Does Better Information Reduce Gender Discrimination in the Technology Industry?*

Abdelrahman Amer, Ashley C. Craig and Clémentine Van Effenterre

March 2023

Abstract

Does information provision mitigate gender biases in how individuals are assessed in interviews? We investigate this question looking at a high-skilled online labor market in the technology industry where women are particularly underrepresented. Leveraging over 60,000 trial mock interviews from an online peer-to-peer platform for software engineers, we first document that women receive lower subjective coding ratings on the platform than men. In 2017, the platform introduced a new automated code evaluation device, which allowed interviewers to learn about interviewees’ objective coding performance in real time, and use this as input in their final subjective evaluations. We exploit the fact that the device was progressively made available for pairs of users chosen at random, and provide evidence that this objective measure of performance does not reduce the raw gender gap in peers’ subjective ratings. Additionally, we find that for the same level of objective performance, women are evaluated more harshly than men. Our results are not explained by changes in the composition of users on the platform, endogenous matching between users and coding problems, selection or gender differences in performance. To explore mechanisms behind the residual gap in ratings, we run a follow-up online experiment to directly assess whether software developers evaluate a piece of code differently if they know the gender of the coder. Using a large set of de-identified code blocks written by men and women, we investigate whether residual gender gaps in subjective ratings are due to unmeasured differences in code quality, or gender bias.

JEL codes: J16, M51, C93

Keywords: Discrimination; Gender; Coding; Experiment.

*Amer: University of Toronto, 150 Saint George Street, Toronto, ON M5S 3G7, Canada (e-mail: abdelrahman.amer@mail.utoronto.ca) Craig: University of Michigan, Department of Economics, 238 Lorch Hall, 611 Tappan Ave, Ann Arbor, MI 48109 (e-mail: ashcraig@umich.edu) Van Effenterre: University of Toronto, 150 Saint George Street, Toronto, ON M5S 3G7, Canada (e-mail: c.vaneffenterre@utoronto.ca). This paper greatly benefited from discussions and helpful comments from Iris Bohnet, Katherine Coffman, Nicole Fortin, Maria Guadalupe, Peter Hull, Corinne Low, Roland Rathelot, Alexandra Roulet, Nina Roussille, Heather Sarsons, Basit Zafar, and from participants at numerous seminars. We are grateful to the Pivotal fund, the NBER Digitization Program and the Russel Sage Foundation for continuous support. This project received IRB approval at the University of Michigan and at the University of Toronto.
Introduction

Discrimination in labor markets has received considerable attention from policy makers and economists in the past decade (Bertrand and Duflo 2017). Economists in particular have identified imperfect information as one way to rationalize differential treatment of members of different groups such as men and women (Phelps 1972, Spence 1973, Arrow 1973b). In many industries, the typical hiring process comprises several stages, each of which may occur under imperfect information. Recruiters extract information about a candidate through resumes, referrals, test results, interviews or simulation assessments to observe a candidate performing a task in a realistic work context. After hiring, assessment of ongoing performance also affects retention and promotion decisions. At each one of these stages, there could be gender disparities.

The way economists typically detect gender bias in hiring consists in looking at performance measures, e.g. whether gender differences in hiring are "justified" by differences in assessments of performance during recruitment. If managers do not choose men over women with strictly higher assessed performance, they are not biased. A near-universal difficulty in these studies is that assessments of performance may themselves be biased (Bohren et al., 2022). For example in the technology industry, where coding tests are often used as screening task in the hiring process, the score on the coding task may penalize women who actually performed similarly.

Unlike studies who focus on bias in the selection of candidates, our paper is the first to document biases in how individuals are assessed in interviews, and to ask whether information provision can mitigate these biases. We investigate this question looking at a high-skilled online labor market in the technology industry where women are particularly underrepresented (Beede et al. 2011, Ashcraft et al. 2016) and where these performance assessments are often made by expert peers who are themselves programmers. We have partnered with an online peer-to-peer platform that provides software engineers with an opportunity to practice their interview and coding skills.

One key feature of our setting is that we can measure gender gaps in subjective evaluations, compare them to differences in actual performance assessed on the same scale, and attempt to reduce these gaps by providing an additional objective measure of code quality and by gender-blinding. The platform granted us access to users’ ratings by their peers on code quality, creativity, likability and overall performance.
In addition, each record contains data on the coders’ educational backgrounds, demographic characteristics and self-assessments of their level of preparedness. While we do not directly observe users’ subsequent labor market outcomes, our plan is to explore the labor market consequences of these differences in subjective ratings by matching our experimental data with labor market information from Revelio Labs (in progress). In 2017, the platform introduced a new automated code evaluation device, which allowed interviewers to learn about interviewees’ objective coding performance in real time during the mock interview, and to use this as input in their final subjective evaluations. The device was progressively made available for 7 percent of new pairs of users chosen at random. When activated, it indicated diagnostics such as whether the code ran, and whether it produced correct answers for various test cases. This change in information allows us to disentangle the signal of skills (objective rating) and how it is interpreted (subjective rating).

We first document that, prior to the introduction of the device, women received lower subjective ratings than men for coding ability and problem solving, though not for communication. These gender gaps in peers’ assessments of coding ability and problem solving correspond to around 12 percent of a standard deviation, and remain consistent when we control for the interviewee’s and interviewer’s characteristics including their gender, as well as the coding problem’s difficulty. We find that the introduction of the device does not reduce the gender gap in subjective ratings. Both men and women’s subjective ratings improve after candidates are provided with objective coding score. Our results cannot be explained by changes in the composition of users on the platform, endogenous matching between users and problems, selection or gender differences in objective performance only.

To interpret our results, we develop a simple model of statistical discrimination in the spirit of Lundberg and Startz (1983) in which the role of an interviewer is to estimate the ability of each job candidate based on imperfect signals. In our model, holding fixed an interviewer’s prior beliefs about the distributions of coding ability among men and women, providing more information increases the weight the interviewer places on the signal they observe, and reduces the role for preconceptions about gender differences in ability. However, the effect of providing more information depends on whether the interviewer believes that the gender gap in coding ability is larger than
it is in reality. We therefore explore gender differences in objective performance and whether they fully explain differences in ratings. The residual gender gaps in subjective ratings controlling for objective performance measure and other characteristics is larger than 6 percent of a standard deviation. Additionally, we find that both high and low performing women receive systematically lower subjective coding and problem solving ratings than men who perform equally well.

To further disentangle the mechanisms behind the persistence of the gender gap in ratings, we use the quasi-random assignment of coding problems to investigate how the impact of the device introduction on gender gaps in subjective ratings varies depending on the difficulty and ambiguity of the problem solved. We find that for a given problem difficulty, the new procedure improves men’s ratings more than those of women. Overall, these results are consistent with previous studies looking at differences in updating by group (Sarsons, 2022).

The fact that women continue to receive lower ratings than men could indicate the presence of bias, as opposed to evaluators simply having incorrect perceptions about code quality which could have been corrected with more information (Bohren et al., 2019b). We run an online experiment to directly assess whether software developers evaluate a piece of code differently if they know the gender of the coder (in progress). Using a large set of de-identified code blocks written by a set of men and women on the platform, we ask: (i) whether there are perceived differences in the quality of the code written by men and women; and (ii) how those perceived differences change when the evaluator is aware of the gender of the coder. This allows us to test whether the residual gender gaps in subjective ratings stem from unobservable dimensions of performance correlated with gender or from bias. While we are aware of other experiments which consider the effect of blinding with respect to gender alone (e.g., Goldin and Rouse 2000), and the effect of anonymous author code review in a given company (Murphy-Hill et al., 2022), this is the first study investigating whether anonymous author code review reduces gender disparities in code review outcomes. This will also shed light on whether blind recruitment is enough to improve diversity in hiring. Our very preliminary results suggest that while women receive higher ratings on average than men, they only do in the blind condition when their gender is not revealed.

In addition, we plan to explore the labor market consequences of these differences
in subjective ratings by matching our experimental data with labor market information from Revelio Labs (in progress). The raw data contain information about job titles, seniority, and salary information from LinkedIn at each point in time since 2008, and detailed information about these individuals and the companies at which they work. The data will allow us to validate our measures of performance, and study the potential consequences of the gender differences in subjective ratings that we documented.

With this research, we contribute to several literatures. First, we contribute to the labor economics literature on the role of information in the hiring process. Using methodology such as resume audit studies, previous authors have established the existence of discrimination in the labor market (Bertrand and Mullainathan 2004, Neumark 2012, Kroft et al. 2013, Farber et al. 2016). However, it has proven difficult for such studies to separate out rational statistical discrimination, statistical discrimination with incorrect beliefs, and taste-based discrimination. A recent contribution by Bohren et al. (2019b) conceptualizes this identification problem when isolating the source of discrimination, and tests it experimentally. With respect to this study, we are looking at how inaccurate beliefs translate to broader employment outcomes by connecting our subjective ratings with real labor outcomes data using the Revelio Labs data (in progress). Finally, by providing real code excerpts to external evaluators, we attempt to minimize deception prevalent in audit studies, and to make a methodological contribution to experimental studies investigating group-level labor market disparities (Kessler et al., 2019).

Another line of research has investigated factors behind the slow progression of women in high-paying occupations (Bertrand et al. 2010, Goldin 2014, Roussille 2020), and to a growing literature documenting potential causes of under-representation of women in the technology industry specifically (Terrell et al., 2017; Murciano-Goroff, 2018; Miric and Yin, 2020; Boudreau and Kaushik, 2020). Part of the explanation may lie in how information about ability is interpreted in occupations that require different skills. However, ability and performance are usually hard to quantify in high-skilled labor markets. Compared to previous studies which rely on measures of performance such as billable hours for lawyers (Azmat and Ferrer, 2017) or patients’ death for surgeons (Sarsons, 2022), we have access to a problem-specific objective measure of performance for computer programmers. Closer to our paper is the contemporaneous
study by Feld et al. (2022). showing that providing managers with information about non-coding-related skills (aptitude, personality) of job applicants eliminates their perception of a gender gap in performance. While this paper also looks at the role of inaccurate beliefs, and the use of additional information to correct them, our main findings differ. Our evidence so far suggests that evaluators’ inaccurate beliefs about the relative performance of men and women cannot be corrected so easily. Our study focuses on biases in the performance evaluations of the code of workers with similar background and expertise, which matters for assessing performance in that interview. Focusing on the horizontal relationship between workers with similar qualifications is directly relevant to both hiring and promotion decisions in this industry, as horizontal hiring and referrals are prevalent in many high-wage sectors (Oyer and Schaefer, 2011) and contribute to persistent between-group inequality (Miller and Schmutte, 2021; Sarsons, 2022).

Finally, we also contribute to the literature on digitization of labor markets. The global reach of online platforms enables employers to access a larger and potentially more diverse pool of workers (Brynjolfsson et al., 2003). Design choices relying on new technologies have been seen as a way to help mitigate systematic biases that occur in reviews and reputation systems at play in the hiring process (Cowgill, 2018; Bohnet, 2016). However, the increasing use of algorithms to automate decision-making has sparked concerns that these automated choices may produce discriminatory outcomes (Lambrecht and Tucker, 2019; Chan and Wang, 2018; Edelman et al., 2017; Fisman and Luca, 2016). There is even some evidence that algorithmic tools can lead to worse hiring decisions (Hoffman et al. 2018). With this project, we aim to shed light on how an automated evaluation of quality affects the assessment of job applicants and hiring decisions in an online labor market.

The remainder of the paper proceeds as follows. We provide institutional background in Section 1. We propose a simple model of statistical discrimination in Section 2. Key descriptive statistics are presented in Section 3. The experimental design is presented in details in Section 4. We present our main results in Section 5 and robustness checks in Section 6. We discuss potential mechanisms in Section 7. Section 8 concludes.
1 Institutional Background

1.1 Recruitment in the Tech Industry

Recruiters of programmers are in the unique position of being able to test a prospective employee’s ability to solve problems using skills that they would require to succeed in the workplace. For many leading technology companies such as Google and Atlassian, interviews are comprised, at least in part, of coding challenges designed to test the relevant skills. A variety of specialized platforms have been developed for this purpose, including CoderPad, Coderbyte, HackerRank, Codility, Pramp. These companies vary in their business models, ranging from interview practice platforms to those that actively source and screen candidates for specific employers.

1.2 The Platform

A user’s experience begins when they log on and provide information about their background and experience, including their proficiency with the available programming languages. The user is then offered the opportunity to schedule an interview during one of many fixed time slots, with the platform suggesting slots which already have users with similar profiles. When that time arrives, users within the time slot are paired based on their similarity scores using Edmunds’ Blossom algorithm.¹

Each pair of users who are matched interview each other in turn. Depending on the language and self-reported ability and experience of the interviewee, a coding problem is assigned. Candidates can participate in as many different practice interviews as they like and each time, will be paired with a different counterpart. The interviewee then proceeds to solve the coding problem in an online text editor that both sides can see. At the same time, the users communicate via video chat (see Figure A1). Once the interview finishes, the interviewer and interviewee swap roles. At the conclusion of their interaction, each of the two users rates the other on their coding quality, creativity, likability and overall performance.

