Decoding Gender Bias in Interviews*

Abdelrahman Amer, Ashley C. Craig and Clémentine Van Effenterre

September 2025

Abstract

Performance evaluations in interviews are central to employment decisions. We combine two field experiments, administrative data and video analysis to study the sources of gender gaps in interview evaluations. Leveraging 60,000 mock interviews on a platform for software engineers, we find that code quality ratings are 12 percent of a standard deviation lower for women. This gap persists after controlling for an objective measure of code quality. Providing evaluators with automated performance measures does not reduce gender gaps. Comparing blind to non-blind evaluations without live interaction reveals no gender gap in either case. In contrast, gaps widen with longer personal interaction and are larger among evaluators from regions with stronger implicit gender bias. Video analysis shows that women apologize more; and interviewers are more condescending and harsher with them. Both correlate with lower ratings. Our findings highlight how interpersonal dynamics can introduce bias into evaluations that otherwise rely on objective metrics.

JEL codes: C93, D83, J16, J71, M51
Keywords: *Discrimination*; *Gender*; *Coding*; *Experiment*; *Information*

*__Amer__: University of Toronto, 150 St. George Street, Toronto ON M5S 3G7, Canada (e-mail: abdelrahman.amer@mail.utoronto.ca). **Craig**: Australian National University, Research School of Economics, Room 2094, LF Crisp Building, 25a Kingsley St, Acton ACT 2601 (e-mail: ashley.craig@anu.edu.au). **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, Rahul Deb, Stefano DellaVigna, Nicole Fortin, Dylan Glover, Maria Guadalupe, Sara Heller, Peter Hull, Pat Kline, Kory Kroft, Corinne Low, Michelle Lowry, Marion Monnet, Peter Morrow, Ricardo Perez-Truglia, Roland Rathelot, Alexandra Roulet, Nina Roussille, Heather Sarsons, Melanie Wasserman and Basit Zafar. We also thank seminar participants at the NBER Entrepreneurship SI and the NBER Gender in the Economy workshop, Ridge WELAC, SOLE, EALE, SEA, and seminar participants at the University of Chicago, Berkeley, UCLA, USC, Sciences Po, CREST, INSEAD, SSE, Bocconi, AMSE, the Queen's & Toronto Workshop, LAGV, Monash, PSE, University of Hawai and ANU. We are grateful to the Pivotal fund, the NBER Digitization Program and the Russell Sage Foundation for financial support. Matthew Jarvis-Cross and Sabrina Wang provided outstanding research assistance. This project received IRB approval at the University of Michigan, the University of Toronto, and the Australian National University. The second experiment was pre-registered on December 14, 2022, ID: AEARCTR-0009816, a pre-analysis plan was uploaded on the AEA RCT Registry website on January 3, 2023, and updated on February 17, 2023. A previous version of this paper was entitled "Does Better Information Reduce Gender Discrimination in the Technology Industry?".

Introduction

Nearly all recruitment processes involve interviews (Levashina et al., 2014). Recruiters also use other assessment tools such as aptitude tests, task simulations, and resume screening, but final hiring decisions typically hinge on personal interactions. Despite extensive research on discrimination as a barrier preventing women from entering high-paying occupations (Bertrand and Duflo, 2017), there is limited evidence on whether similarly situated men and women are assessed differently during interviews. Correspondence studies usually focus on resume screening, and recent evidence of gender bias is rather mixed at this pre-interview stage of the recruitment process (Bertrand and Duflo, 2017; Kline et al., 2022, 2023).

Studying gender bias in interview assessments is inherently challenging. Data are rarely available, and typically come from a single firm if they are. Comparisons across firms fail to account for context-specific interview practices, such as question types and difficulty. Even then, candidate performance will itself vary with the evaluation context, confounding performance with differences in assessment. In addition, there is usually no objective measure of applicant performance against which to benchmark decisions. This is especially true in high-skilled labor markets.

This paper examines gender bias in performance evaluations during interviews, while addressing some of these empirical challenges. We focus on the technology industry, where women are chronically underrepresented (Ashcraft et al., 2016; Roussille, 2024; Cullen et al., 2025). We combine administrative data, field experiments, and video analysis, to highlight the crucial role of live interaction in triggering gender bias in interview evaluations.

We focus on a stable and structured format of performance evaluation—the coding interview—largely used in the recruitment of computer programmers (Behroozi et al., 2020). During these interviews, candidates are asked to solve programming problems in real time, typically in front of an interviewer, as a way to assess technical ability. Examples of firms using these interviews include Amazon, Google, Microsoft, Apple, and Facebook (Laakmann et al., 2024), which by themselves account for 12 percent of software engineering jobs in the United States.[1]

We first use observational data from 60,000 mock interviews on an online peer-to-peer platform based in the United States. The platform offers job applicants the

[1] We calculate this share using data from LinkedIn profiles from 2016 to 2022 (see Section 1.7).

opportunity to practice for technical interviews. In each session, they solve computer programming challenges. Mirroring real interviews, the evaluator interacts with the coder via video, and assesses their performance afterwards.

Female coders receive lower coding and problem solving ratings than men on the platform. These gender gaps correspond to around 12 percent of a standard deviation. They are largely independent of the gender of the interviewer or the type of problem; and remain when we control for interviewees' and interviewers' levels of education, experience, and self-reported preparation. They also persist when we condition on an objective measure of code quality.

We develop a model of discrimination in the spirit of Lundberg and Startz (1983) to help understand these gaps and guide our analysis. We derive testable predictions for four potential mechanisms that could underpin observed gender gaps, and evaluate each hypothesis using two field experiments. First, interviewers may statistically discriminate against women if they believe them to be worse coders than men. Second, there may be differences in skills between men and women. Third, interviewers may engage in taste-based discrimination against women. Finally, stereotypes may be activated when evaluators and coders interact, as would be predicted if discrimination is driven by unconscious or "implicit" bias.

Our first experiment asks whether interviewers incorrectly believe that women write worse code, and statistically discriminate against them based on this false belief. We study the randomized roll-out of the objective code quality measure, which was made available to pairs of participants before ratings were chosen. If voluntarily activated, these "unit tests" assessed whether the code produced correct answers to test cases. The availability of the tests increased ratings across the board without reducing the gender gap. This result allows us to reject the hypothesis that the gender gap in performance evaluations is driven by incorrect beliefs.

We next test our remaining predictions for gender gaps on the platform. First, we examine whether the gaps are driven by differences in code quality or coding styles that were not captured by the unit tests. Second, we test for taste-based discrimination. Third, we investigate whether stereotypes are activated during personal interactions, when gender-specific mannerisms and behavior may trigger unconscious biases.

To evaluate each of these hypotheses, we ran a second experiment in which a stratified random sample of code originally written on the platform was reevaluated by computer science students. Video interaction was not included, which let evaluators

focus on evaluation of the code itself, and allowed us to vary whether gender was observed. The evaluation setting otherwise mirrored the platform. We randomized whether the coder's gender was revealed by their first name (the "non-blind" condition), or only initials were shown so that gender was masked (the "blind" condition). An important and novel feature of our experiment is that precisely the same code blocks from the platform are evaluated in all evaluation contexts. This allows us to rule out differences in performance across conditions due to phenomena such as stereotype threat (Spencer et al., 2016).

First, we compare evaluations of code written by each gender in the "blind" condition. We find no gender gap in the gender-blind evaluations, despite there being a gender gap when the same set of code blocks were evaluated on the platform. This suggests that differences in code quality—or stylistic differences that are penalized for women—do not explain the gender gap on the platform. It also implies that the gap cannot easily be explained by rational statistical discrimination, because this would rely on the existence of a true gender gap in code quality.

We next test for taste-based discrimination (Becker, 1957) in the sense of a fixed penalty for women that is triggered by observing the gender of the coder. In the spirit of seminal work by Goldin and Rouse (2000), we do this by comparing "blind" to "non-blind" evaluations. Because treatment was randomized, and the set of scripts evaluated is precisely the same in each treatment, we can identify evaluator bias without confounding differences. We find no evidence that women are treated differently when gender is made visible and salient by the revelation of their first names.

Our explanation is that stereotypes come into play during personal interaction, in a manner consistent with unconscious bias. Further analyses support this hypothesis. First, the gender gaps in ratings on the platform are twice as large among evaluators who graduated from an institution in geographic areas with more prejudice towards women in science, as measured by Implicit Association Tests (IAT). While IAT scores have been shown to predict bias in settings with sustained interaction (Carlana, 2019), we provide new evidence that awareness of a coder's gender via their first name is not enough to trigger bias, and that bias is amplified by longer interactions.[2] Specifically, a 15 minute increase in the duration of the overall session leads a widening of

[2]On the role of unconscious mental associations and contextual factors in the formation of discriminatory behaviors, see also Bertrand et al. (2005); Reuben et al. (2014); Bordalo et al. (2016); Hangartner et al. (2021); Barron et al. (2022); Cunningham and de Quidt (2022); Kessler et al. (2022); Bellemare et al. (2023); Handlan and Sheng (2023).

the gender gap by 2.6 percent of a standard deviation, controlling for the candidate's own coding duration and their objective performance. We are also able to rule out competing explanations. In particular, we show that there are no gender differences in coding duration or verbal performance which are not reflected in the written code but which nonetheless enter the ratings.

Finally, we look at the content of interactions using video analysis, and isolate behaviors associated with differential treatment. Since our partner platform did not store video recordings, we instead leverage data from a similar platform that has made a subset of mock interview videos publicly available on YouTube. Job-seekers are paired with a professional with experience interviewing candidates for top technology companies. The videos are anonymized and include candidates' voices but not their faces. Using a large language model (LLM), we find differences in verbal and non-verbal aspects of the interaction when the candidate is a woman. Female candidates are more likely to use apologetic language, which is associated with lower subjective coding ratings. They also use rising intonation ("upticks"), a pattern sometimes interpreted as signaling lack of confidence, particularly when produced by female voices (Levon and Ye, 2020; Divakaruni et al., 2023). On the other side, interviewers are significantly more likely to use a tone perceived by the LLM as condescending, harsh, and impatient when the candidate is a woman. They are also less likely to explain, to actively listen, and to build effective rapport. All these behaviors are negatively correlated with final subjective ratings.

This paper builds on the extensive literature on the role of discrimination in hiring (Becker, 1957; Phelps, 1972; Arrow, 1973; Coate and Loury, 1993; Bertrand and Duflo, 2017; Craig and Fryer, 2019). Measuring discrimination requires the researcher to compare decisions for individuals of different groups who perform objectively just as well. For example, correspondence studies are often used to measure bias because they can precisely vary perceived group membership of candidates while holding fixed job-relevant characteristics (Neumark et al., 1996; Bertrand and Mullainathan, 2004; List, 2004; Neumark, 2012; Kroft et al., 2013; Farber et al., 2016; Kline et al., 2022, 2023). However, relatively little is known about bias arising in interview stages that occur after resume screening, largely due to the scarcity of data and difficulty in formulating convincing empirical designs.[3] The critical role played by interpersonal interactions

[3]Radbruch and Schiprowski (2025) use data from admission process of a German study grant and from a consulting firm to document that candidate assessments are negatively influenced by the quality

between the evaluator and candidate in triggering bias in interviews further helps explain why recent studies have found on average little discrimination against female names when personal interactions were absent.

Blind and non-blind evaluations of candidates in more complex recruitment settings have also been used to document the existence of gender discrimination in hiring, following seminal work of Goldin and Rouse (2000). Mocanu (2023) shows that changes in screening tools have important impacts for both employer hiring decisions and job seekers' applications. Studying recruitment in the Brazilian public sector, she provides evidence that women's relative evaluation scores and the female share of new hires increased after impartial recruitment practices were mandated. Our focus is different: We study whether integrating subjective elements into the technical portion of the interview biases evaluations of technical performance. A key advantage of our setting is that we can hold precisely constant the applicant pool, task, and performance of each candidate, while varying the evaluation environment.

A challenge in identifying the sources of biases is that this generally requires that the researcher measure performance, beliefs, or observe changes in decisions as more information becomes available. In a rare example of this approach, Bohren et al. (2019) distinguish taste-based, rational statistical and non-rational statistical discrimination on a Q&A forum by studying how bias changes as prior evaluations become visible. Bohren et al. (2023) implement a similar approach in an online hiring experiment, but directly provide performance information.[4] Our paper builds on this idea. A key advantage of our experiments is that we control the information seen by interviewers—specifically, whether or not participants interact live, and whether gender is revealed via the first name of the coder—all while holding constant the performance of a fixed set of candidates on real coding tasks in a natural labor market setting.

