Men, women, and competition: An experimental test of behavior. *Journal of Economic Behavior Organization*, 72:554–570, 2009.

[351] Sofia B Villas-Boas, Rebecca LC Taylor, and Elizabeth Deakin. Effects of peer comparisons on low-promotability tasks: Evidence from a university field experiment. *Journal of Economic Behavior & Organization*, 158:351–366, 2019.

[352] Michael S Visser and Matthew R Roelofs. Heterogeneous preferences for altruism: Gender and personality, social status, giving and taking. *Experimental Economics*, 14:490–506, 2011.

[353] Frauke von Bieberstein, Stefanie Jaussi, and Claudia Vogel. Challenge-seeking and the gender wage gap: A lab-in-the-field experiment with cleaning personnel. *Journal of economic behavior organization*, 175:251–277, 2020.

[354] Mark Van Vugt and Wendy Iredale. Men behaving nicely: Public goods as peacock tails. *British Journal of Psychology*, 104:3–13, 2013.

[355] Jianxin Wang, Daniel Houser, and Hui Xu. Culture, gender and asset prices: Experimental evidence from the us and china. *Journal of Economic Behavior Organization*, 155:253–287, 2018.

[356] Alice Wieland and Rakesh Sarin. Domain specificity of sex differences in competition. *Journal of Economic Behavior Organization*, 83:151–157, 2012.

[357] Alice Wieland, James Sundali, Markus Kemmelmeier, and Rakesh Sarin. Gender differences in the endowment effect: Women pay less, but won't accept less. *Judgment Decision Making*, 9, 2014.

[358] Alistair Wilson and Emanuel Vespa. Paired-uniform scoring: Implementing a binarized scoring rule with non-mathematical language, 2018.

[359] David Wozniak, William T Harbaugh, and Ulrich Mayr. The menstrual cycle and performance feedback alter gender differences in competitive choices. *Journal of Labor Economics*, 32:161–198, 2014.

[360] Danny Yagan. Supply vs. demand under an affirmative action ban: Estimates from uc law schools. *Journal of Public Economics*, 137:38–50, 2016.

[361] Masayuki Yagasaki and Makiko Nakamuro. *Competitiveness, Risk Attitudes, and the Gender Gap in Math Achievement*. RIETI, 2018.

[362] Niklas Zethraeus, Ljiljana Kocoska-Maras, Tore Ellingsen, B O Von Schoultz, Angelica Linden Hirschberg, and Magnus Johannesson. A randomized trial of the effect of estrogen and testosterone on economic behavior. *Proceedings of the National Academy of Sciences*, 106:6535–6538, 2009.

[363] Y Jane Zhang. Culture, institutions and the gender gap in competitive inclination: Evidence from the communist experiment in china. *The Economic Journal*, 129:509–552, 2019.