Between December 18, 2015, and April 18, 2018, users on the platform engaged in 60,513 interviews. Eighteen percent of these users were female, and 81 percent

¹This algorithm chooses a matching that maximizes the total of the similarity scores of paired users.
were male.\footnote{A small fraction of users could not be classified.} The users mainly hail from English-speaking countries, the US, the UK and Australia but also from Europe, Brazil, Chile, India and Russia (Figure A2). The platform’s user base has grown rapidly over time, starting with only a few users per day in early 2016 to around 150 per day in mid-2018 (see Figure A5). For the period of August 2016-March 2018, Table 1 shows that users were participated on average to 12 sessions. 32 different problems were assigned to the participants.

The platform has several appealing features for the study of gender gaps in performance evaluations in a high skilled labor market compare to a more traditional lab experiment. Drawing from users’ online reviews of their experience of the platform, we note first that the platform provides an environment where tasks are performed under time pressure and where stakes are high. One user writes: "I realized early that my biggest challenge wasn’t the coding problems themselves: it was staying focused while solving them out loud in front of an interviewer with time pressure. [The platform] was perfect for practicing in an environment much more like the real interview."

The platform also mimics the competitive environment in which the software developers are recruited, as they are potentially competing for the same jobs. However, the participants have clear incentives to cooperate, as one user writes: "Doing practice interviews with humans who talk to you was much more valuable than working with a review book or online lists of problems. And [the platform] users I paired with were consistently helpful, polite and professional."

Finally, while evaluations are not anonymous, there is no clear risk of retaliation on the platform. Reputation is not observable as past scores are not public, and the setting is not dynamic: participants do not meet again in the platform after they matched once.

\section{A Simple Model of Statistical Discrimination}

We frame our empirical analysis in terms of a simple model of statistical discrimination in the spirit of Lundberg and Startz (1983).\footnote{See also Aigner and Cain (1977) for a related model, and Fang and Moro (2011) for a more general review of the literature on statistical discrimination.} The role of an interviewer is to estimate the performance, $y_i$, of job candidate $i$ based on an imperfect signal of that
candidate’s performance, \( \theta_i \). In the context of the coding interviews we analyze, ability likely encompasses aspects captured by the subjective ratings for problem solving, coding and communication, but potentially also other dimensions of ability.

For simplicity, we assume that interviewers believe that the performance of candidates of gender \( g \in \{m, f\} \) is normally distributed in the population, with mean \( \mu_g \) and variance \( \sigma_g^2 \).

\[
y_i \sim N(\mu_g, \sigma_g^2)
\]

They may believe (correctly or incorrectly) that the mean, \( \mu_g \), and standard deviation, \( \sigma_g^2 \), differ between male and female candidates in the population.

The signal that an interviewer observes is unbiased, but noisy. Specifically, \( \theta_i = y_i + \varepsilon_i \), where \( \varepsilon_i \) is normally distributed with mean zero and variance \( \sigma_\varepsilon^2 \), and is independent of both \( y_i \) and \( g \). The unconditional distribution of \( \theta_i \) is therefore as follows.

\[
\theta_i \sim N(y_i, \sigma_g^2 + \sigma_\varepsilon^2)
\]

This signal summarizes all of the information available to an interviewer when she assigns a rating, including: verbal interaction, observation of the candidate as she performs the assigned coding task, and any objective measures of code quality.

Rational Bayesian inference based on this noisy signal implies that the interviewer uses her belief about the population as well as the information contained in the signal. Specifically, the interviewer’s estimate of the candidate’s performance (and thus her rating) is a simple weighted average of the signal and the group mean:

\[
E[y_i \mid \theta_i, g] = s_g \theta_i + (1 - s_g) \mu_g
\]

where \( s_g = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_\varepsilon^2} \in (0, 1) \) is the weight placed on the signal.

The role of the interviewer’s \textit{ex ante} belief is greater if the signal is less informative.\(^4\) In the extreme case in which the signal is completely uninformative, the interviewer’s estimate of every candidate’s performance is simply her belief about the mean given the candidate’s gender, \( \mu_g \). In contrast, the interviewer’s beliefs about the population distribution of ability would be completely irrelevant if the signal were perfect.

\(^4\)Alternatively, the interviewer will place more weight on her \textit{ex ante} belief if he or she is confident of that opinion in the sense that \( \sigma_g^2 \) is small.
Statistical discrimination arises if an interviewer’s prior belief differs by gender. The rating assigned to a man will then differ from that assigned to a woman for exactly the same interview performance. If we suppose that interviewers believe the variance of ability, $\sigma^2_g$, to be the same for both genders, then $s_m = s_f = s$; in this case, the gender difference in ratings given a fixed signal realization, $\theta_i$, is given by equation 1.

$$\text{Gender Gap } | \theta_i = E[y_i | \theta_i, m] - E[y_i | \theta_i, f] = (1 - s) (\mu_m - \mu_f)$$

Equation (1) shows that interview ratings will partially reflect the interviewer’s preconceptions about the performance levels of men and women. This implies a gender gap that (in this example) is constant and independent of the candidate’s interview performance. This gender gap is larger if $s$ is smaller—i.e., the signal is noisier so that $\sigma^2_\varepsilon$ is larger, or the interviewer’s beliefs are more strongly held so that $\sigma^2_g$ is smaller.

Since the gender gap in Equation (1) is conditional on interview performance, it constitutes discrimination. If interviewers’ prior beliefs are correct, then a prerequisite for such a gap to exist is that there is a true difference in average coding ability between men and women on the platform. However, it is also possible that the difference between $\mu_m$ and $\mu_f$ reflects a mistaken belief (a “bias”).

**Adding Taste-Based Discrimination**

If there is also taste-based discrimination, then the gender gap in Equation (1) can be adjusted as follows:

$$\text{Gender Gap } | \theta_i = (1 - s) (\mu_m - \mu_f) + \tau_m$$

where $\tau_m$ is the relative adjustment favoring men.

This taste-based adjustment covers at least two interesting cases. First, interviewers may simply discriminate against women. Second, they may refuse to statistically discriminate against women. In this latter case, $\tau_m$ would at least partially offset the statistical discrimination that would otherwise be present.
2.1 Providing Additional Objective Information

The provision of additional objective information to interviewers through the introduction of the device can be thought of as increasing the precision of the signal, \( \theta_i \). Specifically, the intent of the intervention was to reduce the variance, \( \sigma_i^2 \), in the signals that interviewers observe about candidates.

Holding fixed an interviewer’s prior beliefs about the distributions of coding ability among men and women, providing more information in this manner increases the weight the interviewer places on the signal they observe, and reduces the role for preconceptions about gender differences in ability.\(^5\) In Equation (1), \( s \) moves closer to one, and the gender gap conditional on any particular signal realization declines.

The effect of imperfect information on the unconditional gender gap depends on whether interviewers have correct beliefs about the gender difference in ability among the platform’s users, and whether they engage in taste-based discrimination. Letting \( \mu_g^* \) be the true average ability of gender \( g \) candidates, the unconditional gap in performance ratings is given by Equation (3).

\[
\text{Gender Gap} = s \left( \mu_m^* - \mu_f^* \right) + (1 - s) \left( \mu_m - \mu_f \right) + \tau_m
\]

The effect of providing more information (increasing \( s \)) depends on whether interviewers believe that the gender gap in coding ability is larger than it is in reality. In this sense, a finding that the intervention we study reduces the unconditional gap in interview ratings would simultaneously provide evidence of bias, and a cheap and effective solution to that bias. In contrast, an unchanged gender gap would be consistent with interviewers being unbiased in the sense that their prior beliefs are, on average, centered on the truth. Note that it would also be consistent with interviewers refusing to statistically discriminate, even though they believe (correctly or incorrectly) that there are gender differences in average performance.

\(^5\)The distributions of coding ability need not be invariant to the information structure, since less precise information undermines the incentive for an individual to become more productive. See Craig (2019) for an analysis of this issue. Building on Arrow (1973a), Coate and Loury (1993) demonstrate that imperfect information can lead to purely self-fulfilling gaps in productivity.
3 Descriptive Statistics

We first provide descriptive statistics on the profile of users in Table 1. Our setting is characterized by the fact that participants are high-skilled and for the vast majority graduated in STEM fields. One third of participants had Masters degrees, and nearly all of the remaining users held a bachelor’s degree (see Figure A3). Two thirds of users had computer science degrees, with most of the remainder spread between engineering, mathematics, statistics and the hard sciences (see Figure A3 and Figure A4). Women represent about 17 percent of users on the platform. Consistent with evidence from Murciano-Goroff (2018), we find that on average women declare lower level of preparation before the intervention.

Using the universe of connections from January 2016 to July 2017, we are able to document the fact that there is a pre-intervention gender gap in subjective Table 2. Women received lower ratings for coding ability and problem solving, though not for communication. These gender gaps in coding ability and problem solving assessments represent about 12 percent of a standard deviation. They remain consistent when we control for the interviewee‘s and interviewer’s level of education, years of experience and self-declared preparation level, and the gender of the interviewer, consistent with recent studies challenging the notion that female job applicants will be evaluated more favorably when they are paired with female versus male interviewers (Rivera and Owens, 2015). Finally, these gender gaps do not vary substantially by problem difficulty (see Figure B7). They also persist when we add date fixed effects to take into account changes in composition as the platform grew. Of course, without further evidence, this gender gap in ratings is consistent with a gender gap in performance, discrimination, or some combination of the two.

4 Experimental Design

4.1 Intervention

On July 8, 2017, the platform introduced a new feature that provides objective information on code quality to users. This change meant that the platform now provides direct tests of code quality via a series of automated tests (unit tests) during the in-
terview. Users can choose to activate a test by pressing a button (see Figure A1), and the evaluation tool is then visible to both the interviewer and the interviewee. Finally, users can run the tests, and observe pass/fail outcomes. We view this as equivalent to increasing the precision of the signal, $\theta_i$, in our theoretical model.

4.2 Treatment Assignment

The platform introduced the new code quality tests for only a subset of its users. As shown in Figure 2, the intervention was phased in gradually over time in a partially-randomized manner. The share of users treated at least once increases from July 2017 until all users are treated in October 27, 2017. During the staggered implementation (July 2017-Oct 2017), we collect data for 6,401 sessions and respectively 3,167 interviewees.

Figure A6 details how new users are assigned to treatment or control conditions as they enter the platform during the phase-in period. When a new user $i$ is paired to another user $j$, one of two configurations arises. First, for pairs where both $i$ and $j$ are new users or who have only been in the control condition in the past, the pair is randomized into treatment with a 7 percent probability. Once treated, a user always remains in treatment for all future interactions. Thus, any candidate matched with a partner in the treatment condition will automatically be treated as well. If $i$ is matched to $j$ who is already in the treatment condition, $i$ therefore becomes treated without randomization. We deal with this potentially imperfect randomization in Section 6.

In Table 3, we test for balance in the experimental sample. We find that the baseline characteristics are reasonably balanced between the treated and the control group. However, users’ experience with the platform might differ between treatment and control, as treatment is an absorbing state. Therefore, in additional specifications, we choose to also control for date fixed effects, and in certain specification control for the level of presence on the platform.

4.3 Identification Strategy

If all users had activated the objective code quality measures when they were available, our design would have allowed us to directly estimate average treatment effects
by comparing outcomes between users in the treatment and control groups. However, users had to choose to activate the device during the interview, and not all did so. We therefore begin by estimating an Intention-to-Treat (ITT) model:

\[ Y_{it} = \alpha + \beta T_{it} + \theta_t + \epsilon_{it} \] (4)

where \( Y_{it} \) is the score of individual \( i \) for a session on date \( t \), and \( \theta_t \) are date fixed effects. \( T_{it} \) corresponds to a pair of users for whom the new feature was "enabled". The estimation of the \( \beta \) coefficient from Equation (4) corresponds to the ITT. Standard errors are clustered at the date level.

Next, we account for the fact that not all users activated the tests by using treatment assignment as an instrument for actual treatment. This allows us to estimate the treatment effect on the treated (TOT). Specifically, we estimate the following model using two-stage least squares (2SLS):

\[ Y_{it} = \gamma + \delta D_{it} + \lambda_t + \eta_{it} \] (5)
\[ D_{it} = \mu + \pi T_{it} + \zeta_t + \nu_{it} \] (6)

where \( Y_{it} \) is the outcome of user \( i \) at time \( t \); \( D_{it} \) is a dummy variable indicating whether the user activated the objective tests; and \( T_{it} \) is an indicator of whether the pair was assigned to treatment; and \( \lambda_t \) and \( \zeta_t \) are time fixed effects. Standard errors are again clustered at the date level.

5 Results: A Persistent Gender Gap in Evaluations

We begin our analysis studying the impact of the introduction of the device on activation decision and on gender gaps in subjective ratings. We then look at whether differences in objective performance affect differences in ratings.

5.1 Activation and Effect on Subjective Ratings

Estimates of Equation (4) and Equation (5) are shown in Figure 3, and corresponding Table 4. Panel A shows results for all users, then Panels B and C show results for
men and women separately. For each outcome, the first column of the top sub-panel present ITT estimates of Equation (4), and the second column presents 2SLS estimates Equation (5). The first stages Equation (6) are summarized in the lower sub-panels.