Finally, our paper relates to the literature on the role of personal interactions on career outcomes. Recent studies have shown how face-to-face interactions are an important feature of workplaces, as physical proximity affects young workers' ability to build skills (Emanuel et al., 2023), and social interactions impact networking for career advancement, which can contribute to the gender wage gap (Cullen and Perez-Truglia, 2023). However, personal interactions can also trigger workplace hostility (Collis and

of the previous candidate in the interview sequence. Shukla (2024) shows that caste discrimination in India arises only when caste is revealed during personal interviews.

[4]In the technology sector, Feld et al. (2022) and Avery et al. (2023) show that providing recruiters with more information can reduce gender gaps in settings without live interaction.

Van Effenterre, 2025). This might be particularly salient during the interview process and have consequences for disparities in hiring outcomes. Dupas et al. (2021) show that women are interrupted more than men in economics seminars. Using a machine-learning algorithm, they show that these interruptions tend to have a negative tone about half the time. Our data and empirical designs provide an unusual opportunity to look at mutual interaction during interviews, and to assess how interaction affects technical assessments in a setting where we can measure performance and there there is no scope for it change endogenously. We contribute to this literature by highlighting the role of interviews as environments in which behavioral cues can surface, which bias evaluations of women's technical ability.

The remainder of the paper is structured as follows. We describe the institutional context and administrative data in Section 1. The model is presented in Section 2. The two experiments are presented in Sections 3 and 4. We more closely evaluate the role of personal interaction in Section 5, and conclude in Section 6.

1 Live Coding Interviews

Technology companies conduct live coding interviews to screen job applicants (Laakmann et al., 2024). These interview questions are to a large degree standardized and aim to test applicants' abilities to understand and apply basic coding concepts. The prevalence of such interviews has led to the proliferation of test preparation platforms such as Coderbyte, Exponent, HackerRank, and Pramp. Similar to test preparation services for the SAT, these companies offer a collection of coding interviews to prepare candidates during the screening process. Our data comes from one of several platforms that have been developed for this purpose.

We use administrative data from the platform for both our experiments. The data allow us to observe a variety of metrics regarding coders' performance and evaluations. What distinguishes our data from other peer evaluation datasets is the ability to link coding evaluations back to the code that was evaluated.

1.1 Interactions on the Platform

A user's experience on the platform begins when they sign up and provide information about their background and experience, including their proficiency with available programming languages. They then schedule an interview during one of many fixed time slots, with the platform suggesting slots which already have users with similar

profiles. When the time arrives, users within the time slot are matched.[5]

The paired users interview each other in turn. Depending on the language, self-reported ability and experience of the users, one of the coding problems is assigned. The interviewee solves the coding problem in an online text editor that both sides see while the users communicate via live video chat (see Figure A1). Once the interview finishes, the interviewer and interviewee swap roles. At the end of their interaction, each user rates the other on their coding quality, communication, hireability, likability, and problem solving.[6] The platform therefore provides an environment where realistic time-constrained tasks are performed and evaluated. This allows the study of gender gaps in performance evaluations in a high-skilled labor market setting where personal interactions can be of high importance. In fact, users' online reviews underscore the importance of such interactions. For example, 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."*

1.2 Description of The Platform Data

Figure A2 presents a detailed timeline of data coverage. Our first experiment (Section 3) occurred during the period of covered by the first part of our dataset, which contains 60,513 interviews from December 18, 2015 to April 18, 2018. Candidates could participate in multiple interviews. Each time, they are paired with a different counterpart. During this period, users had participated in 12 sessions so far on average.

Descriptive statistics for the population of users are shown in Table A1. Participants are high-skilled, and the vast majority graduated in STEM fields. Almost 45 percent had Master's degrees, and nearly all others had a Bachelor's degree. Two thirds of users had computer science degrees, with most of the rest spread across engineering, mathematics, statistics and the hard sciences. Sixteen percent of users were female.

[5] Users are paired based on their similarity scores using Edmunds' Blossom algorithm, which chooses a matching that maximizes the total of similarity scores of paired users.

[6] While we do not know the order of interviews, evaluations are submitted after the entire session ends, which rules out the possibility for anchoring and mitigates the risk of retaliation.

Consistent with evidence from Murciano-Goroff (2018), we find that women declare lower levels of preparation on average.

Our second experiment (Section 4) uses platform data from a more recent period, from April 2018 to May 2022. Crucially, this more recent dataset contains the full code script written by interviewees on the platform. This allows us to provide real, user-written code for evaluation in Experiment II.

1.3 Description of the Revelio Data

We linked the interview data to labor market data from Revelio Labs. This includes data from hundreds of millions of LinkedIn profiles, combined with other sources. For close to the universe of computer science (CS) graduates in the US labor market, we observe job titles, employers, and salary estimates.[7] We match platform participants with a Bachelor's or Master's degree to individuals in the Revelio data who attained a CS-related degree from a US institution. Matching is based on exact first and last name, and degree type. The final sample consists of 5,126 matched CS graduates from 2016 to 2023. The average starting salary of this sample is $81,000, which compares to data from Glassdoor.[8]

1.4 Gender Gaps in Evaluations of Code Quality

Figure 1 shows the gender gaps in evaluations on the platform in the pre-intervention period. Women received 12 percent of a standard deviation lower ratings for code quality and problem solving on average, with no difference in scores for communication. The gender gaps remain largely unchanged when we control for the interviewee's and interviewer's level of education, years of experience and self-declared preparation level (see Table A2). They do not vary with the gender of the interviewer on average, nor do they vary substantially by problem difficulty (column 4). They also persist when we add date fixed effects to take into account any changes in composition as the platform grew (column 5).

1.5 External Generalizability

The Revelio dataset enables us to evaluate how representative platform users are of the broader population. In Table A3, we compare platform users to graduates of com-

[7] One concern is that there may be some sample selection. However, we have reason to believe that coverage is high for CS graduates in the United States. See Appendix B for further discussion.

[8] Computer science graduates sort into various occupations, but according to the Bureau of Labor Statistics, they primarily become software developers. Data from GlassDoor shows that the average entry level salary for Software Developers is around $85,000 in 2023.

puter science from 2016 to 2017 in the Revelio LinkedIn database. Focusing on the period prior to the first intervention in July 2017, platform users are quite similar to the Revelio sample in terms of gender composition, but are slightly more educated on average in the sense that the share of individuals with a Master degree is higher on the platform. In the post-intervention years for which we have access to race, our platform sample is also more likely to be non-white than the Revelio sample, but the two datasets remain comparable in their shares of female users. As a test of external generalizability, we replicate our analysis of gender gaps in coding evaluations after reweighting to ensure that the sample more closely matches the universe of LinkedIn profiles. The results are very similar (see Table A4).

1.6 Gender Gaps Controlling for Objective Code Quality

During the period covered by our data, the platform introduced a new diagnostic tool to verify the quality of code written during the interview. This tool provided automated ("unit") tests which assessed whether the code produced the correct answers for test cases. In Section 3, we use the randomized roll-out of this tool to test whether the gender gap in code quality ratings is driven by incorrect beliefs that women are less competent coders than men.

Once introduced, the automated tool allows us to assess whether the gender gaps in subjective ratings are explained by gender differences in performance as measured by those tests. They do not fully account for the gap. On average, men pass marginally more tests than women, resulting in a slightly higher share of tests passed (Figure A3). These modest performance differences reflect women's slight underrepresentation at the top of the distribution, as shown in Figure 2. However, we find that there is still a substantial gender gap in subjective coding and problem solving assessments when we control for objective code quality. Table A5 shows that the raw gender gap in subjective coding rating is 11 percent of a standard deviation after July 2017 (column 1), and decreases only slightly—to 8 percent of a standard deviation—once we control for objective code quality (column 2).

Because performance on the unit tests is bimodal, we split the sample into two groups: users who passed all unit tests, and those who did not. For each of the two levels of performance, Figure 3 plots average code quality ratings by objective performance by gender. Large gender gaps remain, conditional on objective performance. Although the gender gap in subjective coding ratings is halved for users with the high-

est objective performance, women receive lower subjective coding ratings than men who perform equally well by this measure.

1.7 Verifying the Value of the Unit Tests: Labor Market Outcomes

Higher scores on these unit tests are strongly associated with future labor market performance. To establish this, we linked the interview data to labor market data from Revelio Labs. We use Mincer-type wage regressions of log earnings on individuals' unit test scores, and their characteristics such as gender, race, the highest degree obtained, institution-of-highest-degree, year-of-graduation, and location. Results are presented in Table 2.[9] Going from the 25th to the 75th percentile of unit test scores is associated with a wage increase of 4.5 percent. This compares to a 6 percent residual gender gap in the first salaries of computer science graduates in the Revelio data.[10]

2 A Guiding Model of Discrimination

To assess whether gender gaps we see on the platform reflect statistical discrimination, taste-based bias, unconscious bias, or a mix, we use a model encompassing all mechanisms and compare evaluations across settings to distinguish them systematically.

2.1 Model Setup

The role of an interviewer is to evaluate the ability of job candidate i, who is of gender $g \in \{m, f\}$. The candidate's true ability, y_i, is unobservable. However, the interviewer sees a noisy but informative signal of it, θ_i. In the context of these coding interviews, ability likely encompasses aspects captured by the subjective ratings for problem solving, coding and communication, but potentially also other dimensions of ability. We focus initially on coding ability, as measured by the code quality rating.

Based on this signal, the evaluator forms a belief, $b_i = E(y_i \mid \theta_i, g)$. Finally, the evaluator reports a code quality rating based on this belief. Specifically, ratings are an increasing, monotonic function of the belief, b_i: $r_i = R(b_i, g \mid e)$. It may also depend on the evaluation environment. Specifically, we consider blind ratings ($e = b$), non-blind ratings ($e = nb$) and non-blind settings in which the evaluator and candidate interact live ($e = live$). For simplicity, we assume below that R is linear, but we note that ratings are discretized in practice.

[9] Full details of all aspects of this analysis are available in Appendix B.

[10] This is conservative: Salaries are imputed from job roles, so miss within-role pay variation. We also note that the gender pay gap reflects both supply and demand factors, such as gender differences in preferences for job amenities, job search, earning expectations, negotiation or discrimination.

2.2 Statistical Discrimination

In the spirit of Lundberg and Startz (1983), consider a simple benchmark in which the interviewer can observe the gender of each candidate. Whether they interact live is held constant ($e = nb$ or $e = live$). The interviewer believes the performance of candidates of gender $g \in \{m, f\}$ is normally distributed in the population, with mean μ_g and variance σ_g^2.

$$y_i \sim \mathcal{N}\left(\mu_g, \sigma_g^2\right) \tag{1}$$

The evaluator may believe (correctly or incorrectly) that the mean, μ_g, and standard deviation, σ_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 ε_i is normally distributed with mean zero and variance σ_ε^2, and is independent of both y_i and g. The unconditional distribution of θ_i is as follows.

$$\theta_i \sim \mathcal{N}\left(y_i, \sigma_g^2 + \sigma_\varepsilon^2\right) \tag{2}$$

Rational inference implies that the interviewer combines her belief about the population with the information in the signal. The interviewer's posterior belief, b_i about the candidate's performance is a weighted average of the signal and the group mean:

$$b_i = E\left[y_i \mid \theta_i, g\right] = s_g \theta_i + \left(1 - s_g\right) \mu_g \tag{3}$$

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 *ex-ante* belief is greater if the signal is less informative.[11] In the extreme case in which it is completely uninformative, the interviewer's estimate of every candidate's performance is simply her belief about the mean given the candidate's gender, μ_g. By contrast, the interviewer's beliefs about the population distribution of ability would be irrelevant if the signal had no noise.

Statistical discrimination arises when an interviewer's prior belief differs by gender. The rating assigned to a man will then differ from that assigned to a woman given the same interview performance and any other information seen by the evaluator.

As a benchmark, suppose that interviewers believe the variance of ability, σ_g^2, to be the same for both genders.[12] This implies that $s_m = s_f = s$. Then the gender difference

[11] Alternatively, the interviewer will place more weight on her *ex-ante* belief if he or she is confident of that opinion in the sense that σ_g^2 is small.

