Experimental Overview and Analysis Plan

Overview:

We investigate evidence of bias against female economists in the way the public evaluates the importance of their research. To measure gender bias, we provide participants with a summary of research co-authored by a female and male economist, and ask them to evaluate it, randomizing the information given to the participants about the gender of the author. The experiment incorporates a between-subject design in which participants are randomly assigned to a group where the research is presented to them as being conducted by a female economist and another group where the research is associated with a male economist. We signal gender using gender-specific pronouns, without referring to gender explicitly or mentioning the authors' names. We provide participants with two statements regarding the research being convincible and important and ask them to indicate whether they agree or disagree with the statements on a 5-point Likert scale. The experiment takes approximately 5 minutes. All participants are given £0.75 (approximately $0.91) for completing the survey. In addition, participants can earn up to £0.5 (approximately $0.61) as a bonus based on their attention to the content of the survey. The payments are in line with standard hourly experimental rates. Our focus in this experiment is on the attitudes of men, therefore we only recruit male subjects.

Experimental Design

The experiment consists of three stages. In the first stage, we provide participants with a summary of a study co-authored by a female and male economist. The choice of the paper is a challenge since the paper should be both understandable and interesting for the general public to make them pay attention and provide their thoughts. We choose a paper by Bertrand and Mullainathan (2004) describing a field experiment on labour market discrimination. Published in AER and being a well-known paper among the studies on discrimination, this paper has several advantages: First, the subject of the paper, racial discrimination in the labour market, is familiar to the public. Second, the methodology the authors use is straightforward and understandable. Third, the findings of the paper, which quantify the level of discrimination against African-Americans, are quite striking. These factors make the paper a well-suited candidate for the purpose of our experiment. The summary presented to the participants is a simplified version of the abstract of the paper, which includes the aim, methodology, and the main findings of the paper. In a pilot study, we asked participants to write their ideas about the research in at least 50 words. Careful analysis of the responses suggests that participants indeed understand the main points of the research.

The participants are randomly assigned to two groups. In both groups, we inform the subjects that they will be presented with a paper written by a professor of economics. In each group, the author is signalled to be a male or a female economist, using gender-specific pronouns in the description of the author. We avoid mentioning the name of the author to remove the possibility of racism affecting judgements. Also, we do not refer to gender explicitly. After reading the description of the author, participants listen to a summary of the paper. We have chosen to

provide participants with an audio track describing the research instead of asking them to read a summary, to make it hard for the participants to search the internet and find the paper online. The audio track is less than 45 seconds. We use online text-to-speech software that employs voices created by Artificial Intelligence for converting text to audio. The AI reader is male-sounding in the group where the research is attributed to a male economist and is female-sounding in the group where the research is associated with a female economist. We further signal the gender of the economist in the research summary by using gender-specific pronouns whenever we want to mention the author.

We present participants with two statements about the paper and ask them to express if they agree or disagree with them on a 5-point Likert scale. These statements are "The author provides convincing evidence for the presence of differential treatment by race in the U.S. labor market." (Hereafter, *the research persuasiveness* statement) and "The research has important policy implications and policymakers should act upon them." (Hereafter, *the research importance* statement). The two statements capture two different aspects of the subjects' reactions to the paper; whereas the first statement assesses subjects' opinions about the author being able to convince them, the second statement asks if the subjects think the research is important. Having both the statements allows us to analyse if gender bias in evaluating the research as an important study is related to female economists being less convincing than their male counterparts, or it persists even if people find women equally convincing (or even more convincing) compared to men.

To incentivize the participants to pay attention to the content carefully, prior to presenting the economist's description and the summary, we let the participants know they have a chance to earn a bonus payment based on their performance in answering questions related to the delivered content in a later stage of the survey. These questions are presented on a separate page after the participants evaluate the summary. One question asks participants to indicate which of the three options is true about the paper. This question evaluates participants' comprehension, and we refer to it as *the comprehension question* hereafter. The other question asks participants to indicate what the gender of the economist they read about is. This question assesses whether the participant has noticed the gender of the author while answering the questions. We also allow participants to choose the option "I don't remember" for both the questions. Including these questions allows us to assess the attention of the participant. The bonus payment is split between the two questions in both groups.