**Activation.** The first stage estimates indicate that 71 percent of users enabled the objective code quality tests when they were made available by the platform. This is a strong first stage, and suggests that the code quality ratings were observed and valued by participants on the platform. Additionally, we observe a lower first stage for women (0.678, S.D=0.016) than for men (0.721, S.D=0.016), consistent with evidence of gender difference in feedback aversion.

**Complier Covariates.** Participants on the platform could choose to activate the device, therefore we characterize compliers by observable characteristics in Table 5. As explained in Abadie (2003), these characteristics can be recovered by calculating the fraction of compliers in different subsamples. Column (1) corresponds to the first stage regression for each specific group. Column (2) is the frequency of the group in the estimation sample. Columns (4) and (5) show the two available estimates of complier means in the untreated and treated conditions for the list of baseline characteristics. These results come an IV procedure where the dependent variable corresponds to $X_iD_i$ (Column 4) and $X_i(1-D_i)$ (Column 5), using $T_i$ as an instrument for $D_i$. As discussed by Angrist et al. (forthcoming), the difference between the two estimates is proportional to the difference in characteristics by treatment status used as a balance check. As expected given the balance checks in Table 3, the treated and untreated complier estimates are very similar for all characteristics. Column (5) presents mean characteristics for never-takers for comparison with the compliers. We estimate never-taker means by regressing $X_i(1-D_i)T_i$ on $(1-D_i)Z_i$. The comparisons in Table 5 reveal that for most subgroups, their representation among compliers is similar as in the overall sample. These results confirm the gender gap in activation estimated from Equation 6. The most distinctive feature of the compliers is their gender: compliers are less likely to be women than never-takers. They also have slightly lower level of experience. Along other characteristics such as country of residence, level of education,

---

6Note that always-takers cannot be characterized in our setting because the device is not available for users who are not randomized into treatment.
type of degree, the complier population is very similar to the overall population.

**Subjective Ratings.** The ITT and 2SLS estimates suggest that the availability of objective code quality measures has increased subjective ratings in several dimensions. Both men and women in the treated group receive substantially higher ratings than their peers in the untreated group in problem solving, communication, and hireability ratings. The increases in ratings are generally larger for men, particularly for coding and likability ratings, where the effects are only marginally significant for women. As a result, gender gaps in subjective ratings persist despite the introduction of the device.

5.2 **Objective Performance and Inaccurate Beliefs**

**Objective Performance.** In our model, holding fixed an interviewer’s prior beliefs about the distributions of coding ability among men and women, providing more information increases the weight the interviewer places on the signal they observe, and reduces the role for preconceptions about gender differences in ability. However, the effect of providing more information depends on whether the interviewer believes that the gender gap in coding ability is larger than it is in reality. We therefore explore gender differences in objective performance and whether they fully explain differences in ratings. Figure 5 shows that—conditional on taking the tests, men and women take the same number of tests, but men pass a higher number of tests than women, and therefore the ratio tests solved / tests taken is slightly higher for men.

**Inaccurate Beliefs.** Figure 8 plots the average subjective ratings in coding (Panel A) and problem solving (Panel B) by objective performance (ratio of tests completed over tests passed at 100 or less), separately by gender. This figure shows that both high and low performing women receive systematically lower subjective coding and problem solving ratings than men who perform equally well. The gender gap in subjective ratings is halved for users at the top of the objective performance distribution, but remains statistically significant. These results are confirmed when we control for sociodemographic characteristics of the interviewer and the interviewer as well as date-

---

7 We split the sample in two groups: users who passed all unit tests, and those who didn’t, given the bimodal distribution of the objective performance measure, see Figure B8.
of-interview fixed effects, as shown in Table 7. Compared to the pre-intervention gender gaps, these residual gaps correspond to about 6 percent of a standard deviation.

**Learning.** Interviewers might learn about how the distribution of performance differs between men and women. If evaluators rationally statistically discriminate, this would generally be expected to change the gender gap in their ratings. This seems quite unlikely given the sample sizes that they will observe over the course of the experiment—indeed, calibration of machine learning models suggest that it takes much longer to learn about such gaps in performance (Li et al., 2020)—but it remains a possibility. To test for the role of learning in correcting potential inaccurate beliefs, we look at the persistence of the gender gap conditional on performance when we control for the interviewer’s experience on the platform. Results presented in Table 8 show that the gender gap in subjective ratings does not vanish when we account for the interviewer’s learning on the platform, proxied by the number of past interviews, the number of interviews with female users, or whether the previous interview was with a top performer female users, defined as a female user who performed above the median. Hence our empirical investigation doesn’t support the hypothesis that learning plays a significant role in this context.

The fact that women continue to receive lower ratings than men could indicate the presence of bias, as opposed to evaluators simply having incorrect perceptions about code quality which could have been corrected with more information (Bohren et al., 2019b). However, there could also be other explanations. Specifically, the residual gap in ratings could be driven by: (i) the way women talk about their code in their video interaction, (ii) aspects of code quality not captured by the automated evaluation of it, (iii) taste-based or accurate statistical discrimination. Section 7 expands our results to better understand the mechanisms behind these differences in subjective ratings.

## 6 Robustness Checks

We present a series of robustness checks to confirm the validity of our results in Table 6. We begin in Panel A by estimating the Intention-to-Treat (ITT) model interacted by
gender:

\[ Y_{it} = \alpha + \beta T_{it} + \gamma T_{it} \ast Woman + \theta_i + \epsilon_{it} \]  

(7)

where \( Y_{it} \) is the score of individual \( i \) for a session on date \( t \), and \( \theta_i \) are date fixed effects. \( T_{it} \) corresponds to a pair of users for whom the new feature was "enabled". The estimation of the \( \beta \) coefficient from Equation (4) corresponds to the ITT and \( \gamma \) to the effect on the gender gap in ratings. Standard errors are clustered at the date level.

In Panel B, we introduce month-of-interview fixed effects and date-of-interview fixed effects in Panel C to account for the fact that interviewees might not enter treatment. The magnitude of the treatment coefficient decreases slightly but stays highly statistically significant, while the interacted coefficient with woman remains imprecisely estimated. Finally, we control for a rich set of individual characteristics in Panel D and find virtually the same results.

To make sure our results are not sensitive to the sample period, we also expand our sample to include the pre-treatment period. The size of the coefficients declines slightly but the results are qualitatively the same. We also exploit the staggered introduction of the device in a difference-in-differences framework on the whole period, including month-of-interview fixed effects and find consistent results.

\[ Y_{it} = \alpha + \beta T_{it} \ast Post + \gamma T_{it} \ast Post \ast Woman + \theta_i + \epsilon_{it} \]  

(8)

The results presented in Panel F are remarkably similar to those on the post-treatment period only.

**Endogenous Matching Between Users.** Since the treatment condition is potentially contaminated through the matching process, a naive comparison between treated and control users could provide a biased estimate of the impact of the new feature. To address this threat to the identification, we control our regression results with the propensity score obtained from a matching procedure in Panel G of Table 6. For the matching procedure, we control for month-of-interview fixed effects, and, for both the interviewer and the interviewer, by a dummy variable for each degree level, a dummy variable for each field of study, the number of years of experience, the self-declared...
level of preparedness, as well as gender. Reassuringly, controlling our regressions for the propensity score matching does not affect our results. Additionally, we estimate the propensity score by logistic regression and the Conditional Average Treatment Effect (CATE) directly using a single-equation lasso and find consistent results (results available upon request).

**Changes in the Composition of Users.** Conditional on individual’s covariates and other users’ covariates, treatment assignment should be as good as random. However, we also explore changes in the composition of users on the platform over time. First, the gender composition of the platform users didn’t change drastically after the introduction of the new device, as shown in Figure B10. Second, Figure B11 presents the evolution of first-time users’ characteristics on the platform of over time. We find no evidence of changes after the introduction of the device in terms of users’ probability of being a US citizen, of having a computer science degree, a graduate degree, or no working experience.

7 **Mechanisms**

In this section, we explore and rule out potential mechanisms behind the persistence of the gender gap in subjective ratings.

7.1 **Selection**

A natural questions is whether the treatment increased the pool of qualified women to choose from, and if it can explain our results. We investigate here whether the treatment introduction led to a selection of female applicants. First, the inspection Figure B12 confirms that there are no changes in the characteristics of first-time female users around the date of the introduction of the device on the platform in terms of work experience, educational background or field of study. Second, we look at the evolution of the share of high-performing users among first-time users. We define high-performing first-time users as those who passed all unit tests taken for a given problem during their first interview experience on the platform. Figure B13 plots the

---

8We use the psmath2 command.
shares of high-performing first-time female and male users and shows that they follow a parallel increase over time. While the quality of all first-time users increases over time, it doesn’t increase differentially by gender. As our main specification controls for date-of-interview fixed effects, our results are cannot be explained by positive selection that would affect only one group.

7.2 Gender Differences in Device Use

If men and women were characterized by different abilities to adopt the device, it could potentially explain why the gender gap is not closed after the implementation of the device, but that this effect could take some time to materialize. We explore this possibility by looking at the dynamics of adoption of the device by gender. In Figure 6, we plot the learning curve of both male and female users, measured the number of test passed over time. We dis-aggregate both by number of days and by number of sessions, to account for the fact that women might not be using the platform as frequently as men. Overall, we observe that the learning curves are remarkably similar, and that, if anything, the curve is steeper in Panel B of Figure 6. This suggests that a slower adoption rate cannot in itself explain our results. Additionally, we explore the possibility that the use of the device could be interpreted differently, if users take tests a lot because they have low level of self-confidence for instance, or to the contrary if they want to signal their ability by using the device a lot. Figure 7 shows the average objective coding performance (number of tests completed over test passed) according to the number of tests taken, separately for male and female users, and rejects the hypothesis that men and women’s device uses correlate with differences in objective performance.

7.3 Problem Assignment and Evaluator Type

Additionally, we explore characteristics of the match between participants and problems. If women were systematically assigned easier problems, this could potentially explain why the updating differs by gender. To explore this possibility, we compute the average objective performance of users for a given problem (a high average performance corresponds to a low-difficulty problem). We show in Figure B14 (Panel A)
that problems substantially differ in difficulty levels. Table B2 confirms that, with the exception of interviewer’s years of experience, participants’ characteristics are reasonably balanced across problem’s average difficulty, split by the median ratio of tests solved over tests passed. We also rank problems by the standard deviation of the performance, as shown in Figure B14 Panel B. We define a problem with a high standard deviation of performance as a proxy for a less informative signal about an individual’s performance. As presented in Table B1 columns (1) and (2), the gender of the interviewee does not predict the type of problem assigned, both in terms of difficulty and standard deviation. Additionally, we explore the possibility that the ranking of problems’ difficulty vary by gender. Figure B9 shows the relative ranking of problems’ difficulty by gender. The ranking is proxied by the average performance of users of the same gender for each problem. The orange vertical lines indicate any positive (negative) deviation upward (downward) of female users’ ranking compared to male users’ ranking. We conclude that the two rankings are overall similar. Finally, we explore whether women are more likely to be matched with harsh evaluators. We define a harsh evaluator as an interviewer whose average coding ratings (excluding the session’s rating) is below the median. As presented in Table B1 columns (3) and (4), female users are not more likely to be matched with a harsh evaluator.

7.4 Gender Differences in Communication Styles

A possible explanation for the persistence of the gender gaps in subjective ratings is that men and women could talk about their codes in different ways, and that it could affect the judgement of the evaluator, especially if women appear self-confident in their answers. While we cannot provide direct evidence that this is not the case, we can investigate how women’s communication ratings evolve across the objective performance distribution. Figures 9 plots the average subjective ratings in communication (Panel A) and likability (Panel B) by objective performance (ratio of tests completed over tests passed at 100 or less), separately by gender. While both high and low performing women received systematically lower subjective coding and problem solving ratings than men who perform equally well (Figure 8), Figures 9 shows that the communication and likability ratings of men and women are comparable across the objective performance distribution. This suggests that for a given objective perfor-
mance, gender differences in communication styles are unlikely to explain persistence in gender gaps in coding subjective ratings.

7.5 Problem Difficulty and Precision of the Signal

Finally, we investigate whether the persistence of the gender gaps can be explained by belief updating. We first investigate whether our results vary by problems’ characteristics. We use the quasi-random assignment of the 31 coding problems to investigate how gender gaps in ratings vary depending on the difficulty and ambiguity of the problem solved. We compute again the average objective performance of users for a given problem (a high average performance corresponds to a low-difficulty problem).