[12] We consider the implications of relaxing this assumption in Appendix D.1. Note that differing prior variances—holding fixed the mean—leads to lower ratings for the high-variance group at the high end (for the same signal) but higher ratings at the low end.

in beliefs about code quality for a given signal realization, θ_i, is:

$$\text{Gender Gap} \mid \theta_i = E\left[y_i \mid \theta_i, m\right] - E\left[y_i \mid \theta_i, f\right] = (1-s)\left(\mu_m - \mu_f\right). \tag{4}$$

Equation (4) shows that beliefs—and thus interview ratings—will reflect the interviewer's preconceptions about the performance levels of men and women. Fixing the candidate's interview performance, this implies a gender gap in evaluations. The gap is larger if the signal is noisier so that σ_ε^2 is larger, or the interviewer's beliefs are more strongly held so that σ_g^2 is smaller.

Since the gender gap in Equation (4) is conditional on interview performance, it constitutes discrimination. Nonetheless, it is referred to as *rational* if interviewers' prior beliefs are correct. In this case, 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 μ_m and μ_f reflects a mistaken belief (a "bias"). This is *non-rational* statistical discrimination.

2.2.1 Testing for Non-Rational Statistical Discrimination

Letting μ_g^* be the true average ability of gender g candidates, the unconditional gap in beliefs is the expectation of Equation (4) over the signal distribution.

$$\text{Unconditional Gender Gap} = s\underbrace{\left(\mu_m^* - \mu_f^*\right)}_{\textbf{True gap}} + (1-s)\underbrace{\left(\mu_m - \mu_f\right)}_{\textbf{Believed gap}} \tag{5}$$

The effect of providing more information is that s increases. Holding fixed an interviewer's prior beliefs about the distributions of coding ability among men and women, the interviewer then places more weight on the signal they observe, which reduces the role for preconceptions about gender differences in ability.[13] The effect on the gender gap in beliefs depends on whether interviewers believe that the gap in coding ability is larger or smaller than it is in reality. This gives us a testable prediction which motivates Experiment I.

Prediction 1. (Non-rational Statistical Discrimination): *If evaluators believe incorrectly that women are less skilled coders, more precise information about performance should reduce the gender gap, holding the evaluation environment otherwise fixed.*

[13]The distribution of coding quality need not be invariant, since less precise information undermines the incentive to exert effort (Craig, 2023). In our setting, however, the set of coding solutions is fixed.

We note that the additional information provided as part of Experiment I is more accurately represented by the receipt of a binary signal. We study this case in Appendix C, and show that the core insight of Prediction 1 is preserved.

2.2.2 Testing for Rational Statistical Discrimination

If the gender gap is driven by rational statistical discrimination, then the provision of additional information about performance should have little impact on the gender gap in beliefs.

Prediction 2. (Rational Statistical Discrimination): *If evaluators believe correctly that women are less skilled coders than men, both the following must be true:*

(a) Conditional on the written code, women should be penalized relative to men.

(b) There is a difference in code quality between men and women. If information about prior beliefs is available, these should also favor men.

Prediction 2(a) can be tested by comparing blind evaluations to non-blind evaluations of the same code. Testing Prediction 2(b) is difficult in most settings, but we have compelling ways to do so in this setting. First, we measured a gender gap conditional on measures of code quality. We were able to do so using automated quality measures. Second, we can have code re-evaluated in a blind setting in which gender bias is not possible. We do this as part of Experiment II. Third, we collect information about prior beliefs in Experiment II as well.

2.3 Non-Statistical Discrimination

There may also be bias in ratings that is not explained by beliefs. First, evaluators are taste-based discriminators, who universally penalize women relative to men as in Becker (1957). In this case, knowing the coder's gender should introduce bias. Another possibility is that evaluators unconsciously (or "implicitly") discriminate. Bias may then only arise (or will be exacerbated) when gender is made salient through profile photographs or extended live interaction, with less or no bias arising simply because gender is observed. Such context-dependent amplification would be predicted by implicit discrimination and stereotypes (Bertrand et al., 2005). On the other hand, bias could be reduced by interaction. This would be in line with the 'contact hypothesis' (Allport et al., 1954; Lenz and Mittlaender, 2022).[14]

[14]Both possibilities could also be classified as taste-based bias, but they differ from static bias as in Becker (1957). Our aim here is simply to detect this type of context-dependency.

2.3.1 Testing for Taste-Based Discrimination

To test for fixed taste-based bias in the sense of Becker (1957), we can compare blind to non-blind evaluations of the same code, holding all else constant.

$$\text{Bias}(b_i \mid e = nb) = R(b_i \mid g_i = m, e = nb) - R(b_i \mid g_i = f, e = nb) \tag{6}$$

Prediction 3. (Taste-based Discrimination): *Taste-based discrimination implies a gap in non-blind evaluations favoring men, with no such gap in blind evaluations.*

(a) If there is no gender gap in average performance, the comparison of blind to non-blind evaluations reveals the extent of taste-based discrimination.

(b) If there is a gender gap in average performance, the same test reveals both taste-based and statistical discrimination.

2.3.2 Testing for Bias Introduced by Personal Interaction

Finally, we can assess whether bias is amplified or reduced by personal interaction by comparing the gender gaps in non-blind ratings with and without live interaction, while ensuring that the evaluations settings are otherwise as close as possible.

$$\begin{aligned}\Delta\text{Bias}(\textit{live} \text{ vs. } nb) = {} & R(b_i \mid g_i = m, e = \textit{live}) - R(b_i \mid g_i = f, e = \textit{live}) \\ & - [R(b_i \mid g_i = m, e = nb) - R(b_i \mid g_i = f, e = nb)]\end{aligned} \tag{7}$$

Prediction 4. (Bias from Personal Interaction): *If live interaction amplifies or reduces bias, then the gender gap in non-blind evaluations will be higher or lower on the platform than in re-evaluations without personal interaction.*

3 Providing Objective Information

Starting on July 8, 2017, the platform rolled out a powerful new diagnostic to verify the quality of code written on the platform (see Sections 1.6 and 1.7). Because the roll-out was randomized, we can use it to test Prediction 1, that is whether the gender gap in code quality ratings is driven by incorrect beliefs that women are less competent coders than men.

3.1 Intervention

The new tool provided automated ("unit") tests which assessed whether the code ran without errors, and produced the correct answers for test cases. Users could choose

to activate the tests by pressing a button (see Figure D1) and run them at any time. When they did, results of the unit tests were then visible to both the evaluator and interviewee before subjective ratings were chosen.

3.2 Treatment Assignment

Treatment assignment was randomized by the platform. The share of users treated at least once increased from July 2017 until all users were treated in October 27, 2017. During this roll-out period, we have data for all 6,401 sessions and 3,167 interviewees. When a new user i was paired to another user j, there were two possibilities. First, if both i and j were new users or had only been in the control condition in the past, the pair was randomized into treatment with a 7 percent probability. Once treated, a user always remained in treatment for future interactions. Second, any candidate matched with a partner who was already in the treatment condition was themselves treated (without randomization).

This nonstandard randomization motivates robustness tests in Section 3.5. However, we note that baseline characteristics are quite balanced between the treated and the control groups, as shown in Table D3. The main concern is that users' experience with the platform might differ between treatment and control, as treatment is an absorbing state. Therefore, in additional specifications, we control for date fixed effects, and in some specifications control for the likelihood of being treated.

3.3 Differences in Activation

Either the interviewer or interviewee could choose whether to activate the device during the interview, and not all did. We account for this using two-stage least squares (2SLS). We start with an Intention-to-Treat (ITT) model for men and women separately:

$$Y_{it} = \beta_g T_{it} + \theta_t + \epsilon_{it} \tag{8}$$

where Y_{it} is the score of individual i on date t, and θ_t are date fixed effects. $T_{it} = 1$ if the feature was enabled for a pair of users, and 0 otherwise. The ITT for gender $g \in \{m, f\}$ is β_g from Equation (8). Standard errors are clustered at the date level.

To account for differences in activation, we then also estimate the treatment effect on the treated (TOT) for each gender by using treatment assignment as an instrument for actual treatment. Specifically, we estimate the following model using 2SLS:

$$Y_{it} = \delta_g D_{it} + \lambda_t + \eta_{it} \tag{9}$$

$$D_{it} = \pi_g T_{it} + \zeta_t + \nu_{it} \tag{10}$$

where Y_{it} is the outcome of user i at time t; D_{it} is a dummy for whether the user activated the tests; T_{it} is an indicator of whether the pair was assigned to treatment; and λ_t and ζ_t are time fixed effects. Standard errors are clustered at the date level.

The coefficients of interest here are treatment effects by gender (β_g and δ_g). The sample includes all platform users between July 8, 2017 when the automated coding measure was first introduced, and October 27, 2017 when its access was generalized.

3.4 Result: No Reduction In The Gender Gap

We begin our analysis studying the activation decision and the impact of the new information on gender gaps in subjective ratings. We then look at whether differences in objective performance are related to differences in ratings.

Estimates from Equation (8) and (9) are shown in Table 1. 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 (8). The second column presents 2SLS estimates. The first stages are summarized in the lower sub-panels. Appendix D.2 provides information about the compliers.

First Stage: Activation. 71 percent of users enabled the objective code quality tests, when available. This strong first stage suggests that the code quality ratings were observed and valued by participants.[15]

Treatment Effects on Subjective Ratings. Both men and women in the treated group receive higher ratings than their peers in the untreated group for all the ratings. The largest effects are on dimensions where the unit tests likely shed the most direct light, including the code quality and problem solving ratings. We also see improvements in communication ratings, which may reflect improvements in how participants talk about their code when more information about quality is available. Likeability ratings increase slightly. On net, we see an improvement in assessments of hireability.

[15]We observe a slightly weaker first stage for women (0.678, S.D=0.016) than for men (0.721, S.D=0.016). This is a small difference, but could reflect relative under-confidence of women (Mobius et al., 2022) or attention discrimination (Bartoš et al., 2016). We cannot distinguish these two hypotheses because we cannot observe whether the evaluator or interviewee activated the tests.

Despite the increase in overall ratings, treatment did not disproportionately increase ratings for women. Instead, the increases in ratings are generally slightly larger for men, although our estimates are noisy. This is especially the case for coding and likability, where the effects are only marginally significant for women. In summary, gender gaps in ratings persist following the introduction of the unit tests. This contradicts Prediction 1 (Non-rational Statistical Discrimination), suggesting that non-rational statistical discrimination cannot explain the gender gaps on the platform. Rational statistical discrimination (Prediction 2) remains a possibility, subject to further tests below.

Why Would Ratings Increase? Our results indicate that the gender gaps persisted with more information, although ratings increased across the board. We evaluate alternative explanations for increase in ratings in Appendix D.1. Our leading explanation is that evaluators were unduly pessimistic for all coders, and potentially more about men than women. As we discuss in Section 4, we find some evidence consistent with this pattern when we collect information about prior beliefs in Experiment II.

3.5 Robustness Checks

Table D1 provides robustness checks to probe the validity of our results. Panel A shows a baseline in which we estimate the ITT model interacted by gender. In Panels B and C, we add month-of-interview, and then date-of-interview fixed effects. These adjust for changes in the share of users treated over time, and changes in user composition. The interaction of treatment with gender remains imprecisely estimated, still suggesting a slight widening of the gender gap. We control for individual characteristics in Panel D and find the same results. Including interviewee-fixed-effects in Panel H attenuates the treatment coefficients, with the interaction coefficient γ statistically insignificant. To ensure our results are not sensitive to the sample period, we expand our sample to include the pre-treatment period: The coefficients shrink slightly but the results are similar. To address the risk of endogenous matching between users, we control for a propensity score measuring the likelihood of being assigned to treatment.[16] The results are shown in Panel G of Table D1. Controlling for the propensity score does not affect our results.

We next show that differential assignments of problem and evaluator by gender are unlikely to drive our results. Table D2 shows that women are not more likely to be

[16]To estimate the propensity score, we use month-of-interview fixed effects and (for both the interviewer and interviewee) 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, and gender.

assigned a difficult problem (column 1), and are not more likely to be matched with a harsh evaluator—defined as interviewers whose average coding ratings (excluding the focal session's rating) is below the median (Columns 3 and 4).