In the next stage, we try to elicit participants' explicit beliefs about economists being reliable experts, and female economists being less reliable than male economists, by requiring the participants to answer two questions that explicitly ask them about their views on the reliability of economists' opinions in general, and female economists specifically. Particularly, we ask subjects to indicate their level of agreement/disagreement with the statement "In general, economists' opinions regarding economic issues are reliable and worthy of note", on a 5-point Likert scale. We randomize how we present the second question about explicit opinions on female economists. Participants randomly read the statement "It is possible that on average, male economists are more reliable than female economists" or "It is possible that on average, female economists are more reliable than male economists", and are asked to indicate their opinion about

it on a 5-point Likert scale. Through this randomization, we can infer if the subjects are trying to avoid appearing biased by assessing whether the respondents' answers differ as a result of changing the phrasing of the question.

In the final stage, we ask the participants a set of demographic questions that might interfere with their answers in the first stage. Particularly, we ask participants to indicate their gender, age, race, level of education, and political attitude. We also ask them if they consider themselves knowledgeable regarding economics. In sum, our design allows us to measure biased beliefs toward female economists as the main hypothesis. We can further control for demographic variables such as the gender of the participant.

Participants

Our participants will be recruited from the Prolific platform. We choose the participants from the pool of subjects located in the USA. We also limit the subjects to those who have an approval rate of at least 95% and have a minimum of 100 previous submissions. In line with the aim of our experiment, we recruit only male subjects. We also exclude participants who had taken part in the previous pilot studies. We estimate that the experiment will take about 5 minutes. The maximum time for completion of the experiment will be 30 minutes, a restriction set by the Prolific platform.

Exclusion Criteria

We exclude participants who fail to answer the comprehension question, as it is an indication of not paying enough attention to the content of the research. We also exclude any participant whose answer to the question regarding the gender of the author is wrong. We will remove those who don't remember the gender of the author in one set of analyses but keep them in another, because if the bias is subconscious, they don't need to remember the gender of the author consciously. It is also possible that participants pretend not to remember the gender to mask their biased beliefs.[1]

Hypotheses

<u>Primary hypotheses:</u>

Male subjects are less likely to believe that the research is important if they think it is being conducted by a female economist.

<u>Secondary hypotheses:</u>

Subjects tend to avoid appearing gender biased, as they answer differently to a statement implying male economists are more reliable than female economists, if the statement is phrased the other way around.

---

[1] We included a separate attention test, unrelated to the content of the research, in our pilot study. 100% of participants got it right. We remove it in the final experiment as the comprehension question about the study is a tougher attention check, and therefore it is a binding constraint.

Note that our secondary hypothesis is a robustness check for our main hypothesis. If there is a difference in the way the subjects respond to the statements, this is a signal that we are underestimating the amount of bias in the main hypothesis.

Analyses

We start our analysis by providing descriptive statistics of the collected data. We will summarize the data by using histograms wherever possible.

We provide the results graphically using bar charts. We run t-tests to test if there is a difference between the average agreement level in responding to the main outcome (the *research importance* statement) across the two groups.

We will also run regressions to test for the primary hypothesis. We control for the responses to the *research persuasiveness* statement. In robustness checks, we will control for other demographic variables and beliefs about the reliability of economists. We will also check for the robustness of our results to removing participants who did not remember the gender of the economist in attention checks. Note that to minimize concerns about multiple comparison testing, we have only a single primary hypothesis and this will be the focus of our paper. However, we will also consider our secondary hypothesis and controls as mentioned.

References

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. American economic review, 94(4), 991-1013.