For each problem, we estimate Equation (5) separately by gender. Results are presented in Figures B15. First, Figure B15 Panel A shows that the updating is different by gender, across problems of different levels of difficulty. One key parameter of our model is the precision of the signal. Following Bohren et al. (2019a), we explore variations in results across the level of precision of the signal, proxied the standard deviation of performance. Figure B15 Panel B shows that the higher the standard deviation (the less precise the signal), the lower the updating. To gain precision, we group problems by difficulty and precision levels, and estimate again Equation (5) separately for each group and each gender. Results are presented in Figure B16. In Panel A, we document an asymmetric updating pattern by gender and problem difficulty. For men, the improvement in ratings is larger for low-difficulty problems than for high-difficulty problems, although we cannot formally reject that the effects are equal across problems of various difficulty levels. We provide suggestive evidence of a reversed effect for women: the treatment effects are imprecisely estimated for both groups of problems, but the magnitude of the effect is larger for high-difficulty problems. Overall, these results are consistent with previous studies looking at differences in updating by group (Sarsons 2022). Figure B16 Panel B confirms that for low standard-deviation problems (when the signal is more precise), the treatment effect on subjective ratings is higher for both genders, despite being consistently lower and imprecisely estimated for women. These results also provide an indirect test for inattention: if the users were not paying attention to the introduction of the device, they would not have adjusted their beliefs about users’ performance differently according to the precision of the sig-
7.6 Gender Differences in Coding Styles

We propose a follow-up experiment to rule-out the role of gender differences in coding styles. Our goal is to investigate whether, conditional on objective performance information, residual gender gaps in subjective ratings are due to unmeasured differences in code quality, or gender bias. We use a large set of de-identified code blocks written by a set of men and women on the platform. An example of such a code block is shown in Figure C17 Panel B, together with the question (Panel A) and the unit tests (Panel C). This spans coders of different skill and problems of different levels of difficulty. For each code block, we have access to the platform’s objective measures of performance including sub-test results (e.g., whether it runs, whether it produces correct answers to unit tests etc.). Descriptive statistics and demographics of each step of the sample construction are presented in Table C4 and in Table C3. In Table C5, we replicate the results of Table 7 for this sample of codes and find an even larger gender gap in subjective ratings when we control for objective performance.

Using these data, we ask evaluators to judge the quality of the code using the same Likert scales as on the platform. The evaluators are recruited among students computer science who have familiarity in the relevant programming languages.

Depending on the treatment condition, the evaluator will be aware of the gender or other basic information about the programmer who wrote the code (but will never be given identifying information). An example of each treatment condition is presented in Figure C18. Using these evaluations, we ask: (i) whether there are perceived differences in the quality of the code written by men and women; and (ii) how those perceived differences change when the evaluator is aware of the gender of the coder. This lets us test whether there are any unobservable dimensions of performance that are correlated with gender and driving the residual gender gaps that we see in our data.\footnote{We tested whether an AI tool (Chat GPT) was able to predict the gender of the coder of a code when the first name was not displayed, and it was not able to form that prediction.} To examine particular dimensions of performance, we also pre-registered different dimensions of the code (length, time of program execution, number of comments, maintainability of code). Details of the experimental design are presented in
Appendix C.

When gender is unobservable, the evaluator can no longer condition her belief on the gender of the applicant. To form a belief about his or her performance, the relevant belief is therefore the interviewer’s perception of the pooled ability of men and women. Let $\lambda_g$ be the fraction of participants of gender $g \in \{m, f\}$, assume that performance of each gender is normally distributed. Then the pooled belief is:

$$y_i \sim N(\mu, \sigma^2)$$  \hspace{1cm} (9)

where $\mu = \lambda_m \mu_m + \lambda_f \mu_f$ and $\sigma^2 = \lambda_m \sigma_m^2 + \lambda_f \sigma_f^2 + (\lambda_m \mu_m^2 + \lambda_f \mu_f^2 - \mu^2)$.

Conditional on the signal, $\theta$, the posterior belief of a worker’s performance is then:

$$E[y_i \mid \theta_i, g] = \tilde{s}\theta_i + (1 - \tilde{s})\mu$$  \hspace{1cm} (10)

where $\tilde{s} = \frac{\sigma^2}{\sigma^2 + \epsilon^2} \in (0, 1)$ is the weight placed on the signal. Therefore the unconditional gender gap is:

$$\text{Gender Gap} = \tilde{s} (\mu_m - \mu_f)$$  \hspace{1cm} (11)

It is thus clear that there cannot be a gender gap here unless there are true differences in productivity between the two groups.

7.6.1 Outcomes

Our primary outcome is evaluators’ subjective ratings of the code quality. For each block of code, respondents will be asked to rate problems on a scale from 1 to 4. For all primary hypotheses, we will use these responses as our main dependent variable. Our main outcome variable is quite different from call-back rates, which are traditionally used in correspondence studies. First, as discussed by Kessler et al. (2019), call-back rates depend on employers’ interest in a candidate, but also the likelihood that the candidate will accept the job: an employer will not pursue candidates who will be unlikely to accept a position if offered. This is potentially important in the context of workers shortages in the tech industry. Second, callback rates only identify preferences at one point in the quality distribution. In our setting, we will learn about evaluators’ preferences at various levels of the performance distribution, and we focus
on an unusually high-skill segment of the labor market. On the other hand, subjective ratings are harder to directly interpret than call-back rates. However, the Revelio data on future labor outcomes (in progress) will address this limitation.\(^{10}\)

We also have a secondary outcome: evaluators’ prediction of the candidate’s score from the automated evaluation tool. Specifically we ask them how many unit tests out of 10 unit tests do they think were passed. A third outcome variable is evaluators’ prediction of whether a human evaluator decided that this coder passed or failed the interview. Finally, we ask evaluators what is the percent chance that the candidate was later invited for an interview for a role involving coding, which allows us to draw a more direct link between our findings and hiring outcomes.

Additionally, we measure how much time respondents spend on each question to measure fatigue and inattention, and how this varies over time. For example, if discrimination is significantly larger in the latter half of group of code blocks, that would provide suggestive evidence that implicit bias plays a role in our findings.

To measure participants’ priors, we exposed them to three different vignettes before the perform their evaluation tasks. We ask them to predict the potential performance of three different hypothetical coders. We cross-randomize the first name (alternating gender) and the skill level for each vignette (see Appendix C).

### 7.6.2 Incentives

Incentives in our experiment differ from traditional correspondence studies. In part, this is due to our effort to reproduce the incentives and environment faced by participants on the platform. However, it also presents other advantages. First, we do not rely on deception. Participants are clearly informed that these code blocks have been written by real software developers without manipulation, despite the fact that we may not reveal all information. A drawback of this design is that we must inform subjects that responses will be used in research, which could lead to experimenter demand effects (De Quidt et al. 2018), but we think providing real code excepts will reinforce the credibility of our design and encourage participants to exert effort in the evaluation process. A risk is that we ask evaluators to provide subjective ratings on

\(^{10}\)Note that, while our study models only part of the hiring process, bias at an earlier stage such as the coding interview would show up as structural bias in subsequent rounds (Pincus, 1996; Bohren et al., 2022). This remains true even if direct discrimination is present at subsequent stages.
several code blocks, which could lower effort and attention over time. To overcome this, we include incentive-compatible questions where individuals are asked to predict scores from the original evaluators on the platform. The ten best evaluators can win a $500 cash prize. We also provide a symbolic but potentially powerful incentive selecting a set of evaluators to the Creative Destruction Lab 2023 Super Session which brings together world-class entrepreneurs, investors and scientists with high-potential startup founders. CDL Super Session days will provide real networking opportunities and exposure to key players in the industry. Finally, we propose individual feedback on evaluation performance. We expect this to increase the incentive for participants to accurately evaluate the code blocks. University students are not in the position to hire workers or co-workers. Therefore, any residual gender gap in ratings across the blind and non-blind conditions cannot be attributed to homophily, but will reflect valuations of a candidate’s performance only. It is therefore a lower bound for overall discrimination in settings where the evaluator will have ongoing interactions with workers they hire.

7.6.3 Hypotheses Tested

Primary
- H1: Code blocks are evaluated differently if the gender of the coder is known.
- H2: Code blocks written by women are evaluated differently when we reveal the gender of the coder, with the gender gap increasing.
- H3: Individual gender bias varies significantly across evaluators.

Secondary
- H4: The gender identity of the evaluator affects their bias.
- H5: The difficulty of a given coding problem affects evaluator bias.
- H6: The level of the coder’s performance affects the degree of bias.
- H7: Prior bias as assessed by the vignettes correlates with the evaluator’s bias in ratings.
- H8: The characteristics of a given coding problem affects the evaluator’s bias.
- H9: The race of the coder affects the degree of gender bias.

Because we are testing multiple hypotheses, we use techniques that limit the false discovery rate such as correcting the p-values following the standard approach (e.g., Benjamini et al. 2006; Anderson 2008).
7.6.4 Econometric Specifications

To test H1, we will use the following specification:

\[
Y_{ij} = \beta_0 + \beta_1 \times T_{ij} + \beta_2 \times T_{ij} \times NBB_i + \beta_3 \times NBB_i + \sum_{j=1}^{p} \gamma_j (order \times 1_j) + \pi_p(j) + \delta_i + \epsilon_{ij}
\]  

where \( Y_{ij} \) is a discrete variable from 1 to 4 which captures the ratings of evaluator i for code block j; \( T_{ij} \) is an indicator for whether gender is revealed to the evaluator; \( NBB_i \) is an indicator for the randomly assigned "non-blind-blind" condition of the code that the evaluator sees; \( \pi_p(j) \) are problem fixed-effects; and \( \delta_i \) are evaluator fixed effects.

In some specifications, we include controls. Since code blocks characteristics are randomly drawn, including these variables in the analysis should not affect our estimates but could increase precision. Standard errors will be clustered at the evaluator level.

In Equation (12), the coefficient of interest in \( \beta_1 \), which measures the average differences in subjective ratings for code blocks where the gender of the coder is revealed or not, controlling for the order \( NBB_i \). The coefficients \( \gamma_j \) capture the effect of the order in which the code was evaluated, to account for learning and fatigue.

To test H2, we will use the following specification, which is very similar to 12 but interacts the key variables with gender indicators:

\[
Y_{ij} = \beta_0 + \beta_1 \times T_{ij} \times female\_coder_j + \beta_2 \times T_{ij} \times NBB_i + \beta_3 \times female\_coder_j + \beta_4 \times NBB_i + \sum_{j=1}^{p} \gamma_j (order \times 1_j) + \pi_p(j) + \delta_i + \epsilon_{ij}
\]  

The coefficient of interest is \( \beta_2 \), which measures the differential effect of revealing the gender of the coder on subjective ratings, depending on what that gender is.

To test H3, our strategy entails first estimating the difference between male and female average gaps between a non-blind and a blind evaluation, following the literature on teacher grading bias (Terrier, 2020; Lavy and Sand, 2018):

\[
E[Y_i|NB] - E[Y_i|B] = \beta_0 + \beta_1 \times female\_coder_i + \beta_2 \times NBB_i + \pi_p(j) + \epsilon_i
\]  

Here, \( \beta_1 \) is the coefficient of interest. The gender bias is defined as the average gap between non-blind and blind ratings for female-written codes, minus this same gap for male-written code. Because a gender bias is estimated for each evaluator, the number of observations for each estimation is limited. To address the sampling error, we plan
to adapt the framework by Kline and Walters (2021). The authors use empirical Bayes (EB) analysis and data from correspondence studies to identify higher moments of the distribution of job-level callback rates as a function of the number of resumes sent to each job, and propose shape-constrained estimators of these moments. The key behavioral restrictions of their framework can be applied to our Likert-scale measure, so we will use their approach to investigate heterogeneity in gender bias.

To test H4 to H9, we will use variant of Model (13) where treatment effect on gender bias is interacted with, respectively, the gender of the evaluator, the difficulty and characteristics of the code, the coder’s performance, the evaluator’s bias measured through their priors, and the race of the coder.

Our very preliminary results suggest that ratings in the experiment are correlated with those on the platform. Additionally, while women receive higher ratings on average than men, we have some suggestive evidence that they only do so in the blind condition when their gender is not revealed. While these results are still preliminary given that the experiment is ongoing, they suggest that gender biases might be driving our main results.

8 Conclusion

We explore whether providing objective information about a candidate’s performance can affect her subjective evaluations. Using the introduction of a device on a platform that provides candidates with an objective score of their coding performance, we conclude that the introduction of objective measures of performance does not close the gender gap in subjective ratings. Our results are not driven by changes in the composition of users, gender differences learning or matching between users. Using the quasi-random assignment of problem of various difficulty level, we show that for a problem of a given difficulty, the new procedure tends to benefit men more than women. We explore whether ex-ante gender differences in performance can fully explain this result. We provide evidence that women, including high-performing ones, are evaluated more harshly than their objective performance would suggest. An online experiment investigates whether gender difference in coding styles can explain these results.
References


Ashcraft, Catherine, Brad McLain, and Elizabeth Eger, Women in tech: The facts, National Center for Women & Technology (NCWIT), 2016.


Figures and Tables

**Figure 1: Pre-intervention gender gaps – Whole sample**

![Bar chart showing gender gap in perceived performance](image)

*Notes:* This figure shows the gender gap in peer-rated performance in five categories for normalized variables: coding, communication, hirability, likability and problem solving, for the whole sample. Stars above a category indicate statistical significance of the gap at the one percent level, and the 95-percent confidence intervals of each bar are shown in gray.