Conditional on an individual's covariates and their partner's, treatment assignment should be nearly as good as random, especially because the matching algorithm used by the platform uses the same characteristics. Nonetheless, we explore changes in user composition over time and in response to treatment. The results are reassuring.[17] Figure D2 shows that the gender composition of users did not change with the introduction of the unit tests, and Figure D3 confirms that there are no changes in the characteristics of first-time female users around time the tests were introduced. Finally, Figure D4 plots the shares of high-performing first-time female and male users and shows that they follow a parallel increase over time. Thus, the quality of first-time users increases over time, but not differentially by gender.

Given the small gender differences in activation of the unit tests, we explore the possibility that there is differential selection by gender into activation. A potential reason for this to occur would be if one group were less likely to take the tests due to lower self-confidence. We assess this in Figure D5, which shows the share of unit tests passed versus the number taken, separately for male and female users. It shows that use of the tests varies similarly with objective performance for men and women.

4 Blind and Non-Blind Code Evaluation

To distinguish between the remaining mechanisms highlighted in Section 2, we used coding solutions written by platform users in another randomized experiment. This experiment used a within-subject design, with new evaluators asked to assess code written by men and women in a "blind" setting where gender was masked, and a "non-blind" setting in which gender was revealed via the coder's name.[18]

We first test whether the gender gap reflects code quality not captured by unit tests, by examining ratings under the blind condition where gender is unobserved. We then assess taste-based discrimination by comparing ratings for the same code in blind ($e = b$) and non-blind ($e = nb$) settings. Finally, we evaluate the role of live interactions by contrasting platform ratings with non-blind experimental evaluations

[17] Our main specifications nonetheless control for date-of-interview fixed effects to minimize any concern that such changes could affect one gender more than the other.

[18] The RCT was pre-registered on December 14, 2022. ID: AEARCTR-0009816. The pre-analysis plan is on the AEA RCT registry (updated: Feb 17, 2023).

of identical code conducted without in-person contact.

Evaluators for this second experiment were not drawn from the set of users on the platform. However, they were selected to be at a similar stage in their careers and to be very similar in characteristics. Specifically, they were mainly Bachelor's and Master's level computer science students with familiarity in the relevant programming languages. Table E3 presents characteristics for the experimental evaluators.

4.1 Empirical Design

To select code blocks for the experiment, we restrict to what we refer to as the experimental sample. We drop observations without unit test scores, keep only the most common programming languages (C++, Java, and Python), restrict to code with length no more than one standard deviation from the mean, and only consider the first attempt in cases where a given participant attempts the same problem twice. Finally, we exclude names that are uncommon or where gender is otherwise ambiguous.

Descriptive statistics from each step of the sample construction are presented in Tables E1. For each coding problem and language pair, we selected code blocks by stratifying by gender, race, and coding performance (whether the code passed all unit tests or not). Within each of these cells, we randomly picked one code block for the experiment. This yields a final sample of 456 code blocks. Table E2 presents summary statistics for this sample.

Each evaluator is assigned four coding blocks in a random order. They evaluate these on the same Likert scales from 1 to 4 as on the platform, but without live interaction. We also asked the experimental evaluator for a prediction of: (1) the share of unit tests the code block passed; (2) whether a human evaluator judged that the coder passed or failed the interview; and (3) the percent chance that the candidate was later invited for a real interview for a role involving coding. This allows us to draw a more direct link between our findings and hiring outcomes.

To measure participants' priors, we exposed them to three different vignettes before they performed their evaluation tasks. We asked them to predict the performance of three different hypothetical coders. We cross-randomized the first name (alternating gender) and the skill level for each vignette (see Appendix E). Finally, to assess the importance of image concerns, we asked participants at the end of the survey to guess the purpose of the study.

4.1.1 Treatment

Of the four blocks presented to an evaluator, two were "blind", and two "non-blind", with the order randomized. Within each arm, one code block was written by a man, and one by a woman. The order was again randomized. In the non-blind condition, gender was revealed via the given name of the coder. In addition, a box was shown with an avatar that revealed gender but no other aspect of a person's identity. In the blind condition, gender was hidden: Only the initial of the given name was seen, with no avatar. An example of each treatment condition is presented in Figures E2a-E3b.

4.1.2 Identification With The Within-Subject Design

The use of a within-subject design to identify treatment effects requires stronger assumptions than between-subject randomization, but can lead to substantial power gains (List, 2025). First, the panel is balanced: despite some attrition, it did not vary by treatment, and Table E4 shows evaluator characteristics are balanced across treatment orderings. Second, treatment-control comparisons must be temporally stable; to address potential fatigue or attention lapses, all specifications control for the number of scripts reviewed. Third, causal transience requires treatment effects not to depend on ordering; we find no such effects but still control for order. Finally, subjects generally could not infer the experiment's purpose, alleviating concerns about priming in the blind condition (see Section 4.3).

4.1.3 Incentives

Participants were incentivized in several ways. First, they were paid a participation fee of $10, plus a piece rate of $10 per script they evaluated. Second, they received bonus payments of $2 for each accurate predictions they make for the objective code quality and hireability measures per code block. Third, the 10 best evaluators could earn a cash prize of $500. Finally, we provided a non-financial but potentially powerful incentive by selecting a set of evaluators to participate in the Creative Destruction Lab 2023 Super Session. This brought real networking opportunities with world-class entrepreneurs, investors and scientists with high-potential startup founders.

4.1.4 Econometric Specifications

Our primary aim is to test whether revealing gender changes the gender gap in ratings. To do so, we use the following specification.

$$
\begin{aligned}
Y_{ij} &= \beta_1 \times \text{Female\_Coder}_j + \beta_2 \times \text{NB}_{ij} + \beta_3 \times \text{NB}_{ij} \times \text{Female\_Coder}_j \\
&+ \beta_4 \times \text{High\_Performer}_j + \beta_5 \times \text{Treatment\_Order}_i \\
&+ \sum_{k=1}^{4} \gamma_{jk} \mathbb{1}(\text{Script\_Order}_j = k) + \pi_{p(j)} + \delta_i + \epsilon_{ij}
\end{aligned} \tag{11}
$$

Here, we indicate treatment by defining $\text{NB}_j = 0$ for blind evaluation j, and $\text{NB}_j = 1$ for non-blind evaluation. Treatment_Order_i is an indicator for the randomly assigned treatment order ("non-blind then blind" condition versus "blind then non-blind"); and $\text{Script\_Order}_j = k$ is used to construct indicators that a given code block was the kth block the coder evaluated, to account for fatigue and learning. High_Performer_j indicates whether the code passed all unit tests or not. We include problem fixed effects, $\pi_{p(j)}$. In some specifications, we include evaluator fixed effects (δ_i) and additional controls. Standard errors are clustered at the evaluator level.

4.2 Results

No Gender Differences In Code Quality. Figure 4 presents our main results and Table 3 the corresponding estimates. The estimate of β_1 shows that code blocks written by women do not receive lower blind ratings, predicted scores or interview chance. If anything, the coefficients are positive, although we cannot rule out zero or small negative coefficients. This blind comparison of evaluations rules out meaningful gender differences in coding style, which could drive gender disparities in interviews but would not be captured by the unit tests. In particular, the confidence interval excludes effects as large as the previously documented gap (12 percent of a S.D). This result also contradicts Prediction 2 (Rational Statistical Discrimination).

No Bias When Gender Is Revealed. Turning to the comparisons of treatments, our estimate of the effect of making evaluation non-blind (β_2 in Equation 11) is negative on average, but the confidence interval includes zero. More importantly, the coefficient on the interaction with Female_Coder_i (β_3) is positive rather than negative, though imprecisely estimated. In this sense, do not find evidence of systematic gender bias that arises when gender is revealed by the first name. This contradicts Predictions 2 (Rational Statistical Discrimination) and 3 (Taste-Based Discrimination), suggesting that these theories do not explain the gender gaps we see.

Prior Beliefs. Experiment II allows us to explore participants' prior beliefs about the coding ability of men and women. Figure E1 shows the distributions of respondents'

prior beliefs about performance on the unit tests. They split by gender and by the skill level reported in the vignette, ranging from a B.Sc to a Master's in computer science with various years of work experience. On average, prior beliefs tend to be similar for men and women, as reflected by the vertical continuous lines which show the mean reported prior. For comparison, the vertical dashed lines show the share of tests actually passed by coders of each gender. This provides an additional test of Prediction 2(b): Rational statistical discrimination would imply more pessimistic prior beliefs about women than men, which we do not find.

There are two additional lessons from Figure E1. First, participants tend to be too pessimistic across the board about the coders in the vignettes, despite having been told that 82 percent of all users pass the unit tests. This could help explain why the introduction of the unit tests in Experiment I increased ratings for both men and women. Second, priors for men and women are quite similar on average, yet men perform slightly better on these tests (although not nearly enough to explain the gap in ratings between male and female coders). In retrospect, this result is again consistent with the results of Experiment I, providing a reason why introducing the unit tests did not succeed in reducing the gender gap in evaluations.

4.3 Comparability of Contexts

Our experiment was constructed to closely mirror the platform. Evaluators were selected to be very similar to those on the platform, both in terms of the stage they were at in their career and other characteristics. The rating scales and code they evaluated were both identical. The main difference is the removal of live interaction.

Despite this close match in characteristics, differences between the samples of evaluators, incentives, or image concerns could in principle contribute to the difference in non-blind gender gaps between the platform and experimental contexts. We explore these possibilities below, and argue that they are unlikely to be driving our results. First, we explore how our results change if we re-weight our regressions to more exactly match the composition of users on the in terms of educational degree and gender (Table E5). Our experimental results are qualitatively unchanged in this reweighted sample, with nearly identical levels of bias in all specifications. Additionally, we can calculate the gender gap in ratings for participants on the platform who have the same student status, and work experience as evaluators in the experiment. Table E6 shows the results: We find a larger rather than a smaller gender gap in the restricted sample,

which suggests that differences in experience and qualifications between samples are not driving our results.

We designed the incentives in our experiment to encourage evaluators to behave as professionally as those on the platform. There are inevitably differences in incentives in the experiment, but there are several reasons to think that participants are motivated to provide accurate assessments of code quality in both settings. First, Figure E4 documents a robust correlation for male-written codes between ratings in Experiment I and platform ratings for the same code in the non-blind version, despite the fact that the relationship is likely attenuated by noise. This supports the idea that evaluators are answering the coding evaluation question in a similar way in both contexts for male coders. The correlation is weaker for female-written code. This may be explained by a reduction in bias, which we later argue arises when live interaction is removed in the experiment. Second, we explore whether our experimental results hold in alternative samples to account for inattention. We restrict to our "high quality" sample, namely restricting to participants who passed the first attention check question, and excluding respondents whose survey completion time falls within the bottom 10th (less than 8 minutes) and top 90th percentiles (4 hours or more). Results are then presented in Table E7, and point to similar effects as in the whole sample.

We designed the study to minimize experimenter demand effects, but evaluate this possibility. At the end of the study, we asked participants to guess its purpose. Of 565 participants, only 22 mentioned discrimination (but not gender), 4 mentioned gender (but not discrimination), and 9 guessed that it was about gender discrimination. Participants largely assumed we were studying the determinants of perceived code quality. Some viewed it as a useful professional opportunity, with several asking whether we had more work of this kind for them. Table E8 shows that our results are robust to the exclusion of participants who correctly inferred that the study was about discrimination or gender.

5 Decoding Personal Interaction

5.1 Unconscious Bias Triggered by Personal Interaction

Our results show that the gender gap only arises when the interviewer and interviewee interact live. This is evident in Figure 5, which compares the gender gaps in standardized coding subjective ratings for the same code in the blind and live settings,

compared to the non-blind setting, and is in line with Prediction 4 (Bias from Personal Interaction). One explanation for this set of results is that bias is only triggered when gender differences in mannerisms and behavior are noticed during live interaction. This aligns with the concept of "implicit" bias (Bertrand et al., 2005; Carlana, 2019; Hangartner et al., 2021; Barron et al., 2022; Cunningham and de Quidt, 2022), which could be viewed as a form of taste-based bias that only manifests with extended interaction. Below, we provide two analyses supporting implicit bias as an explanation.

