Appendix 2

This appendix has been uploaded prior to the collection of stage 2 data.

Details about Stage 2

Stage 2 will be completed by around 600 people (300 Programmers and 300 HR people who work in tech.) who we will refer to as employers. It consists of an approximately 30 minute survey that includes details about: the job advertisement and where it was posted, the python test and about how the Python test score was calculated. At the time when these details are provided, employers must also complete multiple control questions to check their understanding of the task. Employers are then shown 5 male and 5 female profiles in a random order.[1] Each profile contains the first name and initial of the last name of the candidate, highest education level, whether they are currently studying or working, if working their occupation, years of programming experience, and where they learnt programming. An example of a profile is shown in Table 1. For each profile, employers must guess the applicants programming test score (task 1) and also what they think is the guess of a randomly selected person who is similar in terms of occupation to them (task 2). Each employer makes 20 guesses (10 for task 1 and 10 for task 2). Employers are paid based on a binarized scoring rule, where one of the 20 guesses is selected for payment. In addition, in all treatments, there will be 2 bonus guesses (1 for task 1 and 1 for task 2) for one of two randomly selected applicant profiles. Those two profiles are identical except for one having a female name and the other having a male name. These guesses are incentivized separately using the same binarized scoring rule. Following this task, employers answer a set of demographic and attitude questions.

Table 1: Example of a profile.

Name	Benjamin C.
Highest Education level	Graduated 4-year college
Currently Studying	No
Currently Working	Part Time in Software Development
Years of Programming Experience	8
Learned Programming from	University and self-taught

Main Analysis and Outcomes:

1. **Actual Gender Skill Gap:**
   We compare the average python programming score of males relative to females.
2. **Perceived Gender Skill Gap:**
   We compare the average difference in beliefs about the programming test scores of male and female applicants.
3. **Gender bias:**
   Difference between perceived and actual gender skill gap (as described above).
4. **Perceived Gender Gap with Second Order Beliefs**

[1] Note, we reduced the profile number to 10 as a result of time constraints and concern over subject fatigue.

The average difference in second order beliefs (i.e. what employers believe other employers guessed for that profile) of male profiles and female profiles. These second order beliefs can also be considered as a descriptive norm.

5. **Gender Bias in Second Order Beliefs**
   Difference between gender gap in second order beliefs and actual gender skill gap (as described above).

Secondary Analysis and Outcomes

Heterogeneity:

We focus on the following heterogeneity:

- HR vs Programmer in the employer sample
- Control vs Aptitude Test variant vs Personality Test Variant
- Male vs Female employers
- We will collect a set of predictors of the employers including, age, years of experience, gender of their supervisors, as well as the gender composition of their industry and workplace. We explore heterogeneity along these dimensions.

Robustness checks

We will perform the following robustness checks:

- We also estimate primary outcomes 2-5 excluding profiles for which names fail to unambiguously signal candidates' gender. We will identify those names using a separate survey.
- We estimate primary outcomes 2 and 4 (perceived skill gap and perceived gap in second order beliefs) using only the bonus guesses. Those guesses were for profiles of applicants with the same characteristics except for one having a male name and the other a female name.
- As discussed in the original pre analysis plan, if we find evidence for order effects we will perform additional analyses focusing on the first guess of each employer and the first guesses of each employer that showed applicants of the same gender (remember employers see 5 male and 5 female applicants in random order).

Beliefs about gender differences in the population:

We test whether employers have accurate beliefs about the gender difference in Information Communication Technology (ICT) skills in the US population, and whether or not they believe that these are changing over time. To measure the accuracy of beliefs at the population level, we record beliefs about 1) ICT skills among US high school students and 2) gender composition of those who pursue a career in programming in the US. In particular, employers are told about a United States national test on 8th grade students' ICT skills that was conducted in 2014 and 2018. Employers are then asked to guess the score of boys and girls in the 90^{th} percentile as well as boys and girls in the 50^{th} percentile. The scores range between 0-100. Half our employers are

asked about the 2014 study and the other half about the 2018 study. Employers are paid $2 based on the binarized scoring rule. One of the four questions are randomly selected for payment. Employers are given a maximum of 120 seconds to complete the questions. Secondly, we ask employers about the gender composition of programmers in the US according to data from the census bureau. Payment of $2 is based on the binarized scoring rule. Employers are given a maximum of 120 seconds to complete the questions. Comparing these beliefs with the actual figures allows us to infer any inaccurate belief at the population level and in the case of ICT skills, whether people believe that these are changing over time.

Mechanisms:

We include several questions which may help to explain why beliefs may differ from actual test scores. We describe these questions and the potential mechanisms below.

Mechanism 1: Representative Heuristic

Tversky and Kahneman (1983) and more recently Bardalo et al (2016, QJE) argue that people's beliefs may be inaccurate because they use extreme points in the distribution as representative of the group. For this to be an accurate explanation of any gender bias, we would expect at least the following to hold:

1) There are gender differences at the top or bottom end of the distribution

2) Employers believe there are gender differences at the top end or bottom end of the test score distribution

To test for this mechanism, we do two things: to test for 1) we test for gender differences in the extremes of the gender-specific test score distribution. More specifically, we test whether the score of the woman who scored in the 90^{th} percentile (10^{th} percentile) of all female test-takers differs from the score of the man who scored in the 90th percentile (10^{th} percentile) of all male applicants.

To test for 2) in the attitude section of the survey, we ask employers to guess the scores of women in the 90^{th} and 10^{th} percentile of all female test-takers and the scores of men in the 90^{th} and 10^{th} percentile of all male test-takers. Each participant makes 4 guesses. One of those guesses is randomly selected for payment. Subject payment is $2 based on the binarized scoring rule.

Mechanism 2: Attention Discrimination

Bartos et al (2016, AER) show that endogenous costly attention can magnify the impact of prior beliefs about group quality. To test this mechanism we record the time spend on each profile and test whether employers spend more time on male or female applicant profiles.

Mechanism 3: Selection neglect

People could have accurate beliefs about gender differences in programming ability in the population but ignore that job applicants represent a selected sample of the population which may have different gender gaps.

To study whether employers exhibit selection neglect, we ask them about their beliefs about gender differences in coding ability in the population and among applicants separately (in random order to avoid ordering effects). We would see evidence of selection neglect if answers to those two questions would be the same.

As additional suggestive evidence, we also study the patterns that emerge from the respondents' beliefs on the gender differences between US high school students' ICT skills (2014), gender composition of programmers in the US and gender differences in coding ability among the applicants (Male>Female in ICT skills/coding ability/gender composition [i.e. average skill in population] AND Male>Female among applicants would be consistent with selection neglect).

Mechanism 4: Descriptive Social Norms

Beliefs may be influenced by descriptive norms i.e. people may believe something because other people believe the same thing. To assess the influence of this mechanism, we test how correlated own beliefs are with other employers' beliefs (the second order beliefs variable described above). High correlations between own and other beliefs would be consistent with the influence of descriptive social norms.

Updated Sample Size:

In the second stage, we plan to elicit 600 surveys, (300 HR and 300 programmers). The sample will be split evenly between the following (note this is the same block as originally specified)

1. Control / Male first
2. Control / Female first
3. Information (Personality) / Male first
4. Information (Personality) / Female first
5. Information (Aptitude) / Male first
6. Information (Aptitude) / Female first