**Figure 2: Share of treated users on the platform over time**

![Line chart showing share of treated users](image)

*Notes:* This figure shows the evolution over time of the share of users who have been treated at least once on the platform. The red line corresponds to the introduction of the new device on the platform.
**Figure 3: Treatment Effects by Gender**

Impact of treatment by gender

First stage:
- Men: $0.721^{***}(0.011)$
- Women: $0.678^{***}(0.015)$

Notes: This figure shows the estimates of treatment effects by gender for the subjective ratings. The corresponding estimates are presented in Table 4.

**Figure 4: Robustness Checks**

Robustness Check -- Coding Ratings

Notes: This figure shows the results of robustness checks. The corresponding estimates are presented in Table 6.
Figure 5: Gender gap in objective performance after the intervention

Notes: This figure presents the gender gap in objective performance after the intervention in terms of number of tests taken, number of tests solved or failed (right y-axis), and the ratio test solved/passed (right y-axis).
Figure 6: Gender differences in learning

Notes: This figure shows the evolution over time in days (Panel A) and over sessions (Panel B) of the objective coding performance (number of tests completed) of male and female users.
Figure 7: Objective Performance by Number of Tests Taken

Notes: This figure shows the average objective coding performance (number of tests completed over test passed) by how many tests were taken, separately for male and female users.
Figure 8: Subjective Measure by Objective Performance — Coding and Problem Solving

Notes: This figure shows the average subjective ratings in coding (Panel A) and problem solving (Panel B) by objective performance (ratio of tests completed over tests passed at 100 or less), separately by gender.
Figure 9: Subjective Measure by Objective Performance — Communication and Likability

Notes: This figure shows the average subjective ratings in communication (Panel A) and likability (Panel B) by objective performance (ratio of tests completed over tests passed at 100 or less), separately by gender.
Table 1: Descriptive Statistics — August 2016-March 2018

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min.</th>
<th>Max.</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of sessions</td>
<td>25,036</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of interviewees</td>
<td>10,441</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of interviewers</td>
<td>10,232</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of problems</td>
<td>31</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Share of female interviewees</td>
<td>17.82</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Share of female interviewers</td>
<td>17.81</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Panel A: All

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min.</th>
<th>Max.</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Country: USA</td>
<td>0.715</td>
<td>0.452</td>
<td>0</td>
<td>1</td>
<td>49,733</td>
</tr>
<tr>
<td>Interviewee’s degree: computer science</td>
<td>0.669</td>
<td>0.471</td>
<td>0</td>
<td>1</td>
<td>49,731</td>
</tr>
<tr>
<td>Interviewee without working experience</td>
<td>0.273</td>
<td>0.445</td>
<td>0</td>
<td>1</td>
<td>49,732</td>
</tr>
<tr>
<td>Interviewee with a graduate degree</td>
<td>0.451</td>
<td>0.498</td>
<td>0</td>
<td>1</td>
<td>49,733</td>
</tr>
<tr>
<td>Interviewee Preparation Level</td>
<td>2.904</td>
<td>0.798</td>
<td>1</td>
<td>5</td>
<td>49,661</td>
</tr>
</tbody>
</table>

Panel B: Women

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min.</th>
<th>Max.</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Country: USA</td>
<td>0.802</td>
<td>0.399</td>
<td>0</td>
<td>1</td>
<td>8,861</td>
</tr>
<tr>
<td>Interviewee’s degree: computer science</td>
<td>0.650</td>
<td>0.477</td>
<td>0</td>
<td>1</td>
<td>8,861</td>
</tr>
<tr>
<td>Interviewee without working experience</td>
<td>0.304</td>
<td>0.46</td>
<td>0</td>
<td>1</td>
<td>8,861</td>
</tr>
<tr>
<td>Interviewee with a graduate degree</td>
<td>0.516</td>
<td>0.5</td>
<td>0</td>
<td>1</td>
<td>8,861</td>
</tr>
<tr>
<td>Interviewee Preparation Level</td>
<td>2.784</td>
<td>0.792</td>
<td>1</td>
<td>5</td>
<td>8,855</td>
</tr>
</tbody>
</table>

Panel C: Men

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min.</th>
<th>Max.</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Country: USA</td>
<td>0.696</td>
<td>0.46</td>
<td>0</td>
<td>1</td>
<td>40,872</td>
</tr>
<tr>
<td>Interviewee’s degree: computer science</td>
<td>0.673</td>
<td>0.469</td>
<td>0</td>
<td>1</td>
<td>40,870</td>
</tr>
<tr>
<td>Interviewee without working experience</td>
<td>0.266</td>
<td>0.442</td>
<td>0</td>
<td>1</td>
<td>40,871</td>
</tr>
<tr>
<td>Interviewee with a graduate degree</td>
<td>0.437</td>
<td>0.496</td>
<td>0</td>
<td>1</td>
<td>40,872</td>
</tr>
<tr>
<td>Interviewee Preparation Level</td>
<td>2.930</td>
<td>0.797</td>
<td>1</td>
<td>5</td>
<td>40,806</td>
</tr>
</tbody>
</table>
Table 2: Subjective Ratings Pre-Intervention

<table>
<thead>
<tr>
<th>Panel A: All</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min.</th>
<th>Max.</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score in coding</td>
<td>-0.048</td>
<td>1.003</td>
<td>-2.981</td>
<td>1.12</td>
<td>26,306</td>
</tr>
<tr>
<td>Score in problem solving</td>
<td>-0.047</td>
<td>0.984</td>
<td>-2.62</td>
<td>1.264</td>
<td>26,306</td>
</tr>
<tr>
<td>Score in likability</td>
<td>0.075</td>
<td>0.932</td>
<td>-2.738</td>
<td>1.095</td>
<td>26,306</td>
</tr>
<tr>
<td>Score in hireability</td>
<td>0.004</td>
<td>0.998</td>
<td>-3.042</td>
<td>1.046</td>
<td>26,334</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Panel B: Women</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min.</th>
<th>Max.</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score in coding</td>
<td>-0.152</td>
<td>0.995</td>
<td>-2.981</td>
<td>1.12</td>
<td>4,731</td>
</tr>
<tr>
<td>Score in problem solving</td>
<td>-0.15</td>
<td>0.987</td>
<td>-2.62</td>
<td>1.264</td>
<td>4,731</td>
</tr>
<tr>
<td>Score in likability</td>
<td>0.041</td>
<td>0.940</td>
<td>-2.738</td>
<td>1.095</td>
<td>4,731</td>
</tr>
<tr>
<td>Score in hireability</td>
<td>-0.082</td>
<td>1.029</td>
<td>-3.042</td>
<td>1.046</td>
<td>4,736</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Panel C: Men</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min.</th>
<th>Max.</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score in coding</td>
<td>-0.026</td>
<td>1.003</td>
<td>-2.981</td>
<td>1.12</td>
<td>21,575</td>
</tr>
<tr>
<td>Score in problem solving</td>
<td>-0.024</td>
<td>0.982</td>
<td>-2.62</td>
<td>1.264</td>
<td>21,575</td>
</tr>
<tr>
<td>Score in likability</td>
<td>0.083</td>
<td>0.93</td>
<td>-2.738</td>
<td>1.095</td>
<td>21,575</td>
</tr>
<tr>
<td>Score in hireability</td>
<td>0.022</td>
<td>0.991</td>
<td>-3.042</td>
<td>1.046</td>
<td>21,598</td>
</tr>
<tr>
<td>Variables</td>
<td>Control</td>
<td>ITT</td>
<td>Difference</td>
<td>P-value</td>
<td></td>
</tr>
<tr>
<td>---------------------------------</td>
<td>---------</td>
<td>------</td>
<td>------------</td>
<td>---------</td>
<td></td>
</tr>
<tr>
<td>Interviewee female</td>
<td>0.179</td>
<td>0.187</td>
<td>0.007</td>
<td>0.549</td>
<td></td>
</tr>
<tr>
<td>Interviewer female</td>
<td>0.178</td>
<td>0.187</td>
<td>0.008</td>
<td>0.504</td>
<td></td>
</tr>
<tr>
<td>Gender interviewer missing</td>
<td>0.049</td>
<td>0.048</td>
<td>-0.001</td>
<td>0.873</td>
<td></td>
</tr>
<tr>
<td>Country: USA</td>
<td>0.686</td>
<td>0.684</td>
<td>-0.002</td>
<td>0.923</td>
<td></td>
</tr>
<tr>
<td>Interviewee’s deg.: computer science</td>
<td>0.645</td>
<td>0.653</td>
<td>0.008</td>
<td>0.635</td>
<td></td>
</tr>
<tr>
<td>Interviewer’s deg.: computer science</td>
<td>0.643</td>
<td>0.653</td>
<td>0.009</td>
<td>0.578</td>
<td></td>
</tr>
<tr>
<td>Interviewer’s deg.: postgraduate</td>
<td>0.437</td>
<td>0.431</td>
<td>-0.006</td>
<td>0.700</td>
<td></td>
</tr>
<tr>
<td>Interviewee’s deg.: postgraduate</td>
<td>0.441</td>
<td>0.430</td>
<td>-0.012</td>
<td>0.498</td>
<td></td>
</tr>
<tr>
<td>Interviewee’s years of experience</td>
<td>2.943</td>
<td>3.087</td>
<td>0.144</td>
<td>0.224</td>
<td></td>
</tr>
<tr>
<td>Interviewer’s years of experience</td>
<td>2.958</td>
<td>3.090</td>
<td>0.132</td>
<td>0.271</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>1,587</td>
<td>10,004</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Test of joint significance \( F \)-stat: 1.100 \( (p\)-value: 0.377)
Table 4: Results - Full sample

Panel A: All

<table>
<thead>
<tr>
<th>Coding</th>
<th>Problem solving</th>
<th>Likeability</th>
<th>Communication</th>
<th>Hirability</th>
</tr>
</thead>
<tbody>
<tr>
<td>ITT 2SLS</td>
<td>ITT 2SLS</td>
<td>ITT 2SLS</td>
<td>ITT 2SLS</td>
<td>ITT 2SLS</td>
</tr>
<tr>
<td>Treatment</td>
<td>0.147 0.205</td>
<td>0.211 0.295</td>
<td>0.086 0.120</td>
<td>0.198 0.277</td>
</tr>
<tr>
<td>s.d</td>
<td>(0.031) (0.043)</td>
<td>(0.030) (0.041)</td>
<td>(0.033) (0.046)</td>
<td>(0.039) (0.005)</td>
</tr>
<tr>
<td>P-value</td>
<td>0.000 0.000</td>
<td>0.000 0.000</td>
<td>0.012 0.010</td>
<td>0.000 0.000</td>
</tr>
<tr>
<td>N</td>
<td>11,029 11,029</td>
<td>11,029 11,029</td>
<td>11,029 11,029</td>
<td>11,029 11,049</td>
</tr>
<tr>
<td>First stage</td>
<td>0.714</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>s.d</td>
<td>(0.009)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P-value</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>11,591</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F-stat</td>
<td>6084.30</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Panel B: Women

<table>
<thead>
<tr>
<th>Coding</th>
<th>Problem solving</th>
<th>Likeability</th>
<th>Communication</th>
<th>Hirability</th>
</tr>
</thead>
<tbody>
<tr>
<td>ITT 2SLS</td>
<td>ITT 2SLS</td>
<td>ITT 2SLS</td>
<td>ITT 2SLS</td>
<td>ITT 2SLS</td>
</tr>
<tr>
<td>Treatment</td>
<td>0.092 0.135</td>
<td>0.188 0.276</td>
<td>0.054 0.080</td>
<td>0.183 0.269</td>
</tr>
<tr>
<td>s.d</td>
<td>(0.081) (0.114)</td>
<td>(0.073) (0.103)</td>
<td>(0.080) (0.114)</td>
<td>(0.073) (0.104)</td>
</tr>
<tr>
<td>P-value</td>
<td>0.258 0.239</td>
<td>0.012 0.008</td>
<td>0.497 0.482</td>
<td>0.013 0.010</td>
</tr>
<tr>
<td>N</td>
<td>2,049 2,049</td>
<td>2,049 2,049</td>
<td>2,049 2,049</td>
<td>2,049 2,055</td>
</tr>
<tr>
<td>First stage</td>
<td>0.678</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>s.d</td>
<td>(0.016)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P-value</td>
<td>0.002</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>2,151</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F-stat</td>
<td>2069.16</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Panel C: Men

<table>
<thead>
<tr>
<th>Coding</th>
<th>Problem solving</th>
<th>Likeability</th>
<th>Communication</th>
<th>Hirability</th>
</tr>
</thead>
<tbody>
<tr>
<td>ITT 2SLS</td>
<td>ITT 2SLS</td>
<td>ITT 2SLS</td>
<td>ITT 2SLS</td>
<td>ITT 2SLS</td>
</tr>
<tr>
<td>Treatment</td>
<td>0.162 0.225</td>
<td>0.218 0.302</td>
<td>0.093 0.129</td>
<td>0.199 0.276</td>
</tr>
<tr>
<td>s.d</td>
<td>(0.032) (0.045)</td>
<td>(0.033) (0.046)</td>
<td>(0.039) (0.054)</td>
<td>(0.044) (0.061)</td>
</tr>
<tr>
<td>P-value</td>
<td>0.000 0.000</td>
<td>0.000 0.000</td>
<td>0.019 0.016</td>
<td>0.000 0.000</td>
</tr>
<tr>
<td>N</td>
<td>8,980 8,980</td>
<td>8,980 8,980</td>
<td>8,980 8,980</td>
<td>8,980 8,994</td>
</tr>
<tr>
<td>First stage</td>
<td>0.721</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>s.d</td>
<td>(0.016)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P-value</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>9,440</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F-stat</td>
<td>4392.79</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: This table shows results for the ITT and 2SLS models using the whole sample. For each of the five dimensions on which users are rated, the coefficient on treatment in each model is shown in each of the top subpanels, for the whole sample, women and men respectively. The first stages are shown in the lower subpanels. Standard errors are clustered at the date level.
Table 5: Baseline characteristics