Association with IAT Scores. By harnessing the linkage between the platform data and individual-level LinkedIn information, we collect evaluators' higher education institutions. In turn, this allows us to compare ratings assigned by evaluators who attended an institution in geographic areas with high Implicit Association Test (IAT) scores—indicating more prejudice towards women in science (measured from Harvard's Project Implicit)—to those educated in areas with lower IAT scores. Figure 6 shows that the gender gap in coding ratings is significantly larger for interviewers educated in high-IAT regions.[19]

Interaction Duration. We also find that the gender gap in ratings widens with longer live interactions, consistent with unconscious bias as extended exchanges allow mannerisms to surface or may heighten evaluator fatigue.

To study the effects of longer interactions and avoid confounding the duration of the interaction with the individual's own coding proficiency, we separate the duration of the interaction due to own coding duration and partner's coding duration. Results are presented in Table 4. Columns 1 and 2 show that own coding duration is negatively associated with final subjective coding ratings, including when we control for coder's objective score (columns 1 and 2) and coder fixed effect (column 2), with no differential effect by candidate's gender. Next, we show results when we control separately for an interviewee's own coding duration, and that of their interview partner. Our coefficient of interest is Partner Coding Duration $\times$ Female. We find that the gender gap increases when their interview partner's coding duration is longer: A fifteen-minute increase in the length of the session is associated with a gender gap that is 2.6 percent of a standard deviation wider. These results are robust to the inclusion of problem (columns 3 to 6) and coder fixed effects (columns 3 and 5). When we include all range of fixed effects

[19]We define a high IAT area as a metropolitan statistical area with an average IAT Gender-Science score above the US median of 0.31. Estimates by subgroup are presented in Table F1. The distribution of IAT scores across geographic areas in our sample is provided in Figure F1.

including evaluator fixed effects (column 6), the effect is not statistically significant anymore but remains of the same magnitude.

One possible threat to interpreting this result as an effect of extended personal interaction is that partners who take longer are, on average, also of lower ability. If low ability coders are more biased in their evaluations, then the apparent increase in gender bias with partner duration would be confounded. However, column (7) of Table 4 shows that there are larger gender gaps disfavoring women when the evaluator is *higher* ability, as measured by their objective score. This is the opposite direction from what would be required to explain the effect of duration on the gender gap.

To quantify the importance of the gaps we see in these in-person skill assessments, we use the Revelio data to gauge the impact of closing those gaps on employment at six top tech companies (Microsoft, Amazon, Google, Apple, Facebook, and Palantir). These companies are known to incorporate these use in their recruitment processes. Our back-of-the-envelope calculations suggest that closing the gap in subjective coding ratings (12 percent of a standard deviation) predicts a 0.62 percentage point increase in the probability of being hired at one of these companies within two years of an individual obtaining their first computer science related degree. This would be a 2.3 percent increase in the employment of women in software engineering positions at these companies. Further details of this calculation are available in Appendix F.1.

5.2 Competing Explanations for The Role of Personal Interaction

An alternative explanation for the importance of personal interaction is that there are other factors that affect the rating when live interaction is present, which do not manifest in the code itself. While we cannot quantify every aspect of these interactions, we can provide data on several of the most plausible hypotheses.

Coding Time. One possibility is that women take longer to solve a given coding problem. This could be the case if women are slower coders, revise their code more, or receive more help from their interviewer. However, we observe time use on the platform. While there is a rating penalty for slow coders, there are no significant gender differences in coding time (see Figure 7). Controlling for interviewees' coding duration therefore does not reduce the gender gap in ratings (Table 4).

Communication Style. An alternative possibility is that men and women talk about their code differently. If women are less effective at communicating, this could in-

troduce a gender gap that is not there when code is evaluated alone. Figure 8 plots the average subjective ratings for communication by objective performance (share of unit tests passed), separated by gender. While both high and low performing women receive systematically lower subjective coding than men who perform equally well (Figure 3), the communication ratings of men and women are similar across the objective performance distribution. This suggests that gender differences in communication styles are unlikely to explain the persistent gender gaps in coding subjective ratings. Indeed, controlling for communication scores leads to only a small reduction in the gender gap in code quality ratings.

5.3 Decoding Gendered Interactions: Video Analysis

To gain further insight into how personal interactions differ by gender and may trigger gender bias, we analyze interview videos. Our partner platform did not store video recordings, so we instead use data from a very similar platform that has made a subset of mock interview videos publicly available on YouTube. Unlike our partner platform, this second platform pairs job-seekers with a professionals with experience hiring for major technology firms.[20] The videos are anonymized and include candidates' and interviewers' voices but not their faces. Each video is accompanied by the interviewer's subjective coding rating of the candidate's performance. We analyze all 189 of the videos that are available. While the sample is small and selected—users remain anonymous but have consented to making their videos publicly available—our analysis offers suggestive evidence on how language and non-verbal behaviors are associate with both gender and performance.

Review of the videos proceeded three steps. First, gender was manually recorded for the voices in each video. Second, we used Google's Gemini Flash LLM (2.0) to provide quantitative evaluations on discourse markers and non-verbal aspects for both interviewer and candidate, with each video assessed thirty times.[21] The LLM was not asked about gender. Table G1 presents the summary statistics, and shows that approximately 7 percent of the videos feature female candidates.

Using this data, we look at candidate's and interviewers' behaviors from the interview video and their relationship with subjective coding ratings. Figure 9 presents differences in language and non-verbal aspects of the interaction, as observed in the

[20] Note that Table D5 column (5) suggests that the work experience of the interviewer on the platform did not play a significant role on gender gaps in ratings.

[21] Detailed prompts are presented in Appendix G.1.

videos. Convergence across iterations are presented in Figure G1 and G2 respectively. Estimates are obtained from separate linear regressions of each outcome on the gender of the candidate, with standard errors clustered at the interview level. Panel A presents results for the behaviors of the candidates.

Female candidates are significantly more likely to be perceived as friendly. Women are also more likely to exhibit gendered paralinguistic cues: They are significantly more likely to use rising intonation ("upticks"), a pattern sometimes interpreted as signaling lack of confidence, particularly when produced by female voices.[22] They are more likely to apologize (54 percent of a S.D, p-value= 0.051). Point estimates for "confidence", "hesitation", "share speech" (female interviewees speak relatively less during the interview) and women's propensity to ask clarifying questions are not statistically significant.

Next, we analyze interviewers' behaviors depending on the gender of the candidate (Figure 9 Panel B). When the candidate is a woman, interviewers are significantly more likely to be condescending (37 percent of a S.D), harsh (41 percent of a S.D), and impatient (33 percent of a S.D respectively). They are less likely to build effective rapport with the candidate and to be respectful (20 and 15 percent of a S.D respectively). They are also more likely to interrupt (40 percent of a S.D), less likely to explain the problem clearly (-17 percent of a S.D; p-value=0.067) and less likely to actively listen, although this last coefficient is not statistically significant.

Figure 9 Panel C and D shows the relationship between these observed behaviors and the interviewer's subjective coding rating of the candidate's performance, as reported on the platform. Estimates are derived from univariate and multivariate linear regressions of the behavior measures on the interviewer's subjective coding rating, with standard errors clustered at the interview level. Panel C shows results for the candidates' behaviors. We find a significant negative association between the subjective ratings and the use of apologetic language, a positive association with the AI-based measure of confidence and with the "share speech" measure, and a negative one with hesitation. The coefficients for the presence of upticks and the AI-based measure of friendliness are smaller, and non-statistically significant for the multivariate model.

Turning to interviewers, we find that aggressive tone is associated with significantly lower subjective coding ratings (Figure 9 Panel D). This includes our measures of condescension, harshness and impatience. These patterns are more pronounced in

[22] See Levon and Ye (2020) and Divakaruni et al. (2023).

the univariate models as the dimensions are correlated. Explaining, active listening (univariate model only) and effective rapport building are positively associated with final ratings. We find no consistent relationship for interruptions and respect, although the latter may stem from correlation between behaviors in the multivariate analysis.

Our analysis highlights important gender differences in behavioral cues exhibited by both candidates and interviewers. Women are more likely to apologize, which employers may interpret as signaling lower ability (Liu and Mo, 2024), potentially harming promotion prospects (Benson et al., 2024). They are also generally less effective at self-promotion (Murciano-Goroff 2018, Exley and Kessler 2022, Haegele 2024).

The results are also in line with dynamics in high-stake interactions documented by Dupas et al. (2021), who show that female candidates are interrupted more often and face harsher tone in economics seminars used for recruitment. The reciprocal nature of the interaction makes it difficult to disentangle whether hostile behavior prompts increased apologizing among women, or whether women's apologies trigger more hostility. However, our context highlights the role of interviews as environments in which gendered behavioral cues can surface, and potentially bias evaluations of the technical ability of candidates who use different language and mannerisms.

6 Conclusion

We use administrative data, two field experiments and a video analysis, to study gender bias in technical interviews in the technology industry. Across our two experiments, we shed light on the nature of gender bias in a highly-skilled industry where women are chronically underrepresented.

We find that gender bias in performance hinges on personal interaction. Our results are most consistent with the literature on implicit discrimination and stereotypes. Put differently, in line with the sociology literature, biases are more likely to emerge when individuals are "doing gender" (West and Zimmerman, 1987) during personal interaction, rather than when gender is merely revealed by a person's name. This conclusion is further supported by the association of longer interactions with larger gender gaps, and the presence of a strong association between gender gaps and IAT scores where the evaluator was educated. Our video analysis reveals strikingly different patterns of behavior of the interviewer when the candidate is a woman, which may underpin the gender bias that we see arise only when personal interactions are present.

Our results are potentially consequential. Bias in evaluations would lead to systemic bias in hiring decisions even if hiring managers are themselves unbiased (Bohren et al., 2022), or could affect women's promotion in the long-run (Sarsons, 2022). Our back-of-the-envelope calculations suggest that eliminating the in-person coding interview could raise female employment in software engineering positions in top technology companies by 2.3 percent.

It remains an important question for future research precisely which settings and modes of interaction lead to such bias. Some have argued that inter-group contact can reduce biases (Pettigrew and Tropp, 2006), yet implicit bias persists even in settings with extensive contact (Carlana, 2019; Alesina et al., 2023). We go further, and find that sustained interaction with a given individual appears to amplify bias. We also provide evidence on which aspects of the interaction differ depending on the gender of the coder, and how these factors correlate with ratings. More work is needed to understand the effects of the mode of interaction, and the extent to which genuine information is conveyed in personal interaction.

Our analysis suggests innovative ways to mitigate bias in performance evaluation. The gender gap in our setting is eliminated when personal interaction is removed. Decoupling coding evaluations from live interviews may therefore provide a way to reduce biases in the evaluation of cognitive skills, because the technical evaluations will not themselves involve personal interaction. By contrast, the status quo in which interpersonal and technical skills are assessed simultaneously may be leading to assessments that are flawed on both dimensions. This is particularly important, given that technical interviews combine task performance with interpersonal interaction, even though such tasks are rarely performed live in the actual workplace. We caution that it could be more problematic to remove personal interaction entirely: This could harm female candidates who have relatively strong social skills, which are becoming increasingly valued in the labor market (Deming, 2017).

References

Abadie, Alberto, "Semiparametric instrumental variable estimation of treatment response models," *Journal of Econometrics*, 2003, *113* (2), pp. 231–263.

Abramitzky, Ran and Leah Boustan, "Immigration in American Economic History," *Journal of Economic Literature*, 2017, *55* (4), pp. 1311–1345.

_ , Leah Platt Boustan, and Katherine Eriksson, "Europe's Tired, Poor, Huddled masses: Self-Selection and Economic Outcomes in the Age of Mass Migration," *American Economic Review*, 2012, *102* (5), pp. 1832–1856.

_ , _ , and _ , "A Nation of Immigrants: Assimilation and Economic Outcomes in the Age of Mass Migration," *Journal of Political Economy*, 2014, *122* (3), 467–506.

Alesina, Alberto, Michela Carlana, Eliana La Ferrara, and Paolo Pinotti, "Revealing Stereotypes: Evidence from Immigrants in Schools," *American Economic Review*, 2023, *forthcoming*.

Allport, Gordon Willard, Kenneth Clark, and Thomas Pettigrew, *The Nature of Prejudice*, Addison-wesley publishing company Cambridge, MA, 1954.

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

Avery, Mallory, Andreas Leibbrandt, and Joseph Vecci, "Does Artificial Intelligence Help or Hurt Gender Diversity? Evidence from Two Field Experiments on Recruitment in Tech," *Evidence from Two Field Experiments on Recruitment in Tech (February 14, 2023)*, 2023.