<table>
<thead>
<tr>
<th>First Stage Sample mean</th>
<th>Compliers</th>
<th>Never-takers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Treated</td>
<td>Untreated</td>
</tr>
<tr>
<td>Interviewee female</td>
<td>0.678***</td>
<td>0.186</td>
</tr>
<tr>
<td></td>
<td>(0.015)</td>
<td>(0.007)</td>
</tr>
<tr>
<td>Country: USA</td>
<td>0.718***</td>
<td>0.684</td>
</tr>
<tr>
<td></td>
<td>(0.010)</td>
<td>(0.008)</td>
</tr>
<tr>
<td>Interviewee’s deg.: computer science</td>
<td>0.709***</td>
<td>0.652</td>
</tr>
<tr>
<td></td>
<td>(0.011)</td>
<td>(0.008)</td>
</tr>
<tr>
<td>Interviewee’s deg.: postgraduate</td>
<td>0.736***</td>
<td>0.431</td>
</tr>
<tr>
<td></td>
<td>(0.011)</td>
<td>(0.008)</td>
</tr>
<tr>
<td>Interviewee’s years of experience</td>
<td>0.736***</td>
<td>3.067</td>
</tr>
<tr>
<td></td>
<td>(0.021)</td>
<td>(0.045)</td>
</tr>
<tr>
<td>Interviewee Preparation Level (self-declared on 1-5 scale)</td>
<td>0.621***</td>
<td>2.880</td>
</tr>
<tr>
<td></td>
<td>(0.049)</td>
<td>(0.013)</td>
</tr>
</tbody>
</table>

Notes: Column 1 corresponds to the first stage regression for each specific group. Column 2 is the frequency of the group in the estimation sample. Columns 4 and 5 correspond to the estimation of the characteristic in the complier sample, following Abadie (2003) and corresponds to a 2sls regression where the dependent variable corresponds to the endogenous variable multiplied by the indicator of the group.

* p<0.10, ** p<0.05, *** p<0.01
Table 6: Robustness Checks

<table>
<thead>
<tr>
<th>Coding</th>
<th>Problem solving</th>
<th>Likeability</th>
<th>Communication</th>
<th>Hireability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Panel: Baseline</td>
<td>Treatment</td>
<td>0.166***</td>
<td>0.222***</td>
<td>0.099**</td>
</tr>
<tr>
<td></td>
<td>S.E</td>
<td>0.032</td>
<td>0.032</td>
<td>0.039</td>
</tr>
<tr>
<td></td>
<td>Treatment*Woman</td>
<td>-0.099</td>
<td>-0.056</td>
<td>-0.074</td>
</tr>
<tr>
<td></td>
<td>S.E</td>
<td>0.066</td>
<td>0.061</td>
<td>0.084</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>11029</td>
<td>11029</td>
<td>11029</td>
</tr>
<tr>
<td>Panel: with Month FE</td>
<td>Treatment</td>
<td>0.140***</td>
<td>0.212***</td>
<td>0.079**</td>
</tr>
<tr>
<td></td>
<td>S.E</td>
<td>0.029</td>
<td>0.029</td>
<td>0.036</td>
</tr>
<tr>
<td></td>
<td>Treatment*Woman</td>
<td>-0.109*</td>
<td>-0.067</td>
<td>-0.066</td>
</tr>
<tr>
<td></td>
<td>S.E</td>
<td>0.064</td>
<td>0.059</td>
<td>0.082</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>11029</td>
<td>11029</td>
<td>11029</td>
</tr>
<tr>
<td>Panel: with Controls</td>
<td>Treatment</td>
<td>0.168***</td>
<td>0.226***</td>
<td>0.104***</td>
</tr>
<tr>
<td></td>
<td>S.E</td>
<td>0.032</td>
<td>0.032</td>
<td>0.038</td>
</tr>
<tr>
<td></td>
<td>Treatment*Woman</td>
<td>-0.093</td>
<td>-0.061</td>
<td>-0.074</td>
</tr>
<tr>
<td></td>
<td>S.E</td>
<td>0.066</td>
<td>0.060</td>
<td>0.084</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>11029</td>
<td>11029</td>
<td>11029</td>
</tr>
<tr>
<td>Panel: no Date FE</td>
<td>Treatment</td>
<td>0.160***</td>
<td>0.221***</td>
<td>0.100***</td>
</tr>
<tr>
<td></td>
<td>S.E</td>
<td>0.028</td>
<td>0.028</td>
<td>0.033</td>
</tr>
<tr>
<td></td>
<td>Treatment*Woman</td>
<td>-0.106</td>
<td>-0.066</td>
<td>-0.067</td>
</tr>
<tr>
<td></td>
<td>S.E</td>
<td>0.064</td>
<td>0.059</td>
<td>0.082</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>11029</td>
<td>11029</td>
<td>11029</td>
</tr>
<tr>
<td>Panel: Including pre-treatment period</td>
<td>Treatment</td>
<td>0.146***</td>
<td>0.213***</td>
<td>0.082**</td>
</tr>
<tr>
<td></td>
<td>S.E</td>
<td>0.031</td>
<td>0.031</td>
<td>0.034</td>
</tr>
<tr>
<td></td>
<td>Treatment*Woman</td>
<td>0.011</td>
<td>-0.009</td>
<td>0.025</td>
</tr>
<tr>
<td></td>
<td>S.E</td>
<td>0.023</td>
<td>0.024</td>
<td>0.023</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>54077</td>
<td>54077</td>
<td>54077</td>
</tr>
<tr>
<td>Panel: Difference-in-Difference</td>
<td>Treatment</td>
<td>0.131***</td>
<td>0.199***</td>
<td>0.075**</td>
</tr>
<tr>
<td></td>
<td>S.E</td>
<td>0.029</td>
<td>0.029</td>
<td>0.035</td>
</tr>
<tr>
<td></td>
<td>Treatment*Woman</td>
<td>-0.070</td>
<td>-0.008</td>
<td>-0.047</td>
</tr>
<tr>
<td></td>
<td>S.E</td>
<td>0.062</td>
<td>0.056</td>
<td>0.076</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>54077</td>
<td>54077</td>
<td>54077</td>
</tr>
<tr>
<td>Panel: Controlling for Propensity Score Matching</td>
<td>Treatment</td>
<td>0.165***</td>
<td>0.221***</td>
<td>0.099**</td>
</tr>
<tr>
<td></td>
<td>S.E</td>
<td>0.032</td>
<td>0.033</td>
<td>0.039</td>
</tr>
<tr>
<td></td>
<td>Treatment*Woman</td>
<td>-0.099</td>
<td>-0.055</td>
<td>-0.073</td>
</tr>
<tr>
<td></td>
<td>S.E</td>
<td>0.066</td>
<td>0.061</td>
<td>0.084</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>11029</td>
<td>11029</td>
<td>11029</td>
</tr>
</tbody>
</table>

Notes: This table shows results a series of robustness checks. Panel A presents the results of the baseline ITT specification (Treatment) and the interaction with a categorical variable equal to one when the interviewee is a woman. In Panel B we add month-of-interview fixed effects, and date-of-interview fixed effects in Panel C. In Panel D, we control for socio-demographic characteristics. In Panel E we expand our sample to include pre-treatment introduction interviews with month-of-interview fixed effects. In Panel F, we implement a difference-in-differences with month-of-interview fixed effects. Finally, in Panel G, we control for propensity score matching. Standard errors are clustered at the date level.
Table 7: Gender gap in subjective coding ratings, controlling for objective performance

<table>
<thead>
<tr>
<th></th>
<th>Subjective Coding Ratings</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
</tr>
<tr>
<td>Interviewee female</td>
<td>-0.0812***</td>
<td>-0.0638***</td>
<td>-0.0645***</td>
<td>-0.0610***</td>
</tr>
<tr>
<td></td>
<td>(0.0172)</td>
<td>(0.0173)</td>
<td>(0.0173)</td>
<td>(0.0197)</td>
</tr>
<tr>
<td>Objective performance</td>
<td>0.485***</td>
<td>0.456***</td>
<td>0.457***</td>
<td>0.479***</td>
</tr>
<tr>
<td></td>
<td>(0.0141)</td>
<td>(0.0141)</td>
<td>(0.0141)</td>
<td>(0.0171)</td>
</tr>
<tr>
<td>Interviewer female</td>
<td>0.0320*</td>
<td>0.0298</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0165)</td>
<td>(0.0189)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interviewee’s sociodemographic controls</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Interviewer’s sociodemographic controls</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Date FE</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Observations</td>
<td>19,559</td>
<td>19,551</td>
<td>19,551</td>
<td>19,551</td>
</tr>
</tbody>
</table>

Notes: This table shows the estimation of the gender gap in subjective ratings, controlling for objective performance measure (proxied by the ratio of test solved over passed by problem), using a linear regression model in which we progressively add controls. In column 2, we add sociodemographic controls, such as interviewer’s and interviewee’s years of experience, a dummy variable for each level area of education and highest educational level, and self-reported level of preparedness. In column 3 to 5, we control for the gender of the interviewer. In columns 4, we add date-of-interview fixed effects.
<table>
<thead>
<tr>
<th></th>
<th>Subjective Coding Ratings</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
</tr>
<tr>
<td>Interviewee female</td>
<td>-0.081***</td>
<td>-0.081***</td>
<td>-0.084***</td>
<td>-0.0757***</td>
</tr>
<tr>
<td></td>
<td>(0.018)</td>
<td>(0.018)</td>
<td>(0.021)</td>
<td>(0.021)</td>
</tr>
<tr>
<td>Interviewer’s total # of sessions</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interviewer’s # of past sessions</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interviewer’s total # of female interviewees</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Past top female performer</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Objective performance</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Interviewer gender</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Interviewee’s sociodemographic controls</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Interviewer’s sociodemographic controls</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Date FE</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Observations</td>
<td>19,551</td>
<td>19,551</td>
<td>14,677</td>
<td>13,541</td>
</tr>
</tbody>
</table>

Notes: This table shows the estimation of the gender gap in subjective ratings, controlling for objective performance measure (proxied by the ratio of test solved over passed by problem), using a linear regression model in which we progressively add controls. In column 1, we add a control for the interviewer’s total number of sessions, in column 2 we control for the number of previous sessions, in column 3 control for the interviewer’s total number of sessions with a female user, and in column 4 we control for whether the interviewer faced a top female performer during the previous session. All specifications include controls for interviewer’s and interviewee’s years of experience, a dummy variable for each level area of education and highest educational level and for the gender of the interviewer, problem fixed-effects and date-of-interview fixed effects.
Appendix to

Does Better Information Reduce Gender Discrimination in the Technology Industry?

Abdelrahman Amer, Ashley C. Craig and Clémentine Van Effenterre

March 2023

List of Appendices

Appendix A: Institutional details A-2
Appendix B: Additional Results A-5
Appendix C: Follow-up Experiment A-13
Appendix D: Questionnaire A-20
Appendix A  Institutional details

Figure A1: Environment of the platform and treatment

(a) Control  (b) Treatment

Notes: Figure A1(a) presents the website layout for a mock interview on the platform in the control condition. Figure A1(b) represents the treatment condition.

Figure A2: Users across the world

Notes: The map shows the distribution of users across the world.
Figure A3: Users’ level of education

Notes: The figure presents the average level of education of users.

Figure A4: Users’ field of education

Notes: The figure presents the field of education of users.
**Figure A5: Growth of the platform**

Notes: The figure shows the evolution of the number of users on the platform from January 2016 to January 2018.

**Figure A6: Treatment assignment**

Notes: This diagram shows how users are assigned to the treatment or to the control conditions when they enter the platform.
Appendix B  Additional Results

Figure B7: Pre-treatment gender gaps by problem difficulty

Notes: This figure plots gender gaps in subjective ratings for coding and problem solving by problem difficulty in the pre-intervention period. Problem difficulty is computed using the average objective performance of users in the post-intervention period.

Figure B8: Distribution of Objective Performance by Gender

Notes: The figure presents the distribution of the objective performance measure (ratio of test solved / tests taken) by gender.
Figure B9: Ranking of problems by gender

Notes: This figure shows the relative ranking of problems’ difficulty by gender. The ranking is proxied by the average performance of users for each problem. The orange vertical lines show any positive or negative deviation of female users’ ranking compared to male users’ ranking.