Barron, Kai, Ruth Ditlmann, Stefan Gehrig, and Sebastian Schweighofer-Kodritsch, "Explicit and Implicit Belief-Based Gender Discrimination: A Hiring Experiment," Technical Report, CESifo Working Paper 2022.

Bartoš, Vojtěch, Michal Bauer, Julie Chytilová, and Filip Matějka, "Attention Discrimination: Theory and Field Experiments with Monitoring Information Acquisition," *American Economic Review*, 2016, *106* (6), 1437–75.

Becker, Gary S, *The Economics of Discrimination*, University of Chicago Press, 1957.

Behroozi, Mahnaz, Shivani Shirolkar, Titus Barik, and Chris Parnin, "Debugging hiring: What went right and what went wrong in the technical interview process," in "Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering: Software Engineering in Society" 2020, pp. 71–80.

Bellemare, Charles, Marion Goussé, Guy Lacroix, and Steeve Marchand, "Physical Disability and Labor Market Discrimination: Evidence from a Video Résumé Field Experiment," *American Economic Journal: Applied Economics*, 2023, *15* (4), 452–476.

Benson, Alan, Danielle Li, and Kelly Shue, "Potential and the gender promotions gap," Technical Report, SSRN 2024.

Bertrand, Marianne and Esther Duflo, "Field experiments on discrimination," in "Handbook of Economic Field Experiments," Vol. 1, Elsevier, 2017, pp. 309–393.

_ and Sendhil Mullainathan, "Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination," *American Economic Review*, 2004, *94* (4), 991–1013.

_ , **Dolly Chugh, and Sendhil Mullainathan**, "Implicit Discrimination," *The American Economic Review*, 2005, *95* (2), 94–98.

Bohren, J Aislinn, Alex Imas, and Michael Rosenberg, "The dynamics of discrimination: Theory and evidence," *American Economic Review*, 2019, *109* (10), 3395–3436.

_ , **Kareem Haggag, Alex Imas, and Devin G Pope**, "Inaccurate statistical discrimination: An identification problem," *Review of Economics and Statistics*, 2023, pp. 1–45.

_ , **Peter Hull, and Alex Imas**, "Systemic discrimination: Theory and measurement," Technical Report, National Bureau of Economic Research 2022.

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer, "Stereotypes," *Quarterly Journal of Economics*, 2016, *131* (4), pp. 1753–1794.

Carlana, Michela, "Implicit Stereotypes: Evidence from Teachers' Gender Bias*," *Quarterly Journal of Economics*, 03 2019, *134* (3), 1163–1224.

Coate, Stephen and Glenn Loury, "Antidiscrimination Enforcement and the Problem of Patronization," *American Economic Review*, 1993, *83* (2), pp. 92–98.

Collis, Manuela R and Clémentine Van Effenterre, "Workplace Hostility," Technical Report, Working Paper 2025.

Craig, Ashley C., "Optimal Taxation with Spillovers from Employer Learning," *American Economic Journal: Economic Policy*, 2023, *14* (2), pp. 82–125.

_ **and Roland G. Fryer**, "Complementary Bias: A Model of Two-Sided Statistical Discrimination," 2019.

Cullen, Zoë and Ricardo Perez-Truglia, "The old boys? club: Schmoozing and the gender gap," *American Economic Review*, 2023, *113* (7), 1703–1740.

_ , **Bobak Pakzad-Hurson, and Ricardo Perez-Truglia**, "Home Sweet Home: How Much Do Employees Value Remote Work?," Technical Report 33383 2025.

Cunningham, Tom and Jonathan de Quidt, "Implicit Preferences," Technical Report, CEPR Discussion Paper 2022.

Deming, David J, "The Growing Importance of Social Skills in the Labor Market," *The Quarterly Journal of Economics*, 2017, *132* (4), 1593–1640.

Divakaruni, Anantha, Laura Fritsch, Howard Jones, and Alan D Morrison, "Market Reactions to Gendered Speech Patterns: Uptalk, Earnings Calls, and the# MeToo Movement," Technical Report, SSRN 2023.

Dupas, Pascaline, Alicia Sasser Modestino, Muriel Niederle, Justin Wolfers et al., "Gender and the dynamics of economics seminars," Technical Report, National Bureau of Economic Research 2021.

Emanuel, Natalia, Emma Harrington, and Amanda Pallais, "The power of proximity to coworkers: training for tomorrow or productivity today?," Technical Report, National Bureau of Economic Research 2023.

Exley, Christine L and Judd B Kessler, "The gender gap in self-promotion," *The Quarterly Journal of Economics*, 2022, *137* (3), 1345–1381.

Farber, Henry S, Dan Silverman, and Till Von Wachter, "Determinants of callbacks to job applications: An audit study," *American Economic Review*, 2016, *106* (5), 314–18.

Feld, Jan, Edwin Ip, Andreas Leibbrandt, and Joseph Vecci, "Identifying and Overcoming Gender Barriers in Tech: A Field Experiment on Inaccurate Statistical Discrimination," Technical Report, CESifo Working Paper 2022.

Goldin, Claudia and Cecilia Rouse, "Orchestrating impartiality: The impact of "blind" auditions on female musicians," *American Economic Review*, 2000, *90* (4), pp. 715–741.

Haegele, Ingrid, "The broken rung: Gender and the leadership gap," Technical Report, arXiv preprint arXiv:2404.07750 2024.

Handlan, Amy and Haoyu Sheng, "Gender and Tone in Recorded Economics Presentations: Audio Analysis with Machine Learning," Technical Report 2023.

Hangartner, D., D. Kopp, and M. Siegenthaler, "Monitoring Hiring Discrimination through Online Recruitment Platforms," *Nature*, 2021, *589*, 572—576.

Kenneth, J Arrow, "The Theory of Discrimination," *Discrimination in Labor Markets*, 1973, *3*.

Kessler, Judd B, Corinne Low, and Xiaoyue Shan, "Lowering the playing field: Discrimination through sequential spillover effects," Technical Report, mimeo 2022.

Kline, Patrick, Evan K Rose, and Christopher R Walters, "Systemic Discrimination Among Large US Employers," *Quarterly Journal of Economics*, 2022, *137* (4), pp 1963–2036.

_ , _ , and _ , "A Discrimination Report Card," Technical Report, NBER 2023.

Kroft, Kory, Fabian Lange, and Matthew J Notowidigdo, "Duration dependence and labor market conditions: Evidence from a field experiment," *The Quarterly Journal of Economics*, 2013, *128* (3), 1123–1167.

Laakmann, Gayle, Mike Mroczka, Aline Lerner, and Nil Mamano, *Beyond Cracking the Technical Interview* 2024.

Lenz, Lisa and Sergio Mittlaender, "The effect of intergroup contact on discrimination," *Journal of Economic Psychology*, 2022, *89*, 102483.

Levashina, Julia, Christopher J Hartwell, Frederick P Morgeson, and Michael A Campion, "The structured employment interview: Narrative and quantitative review of the research literature," *Personnel psychology*, 2014, *67* (1), 241–293.

Levon, Erez and Yang Ye, "Language, indexicality and gender ideologies," *Gender and Language*, 2020, *14* (2), 123–151.

List, John A, "The nature and extent of discrimination in the marketplace: Evidence from the field," *The Quarterly Journal of Economics*, 2004, *119* (1), 49–89.

List, John A., *Experimental Economics: Theory and Practice*, University of Chicago Press, 2025.

Liu, Lily Liu and Marshall Mo, "Gender Gap in Apologies," Technical Report, unpublished 2024.

Loyalka, Prashant, Ou Lydia Liu, Guirong Li, Igor Chirikov, Elena Kardanova, Lin Gu, Guangming Ling, Ningning Yu, Fei Guo, Liping Ma et al., "Computer science skills across China, India, Russia, and the United States," *Proceedings of the National Academy of Sciences*, 2019, *116* (14), 6732–6736.

Lundberg, Shelly J. and Richard Startz, "Private Discrimination and Social Intervention in Competitive Labor Market," *American Economic Review*, 1983, *73* (3), 340–347.

Mobius, Markus M., Muriel Niederle, Paul Niehaus, and Tanya S. Rosenblat, "Managing Self- Confidence: Theory and Experimental Evidence," *Management Science*, 2022, *68* (11), 7793–8514.

Mocanu, Tatiana, "Designing Gender Equity: Evidence from Hiring Practices and Committees," 2023.

Murciano-Goroff, Raviv, "Missing Women in Tech: The Role of Self-Promotion in the Labor Market for Software Engineers," 2018.

Neumark, David, "Detecting discrimination in audit and correspondence studies," *Journal of Human Resources*, 2012, *47* (4), 1128–1157.

_ , Roy J Bank, and Kyle D Van Nort, "Sex Discrimination in Restaurant Hiring: An Audit Study," *Quarterly Journal of Economics*, 1996, *111* (3), pp 915–941.

Pettigrew, Thomas F and Linda R Tropp, "A meta-analytic test of intergroup contact theory," *Journal of Personality and Social Psychology*, 2006, *90* (5), 751–783.

Phelps, Edmund S, "The statistical theory of racism and sexism," *The American Economic Review*, 1972, *62* (4), 659–661.

Radbruch, Jonas and Amelie Schiprowski, "Interview sequences and the formation of subjective assessments," *Review of Economic Studies*, 2025, *92* (2), 1226–1256.

Reuben, Ernesto, Paola Sapienza, and Luigi Zingales, "How stereotypes impair women's careers in science," *Proceedings of the National Academy of Sciences*, 2014, *111* (12), 4403–4408.

Roussille, Nina, "The role of the ask gap in gender pay inequality," *Quarterly Journal of Economics*, 2024, *139* (3), 1557–1610.

Sarsons, Heather, "Interpreting Signals in the Labor Market: Evidence from Medical Referrals," 2022.

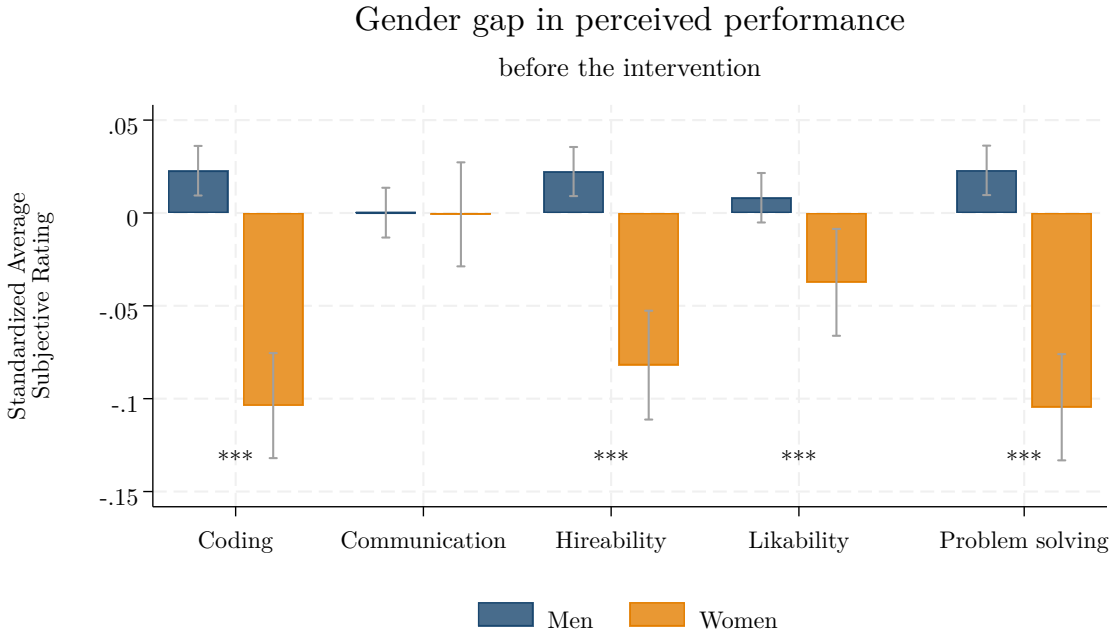
Shukla, Soumitra, "Making the Elite: Discrimination at Top Firms," Technical Report, Working Paper 2024.

Spencer, Steven J, Christine Logel, and Paul G Davies, "Stereotype threat," *Annual review of psychology*, 2016, *67*, 415–437.

West, Candace and Don H. Zimmerman, "Doing Gender," *Gender and Society*, 1987, *1* (2), 125–151.