Table B1: Problems’ and Evaluators’ Characteristics

<table>
<thead>
<tr>
<th></th>
<th>Problem Difficulty (1)</th>
<th>Precision of Harsh Evaluator (2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interviewee female</td>
<td>-0.003 (-0.008)</td>
<td>0.006 (0.008)</td>
<td>0.005 (0.010)</td>
<td>0.005 (0.010)</td>
</tr>
<tr>
<td>Interviewer Gender</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Date FE</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Problem FE</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>N</td>
<td>26,667</td>
<td>26,667</td>
<td>22,582</td>
<td>19,635</td>
</tr>
</tbody>
</table>

Notes: The regression TBC
Figure B10: Share of male and female users over time

Notes: This figure shows the evolution of the shares of female and male users on the platform before and after the introduction of the device.

Figure B11: Evolution of First-Time Users’ Characteristics

Notes: The figure presents the evolution of first-time users’ characteristics averaged by month around the date of the introduction of the device on the platform.
**Figure B12**: Evolution of First-Time Female Users’ Characteristics

![Graph showing the evolution of first-time female users' characteristics](image)

Notes: The figure presents the evolution of first-time female users’ characteristics averaged by month around the date of the introduction of the device on the platform.

**Figure B13**: Share of High-Performing First-Time Female and Male Users

![Graph showing the share of high-performing first-time female and male users](image)

Notes: The figure presents the evolution of the share of high-performing first-time female and male users by month after the introduction of the device on the platform. High-performing users are defined as those passing all unit tests taken for a given problem.
Table B2: Balancing test by problem difficulty – whole sample

<table>
<thead>
<tr>
<th>Variables</th>
<th>Hard</th>
<th>Easy</th>
<th>Difference</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interviewee female</td>
<td>0.173</td>
<td>0.176</td>
<td>0.003</td>
<td>0.583</td>
</tr>
<tr>
<td>Interviewer female</td>
<td>0.175</td>
<td>0.173</td>
<td>-0.002</td>
<td>0.625</td>
</tr>
<tr>
<td>Gender interviewer missing</td>
<td>0.079</td>
<td>0.073</td>
<td>-0.006</td>
<td>0.057</td>
</tr>
<tr>
<td>Country: USA</td>
<td>0.699</td>
<td>0.702</td>
<td>0.003</td>
<td>0.556</td>
</tr>
<tr>
<td>Interviewee’s deg.: computer science</td>
<td>0.641</td>
<td>0.639</td>
<td>-0.001</td>
<td>0.818</td>
</tr>
<tr>
<td>Interviewer’s deg.: computer science</td>
<td>0.642</td>
<td>0.636</td>
<td>-0.006</td>
<td>0.370</td>
</tr>
<tr>
<td>Interviewer’s deg.: postgraduate</td>
<td>0.477</td>
<td>0.469</td>
<td>-0.007</td>
<td>0.302</td>
</tr>
<tr>
<td>Interviewee’s deg.: postgraduate</td>
<td>0.471</td>
<td>0.471</td>
<td>-0.000</td>
<td>0.978</td>
</tr>
<tr>
<td>Interviewee’s years of experience</td>
<td>3.230</td>
<td>3.286</td>
<td>0.056</td>
<td>0.186</td>
</tr>
<tr>
<td>Interviewer’s years of experience</td>
<td>3.321</td>
<td>3.193</td>
<td>-0.128</td>
<td>0.002</td>
</tr>
</tbody>
</table>

N: 
11,984 | 12,080

Test of joint significance: F-stat: 1.800 (p-value: 0.078)
**Figure B14: Variations across Problems**

(a) Problem Average Difficulty

(b) Precision of the Signal

Notes: This figure shows the distribution of average performance by Problem (Panel A) and the distribution of standard deviation by problem (Panel B) measured by the mean and standard deviation the objective coding performance (ratio of tests completed over tests passed).
Figure B15: Correlation between Men’s and Women’s Treatment Effects on Subjective Rating

Notes: This figure shows the correlation between the treatment effect on women and men by problem’s average difficulty and by problem’s standard deviation of performance. The treatment effect are obtained from the estimation of Equation (5) where the dependent variable is the subjective rating in coding, separately by problem and gender.
Figure B16: Men’s and Women’s Treatment Effects on Subjective Rating by Problem

(a) Problem Average Difficulty

(b) Precision of the Signal

Notes: This figure shows the estimates of Equation (5) where the dependent variable is the subjective rating in coding, separately by problem type and gender.
Appendix C  Follow-up Experiment

C.1 Experimental Design

Recruitment  Our subject population is comprised of recent graduates or students currently enrolled in computer science programs. We recruited evaluators through universities’ graduate (and potentially undergraduate) programs. Our recruitment email discloses that we are studying how evaluators judge the performance of software developers but does not explicitly mention gender.

Sample  To construct the sample of code blocks, we leverage a more recent dataset obtained from the platform we partnered with, spanning observations from January 2018 to May 2022 (see Table ??). Like our previous dataset, this dataset contains the subjective ratings and objective measure of coding quality. From this sample, we use first names to identify gender using predictions from genderize.io. This leaves us with 38,322 session-participant pairs, and 10,380 unique participants. Of these, 18 percent are probabilistically identified as female. A novel feature of our dataset is that we can link this information to the code blocks written by each participant in each session. An example of such a code block is shown in Figure C17 Panel B, together with the question (Panel A) and the unit tests (Panel C). Our final sample is stratified by gender and coding performance.

Randomization  Let $N$ be the number of evaluators and $P$ the number of problems by evaluator. As mentioned before, our sample of code blocks is stratified by gender and performance, such that $\frac{P}{2}$ code blocks are written by women, among which $\frac{P}{4}$ are high-score codes according to the platform objective device. Each evaluator $i$ is assigned a set of $P$ problems in a random order. We use a within-subject design. We define $R_j = 0$ for a blind problem $j$ (if the gender of the coder is not revealed), $R_j = 1$ for a non-blind problem $j$ (if the gender of the coder is revealed). For each evaluator $i$, the gender of the coder will be revealed for half of the problems. To account for potential priming effect, we plan to randomize whether the gender of the coder is revealed in the first or in the second half of the study:

1. For half of evaluators, problems will be blind, then non-blind.
∀ i = 1, ..., \(\frac{N}{2}\) \[
\begin{align*}
&\text{for } j = 1, \ldots, \frac{P}{2}, \quad R_{ij} = 0 \\
&\text{for } j = \frac{P}{2}, \ldots, P, \quad R_{ij} = 1
\end{align*}
\]

2. For the other half, problems will be non-blind, then blind.

∀ i = \(\frac{N}{2}\), ..., N \[
\begin{align*}
&\text{for } j = 1, \ldots, \frac{P}{2}, \quad R_{ij} = 1, \\
&\text{for } j = \frac{P}{2}, \ldots, P, \quad R_{ij} = 0
\end{align*}
\]

**Testing the salience of the main treatment** In the piloting phase of the experiment, we asked a random sample of online participants ("evaluator") on Prolific to predict the gender of a participant ("worker") after evaluating a task they completed, mimicking the lay-out of the first name and avatar of our main experiment. While a non-trivial fraction of "evaluators" didn’t pay attention to the gender of the "workers", neither the evaluators’ characteristics nor the workers’ characteristics (including gender, race, and how racially distinctive the first name) are predictive of the accuracy of the gender prediction. Additionally, we tested whether an AI tool (Chat GPT) was able to predict the gender of the coder of a code when the first name is not displayed, and it was not able to form that prediction.

**Measure of Priors** To measure participants’ priors, we exposed them to three different vignettes before the perform their evaluation tasks. We ask them to predict the potential performance of three different hypothetical coders. We cross-randomize the first name (alternating gender) and the skill level for each vignette. The vignette are constructed as follows:

"52% of the codes you will potentially see resulted in a perfect score and passed all the unit tests. We ask your opinion about the potential performance of different hypothetical coders. If your guess is within 5% of the truth, we will send you an additional reward!"

"[First Name] holds [Skills]. According to you, what is the percent chance that [First Name]'s code passed all the unit tests?"
<table>
<thead>
<tr>
<th>Skills</th>
<th>First names</th>
</tr>
</thead>
<tbody>
<tr>
<td>a M.Sc in computer science and has 2 years of work experience</td>
<td>Katie/Tom</td>
</tr>
<tr>
<td>a Ph.D. in mathematics and has no industry experience</td>
<td>Alexa/Mickael</td>
</tr>
<tr>
<td>a B.Sc. degree in computer science</td>
<td>Corinne/Matt</td>
</tr>
</tbody>
</table>
Figure C17: Example of Code — K-Messed Array Sort

Given an array of integers \( arr \) where each element is at most \( k \) places away from its sorted position, code an efficient function \( \text{sortKmessedArray}() \) that sorts \( arr \). For instance, for an input array of size \( 10 \) and \( k = 2 \), an element belonging to index \( 6 \) in the sorted array will be located at either index \( 4, 5, 6, 7, 8 \) or \( 9 \) in the input array.

Analyze the time and space complexity of your solution.

```
// Example
int arr[] = {5, 4, 3, 1, 0, 3, 0, 0, 10, 9};
k = 2

//Const specials:
- [time limit] 3000ms
- [memory limit] 64MB
- arr.length <= 1000
- arr[i] int, 0 <= arr[i] <= 100

//Answer

(a) Question

```

```javascript
function sortKmessedArray(arr, k) {
    for (var i = 0; i < arr.length; i++) {
        let lowerBound = i - k < 0 ? 0 : i - k;
        let upperBound = i + k + arr.length > arr.length - 1 ? arr.length - 1 : i + k;
        let item = arr[i];
        let index = lowerBound;
        for (var j = lowerBound; j <= upperBound; j++) {
            if (item > arr[j]) {
                index = j;
                arr.splice(i, 1);
                if (index > i) {
                    arr.splice(index, 1, item);
                } else {
                    arr.splice(index + 1, 1, item);
                }
            }
        }
    }
    console.log(arr);
    return arr;
}

sortKmessedArray([1, 4, 5, 2, 3, 7, 8, 6, 18, 91, 2]);
```