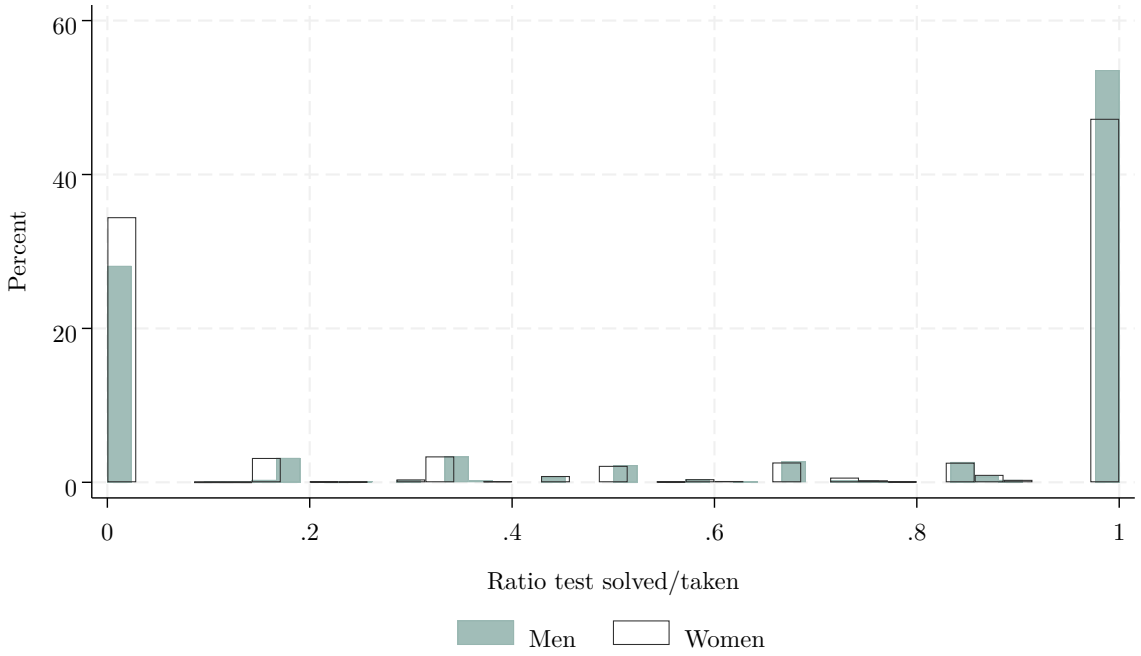
Tables and Figures

Figure 1: Pre-intervention Gender Gaps – Whole Sample



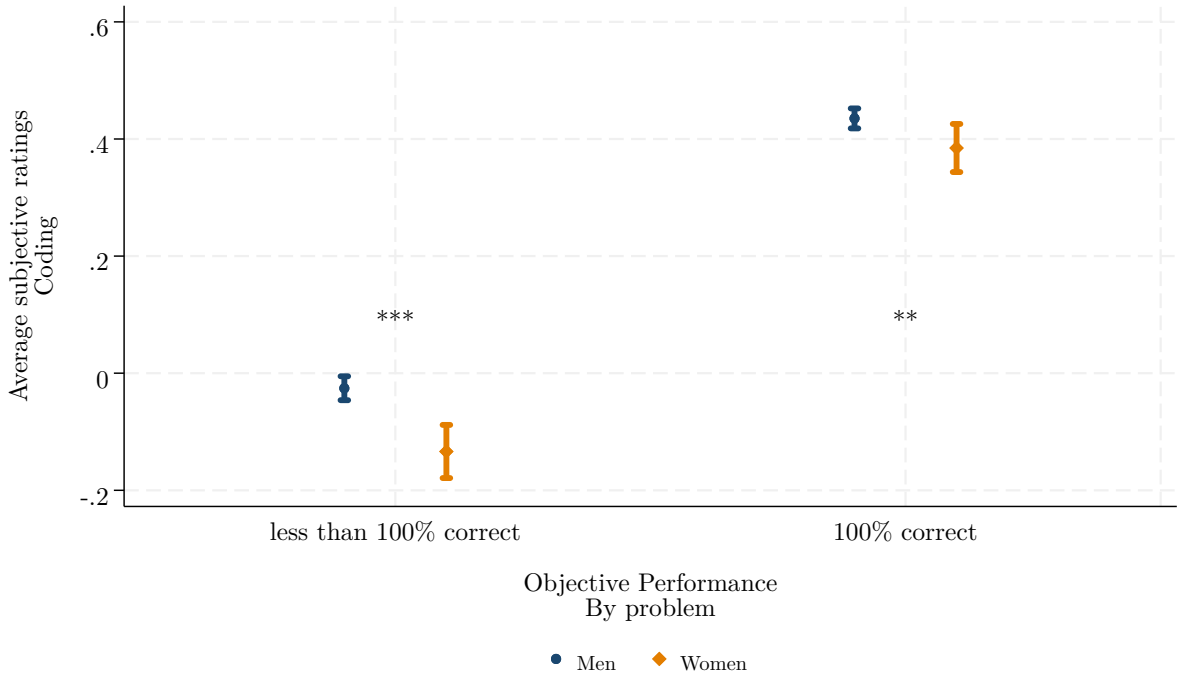
Notes: This figure shows the gender gap in peer-rated performance in five categories for standardized variables: coding, communication, hirability, likability and problem solving, for the whole sample for the pre-intervention period (from December 2017 to July 2017). Stars above a category indicate statistical significance of the gap at the one percent level, and the 95-percent confidence intervals of each bar are shown in gray.

Figure 2: Distribution of Objective Performance by Gender



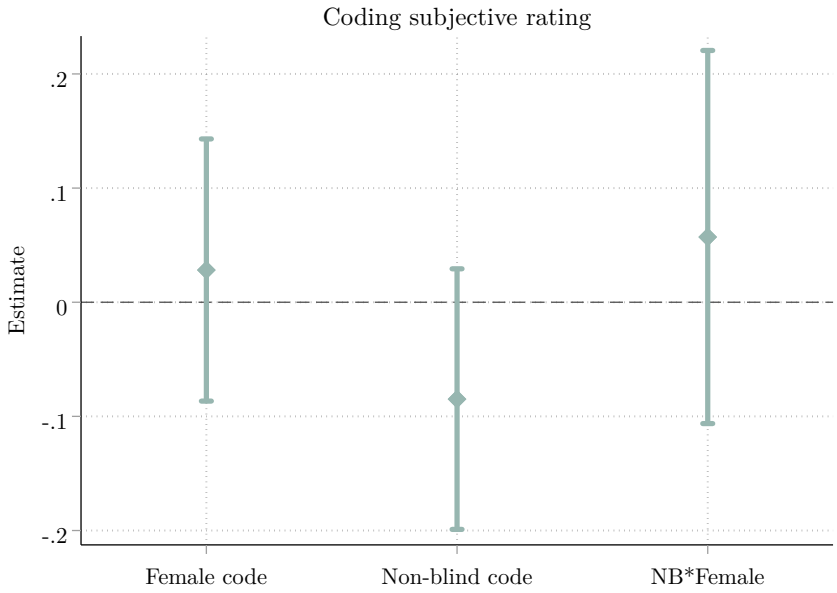
Notes: The figure presents the distribution of our objective performance measure (share of tests passed) by gender. These "unit tests" indicate whether the code ran and produced the correct answers to pre-defined test cases. The sample includes all platform users who activated the objective performance measure from July 2017 to April 2018.

Figure 3: Subjective Ratings by Objective Score — Coding



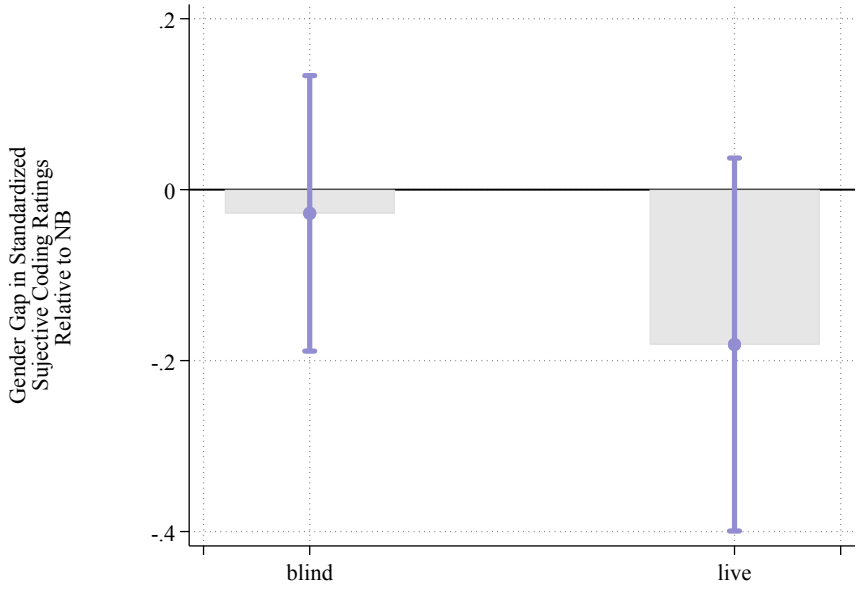
Notes: This figure shows the average subjective ratings for coding for high and low quality code blocks. Reflecting the bimodal distribution of objective performance, we define high quality as passing all tests. Results for men are in blue, and results for women are in orange. The sample includes all platform users who activated the objective performance measure from July 2017 to April 2018.

Figure 4: Blinding Experiment — Effect Of Blinding On Gender Gaps



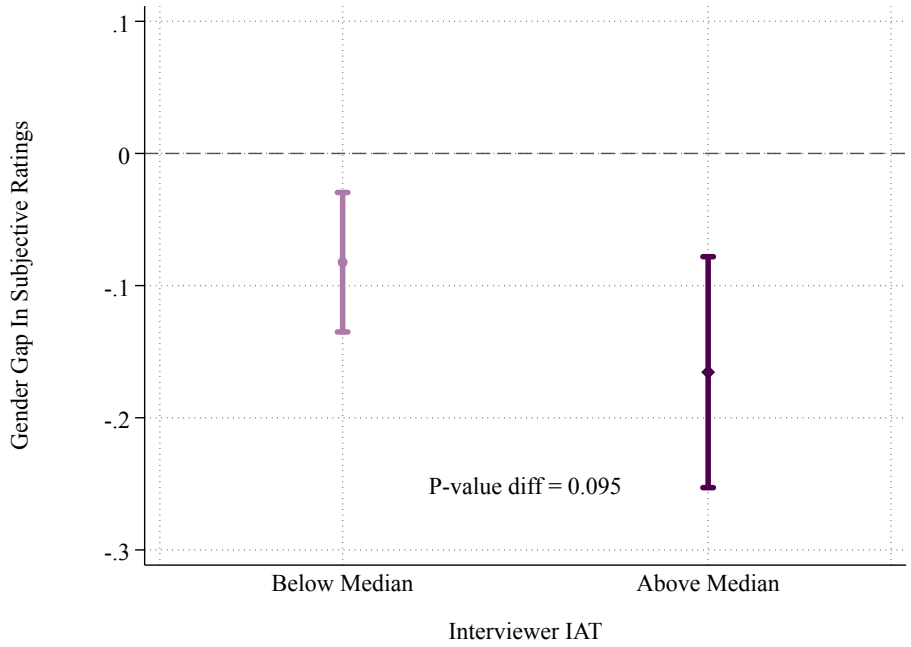
Notes: This figure shows the results from Experiment II (see Section 4). The regression specification is as described in Equation (11), controlling for evaluator fixed effects. The dependent variables are the (standardized) subjective coding ratings. The 95-percent confidence intervals shown are based on standard errors clustered at the evaluator level. Corresponding estimates are presented in the first column of Table 3.