(b) Answer

```
// Tests

describe("Solution", function() {
    if("Test #1 for question \"K-Messed Array Sort\") {function() {
        console.error('<START_ERROR>/>');
        const actual = sortKmessedArray([1, 0], 1);
        console.log('<ACTUAL:result>:', actual);
        console.error('<END_ERROR>/>');
        Test.assertEquals(actual, [1]);
    });
    if("Test #2 for question \"K-Messed Array Sort\") {function() {
        const actual = sortKmessedArray([1, 0, 1], 1);
        console.log('<ACTUAL:result>:', actual);
        console.error('<END_ERROR>/>');
        Test.assertEquals(actual, [0, 1, 1]);
    });
    if("Test #3 for question \"K-Messed Array Sort\") {function() {
        console.error('<START_ERROR>/>');
        const actual = sortKmessedArray([1, 0, 3, 2], 1);
        console.log('<ACTUAL:result>:', actual);
        console.error('<END_ERROR>/>');
        Test.assertEquals(actual, [0, 1, 2, 3]);
    });
    if("Test #4 for question \"K-Messed Array Sort\") {function() {
        console.error('<START_ERROR>/>');
        const actual = sortKmessedArray([1, 0, 3, 4, 5, 7, 6, 8], 1);
        console.log('<ACTUAL:result>:', actual);
        console.error('<END_ERROR>/>');
        Test.assertEquals(actual, [0, 1, 2, 3, 4, 5, 6, 7, 8]);
    });
    if("Test #5 for question \"K-Messed Array Sort\") {function() {
        console.error('<START_ERROR>/>');
        const actual = sortKmessedArray([1, 4, 5, 2, 3, 7, 8, 6, 10, 0], 2);
        console.log('<ACTUAL:result>:', actual);
        console.error('<END_ERROR>/>');
        Test.assertEquals(actual, [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]);
    });
    if("Test #6 for question \"K-Messed Array Sort\") {function() {
        console.error('<START_ERROR>/>');
        const actual = sortKmessedArray([6, 1, 4, 11, 2, 0, 3, 7, 10, 5, 0, 9], 0);
        console.log('<ACTUAL:result>:', actual);
        console.error('<END_ERROR>/>');
        Test.assertEquals(actual, [0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11]);
    });

(c) Tests

Notes: This figure presents an example of code excerpt that will be used in the experiment. Panel A displays the question, Panel B the written code block, and Panel C the series of unit tests that generate the objective measure of performance.

A-16
Notes: This figure presents an example of code in the blind and non-blind conditions for both male and female coders.
Table C3: Descriptive Statistics — Follow-up Experiment

<table>
<thead>
<tr>
<th></th>
<th>Raw Data</th>
<th>Clean Data</th>
<th>Experimental Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of session-participant pairs</td>
<td>482,390</td>
<td>178,717</td>
<td>38,322</td>
</tr>
<tr>
<td>Number of unique participants</td>
<td>97,614</td>
<td>30,633</td>
<td>10,380</td>
</tr>
<tr>
<td>Number of unique problems</td>
<td>39</td>
<td>39</td>
<td>38</td>
</tr>
<tr>
<td>Share non-missing unit score</td>
<td>42.24%</td>
<td>56.47%</td>
<td>100%</td>
</tr>
<tr>
<td>Share of Python scripts</td>
<td>29.76%</td>
<td>37.29%</td>
<td>43.10%</td>
</tr>
<tr>
<td>Share of Java scripts</td>
<td>35.14%</td>
<td>34.91%</td>
<td>44.72%</td>
</tr>
<tr>
<td>Share of C++ scripts</td>
<td>16.89%</td>
<td>9.22%</td>
<td>12.16%</td>
</tr>
</tbody>
</table>

Note: the raw data are as received from Platform. The clean data correspond to scripts with non-missing interviewer rating, feedback and question type. The final sample corresponds to scripts with identified gender and race, and non-missing unit-test score. Participants restricted for those in USA only.
Table C4: Descriptive Statistics — Sample Construction — January 2018-May 2022

<table>
<thead>
<tr>
<th></th>
<th>Obs.</th>
<th>Mean</th>
<th>Median</th>
<th>S.D</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Raw Data</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Full score</td>
<td>203,769</td>
<td>0.34</td>
<td>1.00</td>
<td>0.39</td>
</tr>
<tr>
<td>Num code lines</td>
<td>482,390</td>
<td>44.12</td>
<td>40.00</td>
<td>37.45</td>
</tr>
<tr>
<td>Female</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Non-white</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>Clean Data</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Full score</td>
<td>100,933</td>
<td>0.81</td>
<td>1.00</td>
<td>0.40</td>
</tr>
<tr>
<td>Num code lines</td>
<td>178,717</td>
<td>55.25</td>
<td>48.00</td>
<td>31.89</td>
</tr>
<tr>
<td>Female</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Non-white</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>Experimental Sample</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Full score</td>
<td>38,322</td>
<td>0.82</td>
<td>1.00</td>
<td>0.38</td>
</tr>
<tr>
<td>Num code lines</td>
<td>38,322</td>
<td>45.18</td>
<td>44.00</td>
<td>13.55</td>
</tr>
<tr>
<td>Num code lines - male</td>
<td>31,245</td>
<td>45.23</td>
<td>44.00</td>
<td>13.63</td>
</tr>
<tr>
<td>Num code lines - female</td>
<td>7,077</td>
<td>44.97</td>
<td>44.00</td>
<td>13.17</td>
</tr>
<tr>
<td>Female</td>
<td>38,322</td>
<td>0.18</td>
<td>0.00</td>
<td>0.39</td>
</tr>
<tr>
<td>Non-white</td>
<td>38,322</td>
<td>0.61</td>
<td>1.00</td>
<td>0.49</td>
</tr>
</tbody>
</table>

Table C5: Gender gap in subjective coding ratings, controlling for objective performance — Experimental Sample

<table>
<thead>
<tr>
<th></th>
<th>Subjective Coding Ratings</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
</tr>
<tr>
<td><strong>Interviewee female</strong></td>
<td>-0.123***</td>
</tr>
<tr>
<td></td>
<td>(0.0131)</td>
</tr>
<tr>
<td><strong>Objective performance</strong></td>
<td>1.092***</td>
</tr>
<tr>
<td></td>
<td>(0.0216)</td>
</tr>
<tr>
<td><strong>Interviewer’s FE</strong></td>
<td>No</td>
</tr>
<tr>
<td><strong>Problem FE</strong></td>
<td>No</td>
</tr>
<tr>
<td><strong>Observations</strong></td>
<td>38,322</td>
</tr>
</tbody>
</table>

Note: This table shows the estimation of the gender gap in subjective ratings, controlling for objective performance measure (proxied by the ratio of test solved over passed by problem), using a linear regression model in which we progressively add controls. We progressively add interviewer and problem fixed effects.
Appendix D: Questionnaire

Informed Consent

Overview

You are being asked to take part in a research study being done by a group of researchers from the University of Michigan and the University of Toronto. This is a survey for academic research in social sciences. Your participation is invaluable for our research. If you choose to participate and to complete the survey, you will be financially compensated with a minimum of $50. As a participant, you will be asked to evaluate pieces of code written by others, and answer a short follow-up questionnaire. We expect that participation will take around 60 minutes. In each part, you will receive clear instructions and will be told how your decisions in that part will influence your earnings in the study. You will also have the opportunity to learn about your performance as evaluator.

Non-Deception Statement

This study does not deceive you by providing misleading or incorrect information. All our communications are truthful, but we may not always reveal all information. Specifically, there are different versions of this study. While you will be fully informed about the version of this study that you have been randomly assigned to, you will not be informed about different versions of this study that other participants are in.

Voluntary Participation, Privacy, and Point of Contact

Your participation is completely voluntary. You can agree to take part and later change your mind. Your decision will not be held against you. Note that the data you provide in this study will be anonymized prior to analysis. Your information will be kept entirely confidential and accessed only by the research team, and only as necessary to conduct the research. In the future, this non-identifiable data may be shared with other researchers or published. All information identifying you as a study participant will be destroyed upon the conclusion of the study. However, the anonymized information you provide may be maintained indefinitely.
The principal investigator of this study is Ashley C. Craig from University of Michigan. If you have any questions, concerns, or complaints, or think this research hurt you, talk to the research team at ash@ashleycraig.com. If you have questions about your rights as participants, you can contact the Research Oversight and Compliance Office — Human Research Ethics Program at ethics.review@utoronto.ca or 416-946-3273. You can also contact the University of Michigan IRB (Health Sciences and Behavioral Sciences) at 734-936-0933 or irbhsbs@umich.edu, quoting eResearch #HUM00204184.

The research study you are participating in may be reviewed for quality assurance to make sure that the required laws and guidelines are followed. If chosen, (a) representative(s) of the Human Research Ethics Program (HREP) may access study-related data and/or consent materials as part of the review. All information accessed by the HREP will be upheld to the same level of confidentiality that has been stated by the research team. If you would like a summary of the results of this research (once the study has been completed), please email ash@ashleycraig.com.

Compensation

You will receive $10 if you complete the survey and an additional $10 for each code segment you evaluate. Additionally, we will ask you to make a series of predictions. You will have the opportunity to gain $2 for each accurate prediction. Your total earnings will be distributed within one week after the completion of the survey. If you are interested, you can receive individualized feedback about the quality of your performance as an evaluator.

Based on their performance, the best ten evaluators win a $500 prize. The three best evaluators will also be invited to the Creative Destruction Lab 2023 Super Session in Toronto, which brings together world-class entrepreneurs, investors and scientists with high-potential startup founders. Organized in June 2023, the CDL Super Session days will give you with meaningful networking opportunities and exposure to key players in the industry. If there are ties in evaluation performance, the recipients of the prize and these invitations will be chosen randomly from among the set of evaluators with equal best accuracy scores. You may print a copy of this information for your records.
Yes, I would like to voluntarily participate in this experiment.

I am interested in receiving individualized feedback on my performance as an evaluator.

- Yes
- No

For the purposes of payment and the $500 cash prize, and to be considered for an invitation to the Creative Destruction Lab, please type your email below. We will not use your email for any purposes other than the provision of these rewards.

[ Type here ]

Please make sure you are willing and ready to sit through this study uninterrupted and undistractedly before starting it. We ask you to please focus on the tasks of this study and thank you for your cooperation.

**General Roadmap**

This study consists of 4 evaluation tasks, followed by a few questions. The evaluation parts will ask you to give a score from 1 to 4 for scripts, both of which are solutions to a given coding question. The coding question will be outlined before the script.

**Attention Checks**

Note that this experiment contains attention checks. These questions are there to ensure you are paying attention as you take this survey. The answers to those attention check questions will not be ambiguous, will not be a trick question, and will not be timed. If you answer an attention check incorrectly or not within the provided time, you may be dismissed without pay.

Here is your first attention check. In the space below, please spell the word "human" backwards. Please use all lowercase letters and insert no space between the letters.

[ Type here ]
1. What best describes your present situation regarding your education?
   - I am currently a student
   - I have completed at least one degree
   - I was previously enrolled in a degree program but did not complete it

2. What is your highest level of education (including enrolled)?
   - High School diploma or GED
   - Some college, but no degree
   - Associates or technical degree
   - Bachelor’s degree
   - MA, MSc or MEng
   - PhD
   - Prefer not to say

3. What is or are the area(s) of your highest degree? (multiple answers are allowed)
   - Computer Science
   - Computer Engineering
   - Mathematics
   - Information Systems / M.I.S.
   - Statistics
   - Other Exact Sciences Degree (e.g. physics, chemistry, astronomy)
   - Other Technology Related Degree
   - None
   - Other

4. What is the institution where you received or will receive your highest degree?
   [ Drop down menu ]

5. How would you describe your knowledge of these programming languages?
   Basic-Intermediate-Advanced
   - Python
   - Java
   - C++
6. During this study, you will be asked to evaluate a series of human written code blocks. Please select the coding language you are most proficient in.

- Python
- C++
- Java

Before you start, we want to ask you a series of quick questions. The code excerpts were automatically subjected to a series of unit tests. These determined whether the code ran, and produced correct answers in pre-defined test cases.

Overall, 52% of the code blocks you will potentially see resulted in a perfect score and passed all the unit tests. We ask your opinion about the potential performance of different hypothetical coders. If your guess is within 5% of the truth for coders like those described, you will receive an additional reward!

- Katie/Tom holds a M.Sc in computer science and has 2 years of work experience. According to you, what is the percent chance that Katie’s code passed all the unit tests?
- Alexa/Michael holds a Ph.D. in mathematics and has no industry experience. According to you, what is the percent chance that Alexa’s code passed all the unit tests?
- Corinne/Matt holds a B.Sc. degree in computer science. According to you, what is the percent chance that Matt’s code passed all the unit tests?

BEGINNING OF TASK

We are now going to ask you to evaluate a series of codes. These codes were written by actual software developers. We will provide you with the initial question and their written answers.

For each piece of code, we ask you to give your personal opinion about the quality of code, by providing a rating between 1 (lowest) and 4 (highest). At the end of all code evaluation, we will ask you to explain how you decided on your rating. You will gain a $10 additional bonus for each code you evaluate.
Additionally, we will ask you to make a series of predictions. You will have the opportunity to gain $2 for each accurate prediction.

**Code Block 1**

1. How would you rate the quality of the code (1 lowest, 4 highest)?
   - 1 (lowest)
   - 2
   - 3
   - 4 (highest)

2. Can you let us know why you gave this score to the code?

   Text Box

3. A series of unit tests were used to evaluate this code. How many out of 10 unit tests do you think were passed? If your guess is within 5 percentage points of the truth, you will gain $2 and will increase your chances of participating to the Creative Destruction Lab Meeting and winning one of the $500 prizes.
   - Drop Down menu

4. How confident are you about this prediction?
   - Not confident at all
   - Not confident
   - Somewhat confident
   - Confident
   - Very confident

5. Another human evaluator assessed whether this coder passed or failed based on this coding performance and other factors. We ask you to guess whether that evaluator decided that this coder passed or failed. Please note that 85% of all coders pass. If you guess correctly, you will gain $2 USD, and will increase your chances of participating in the Creative Destruction Lab meeting and winning one of the $500 USD prizes. Based on this code that they wrote, do you think the code passed or failed?
   - Failed
6. How confident are you about this prediction?

- Not confident at all
- Not confident
- Somewhat confident
- Confident
- Very confident

According to you, what is the percent chance that the candidate was later invited for an interview for a role involving coding?

- Cursor between 0 and 100

People often consult internet sites to learn about employment opportunities in tech. We want to know which sites you use. We also want to know if you are paying attention, so please select Glassdoor and Crunchbase regardless of which sites you use. When looking for employment opportunities, which is the one website you would visit first? (Please only choose one).

- LinkedIn
- Hired
- Glassdoor
- Crunchbase
- ZipRecruiter
- TripleByte
- Underdog
- Angel

**Code 2 to 4 — Repeat**

*FOR PILOT ONLY* What is your prediction of the percent chance that the last candidate was a woman?

- Cursor between 0 and 100
Follow-up questions

1. In which country do you currently reside?
   - Canada
   - USA
   - Other (choose)

2. How do you describe yourself?
   - Male
   - Female
   - Non-Binary / third gender
   - Prefer to self-describe: (type)
   - Prefer not to say

3. What is your year of birth?
   - Drop down menu

4. What best describes your employment status of the last three months?
   - Working full-time
   - Working part-time
   - Unemployed and looking for work
   - A homemaker or stay-at-home parent
   - Student
   - Retired
   - Other

5. How many year of working experience do you have?
   - Drop down menu

6. On a scale of 1-4 how prepared do you believe you are able to evaluate others’ code?
   - 1
   - 2
   - 3
   - 4
1. In the box below, explain how you made your decisions today. Please answer in one or more full sentences.

   • Text Box

2. If you had to guess, what do you think was this study about? Please answer in one or more full sentences.

   • Text Box

3. Do you have any comments or feedback related to this study? (optional)

   • Text Box

4. Was there anything confusing about this study? (optional)

   • Text Box

Congratulations, you completed the main portion of the experiment! Once you have completed the questionnaire, you will reach the end of the experiment and learn about your total payment.

   END of Questionnaire