Figure 5: Gender Gap By Evaluation Environment



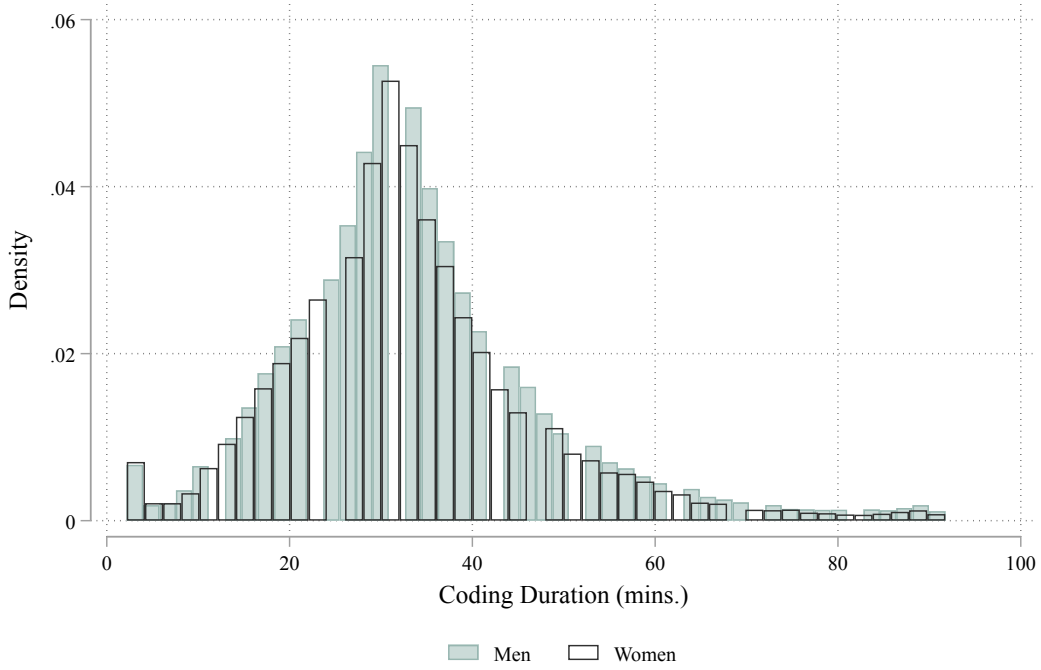
Notes: This figure shows the gender gap in subjective coding ratings relative to the non-blind environment, separately, in the blind and live settings. The latter is the empirical counterpart of Equation (7) in Section 2. Live evaluations of the same code blocks are from Experiment I, while blind, and non-blind, evaluations are from Experiment II. The sample is therefore the set of scripts for which we have evaluations in all three settings. Estimated coefficients are from regressing standardized subjective ratings on a female dummy interacted with blind, for the first coefficient, and similarly in a separate regression for live in the second. For both cases, the reference group were code scripts written by male coders, and evaluated in a non-blind environment from Experiment II. Due to the design of Experiment II, we use multiple evaluations per script for the blind and non-blind contexts, whereas scripts in the live context from the platform have only one evaluation.

Figure 6: Gender Gap By Evaluator IAT



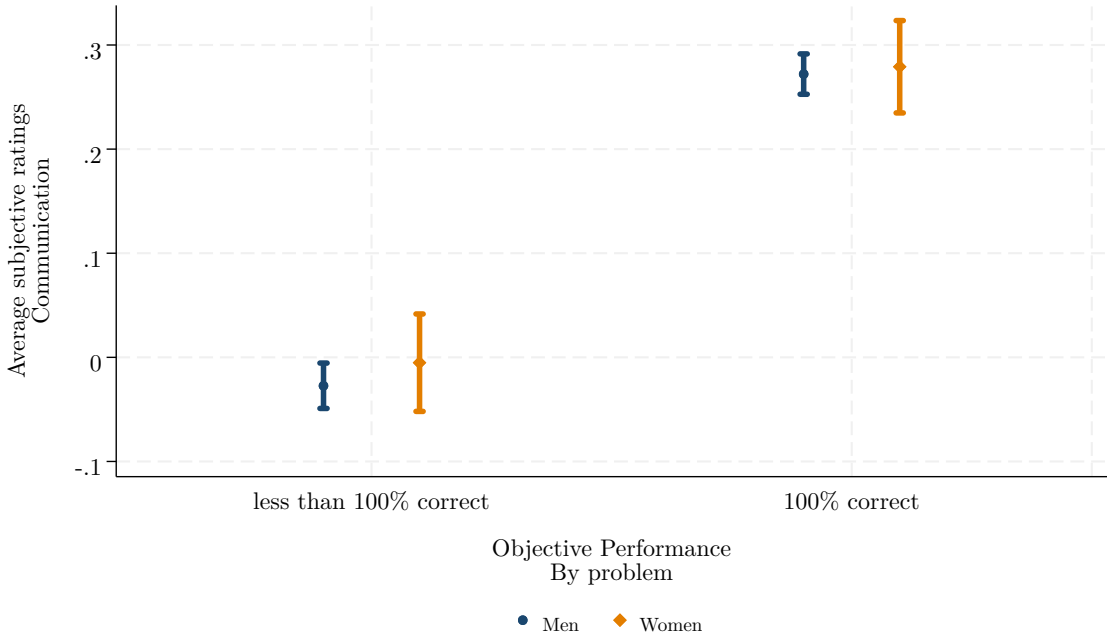
Notes: This figure shows the gender gap in ratings by evaluator's IAT. Average IAT score is calculated at the MSA level. MSAs are then classified as having either below or above median IAT score relative to other geographic areas. The distribution of IAT score is presented in Figure F1. Evaluators' graduating institutions are matched to their MSA allowing us to classify evaluators to below (above) median if they graduated from an institution located in an MSA with a below (above) median IAT score. Evaluators' institutions are obtained from LinkedIn data as described in Section 1.7. IAT scores are from the Gender-Science IAT module for the years 2018 and 2019 of the Harvard Implicit Project. Corresponding estimates are presented in Table F1.

Figure 7: Coding Duration By Gender



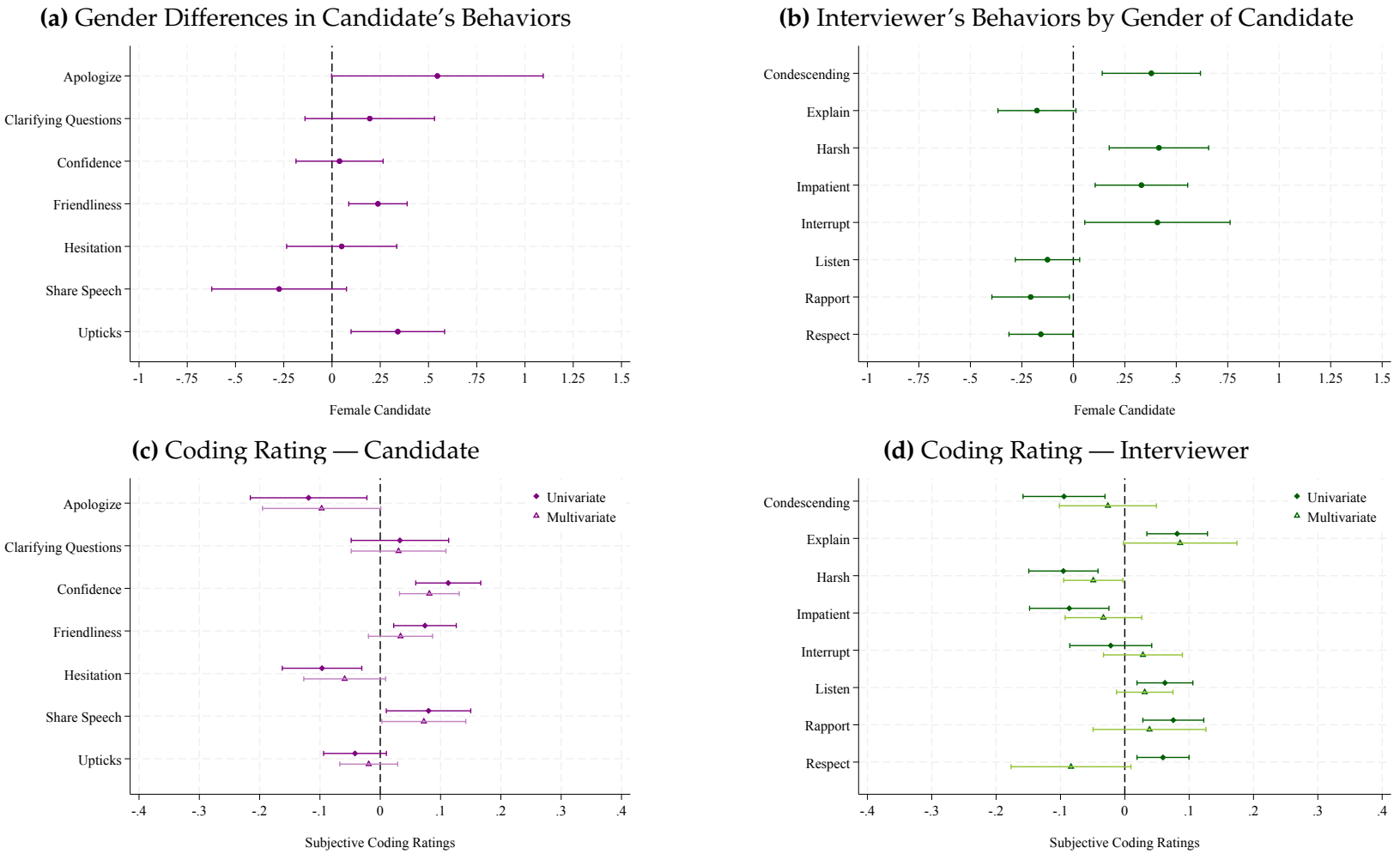
Notes: This figure shows the coding duration in minutes by gender for platform users. The sample includes platform users based in the US with non-missing gender from the January 2018 to May 2022.

Figure 8: Subjective Ratings by Objective Score — Communication



Notes: This figure shows the average subjective ratings for communication for high and low quality code blocks. Reflecting the bimodal distribution of objective performance, we define high quality as passing all tests. Results for men are in blue, and results for women are in orange. The sample includes all platform users who activated the objective performance measure from July 2017 to April 2018.

Figure 9: Video Analysis



Notes: This figure analyzes candidate's and interviewers' behaviors from the interview videos, and their the relationship with the interviewer's subjective coding rating of the candidate, as reported on the platform. Panel A presents gender differences in candidate's behaviors, as observed in the videos. Panel B presents differences in interviewer's behaviors by the candidate's gender, as observed in the videos. Panel C presents the relationship between the behaviors of the candidates and the interviewers' subjective coding ratings of the candidates, as reported on the platform. Panel D presents the relationship between the behaviors of the interviewers and the interviewers' subjective coding ratings of the candidates, as reported on the platform. The candidate's gender is recorded at the video/URL level, based on metadata coded directly from YouTube. Candidates' and interviewers' behaviors are evaluated using Google's Gemini Flash AI (version 2.0), with each video assessed thirty times. For Panel A and B, estimates are obtained from separate linear regressions of each outcome on the gender of the candidate, with standard errors clustered at the URL level. For Panel C and D, estimates are derived from both univariate and multivariate linear regressions of the behavior measures on the interviewer's subjective coding rating, with standard errors clustered at the URL level.

Table 1: Impact of the Introduction of the Automated Measure of Code Quality

Panel A: All

	Coding		**Problem solving**		**Likeability**		**Communication**		**Hirability**	
	ITT	2SLS	ITT	2SLS	ITT	2SLS	ITT	2SLS	ITT	2SLS
Treatment	0.147	0.205	0.211	0.295	0.086	0.120	0.198	0.277	0.169	0.237
s.d	(0.031)	(0.043)	(0.030)	(0.041)	(0.033)	(0.046)	(0.039)	(0.005)	(0.028)	(0.039)
P-value	0.000	0.000	0.000	0.000	0.012	0.010	0.000	0.000	0.000	0.000
N	11,029	11,029	11,029	11,029	11,029	11,029	11,029	11,029	11,049	11,049
First stage		0.714								
s.d		(0.009)								
P-value		0.000								
N		11,591								
F-stat		6084.30								

Panel B: Women Interviewees

	Coding		**Problem solving**		**Likeability**		**Communication**		**Hirability**	
	ITT	2SLS	ITT	2SLS	ITT	2SLS	ITT	2SLS	ITT	2SLS
Treatment	0.092	0.135	0.188	0.276	0.054	0.080	0.183	0.269	0.175	0.257
s.d	(0.081)	(0.114)	(0.073)	(0.103)	(0.080)	(0.114)	(0.073)	(0.104)	(0.080)	(0.113)
P-value	0.258	0.239	0.012	0.008	0.497	0.482	0.013	0.010	0.030	0.024
N	2,049	2,049	2,049	2,049	2,049	2,049	2,049	2,049	2,055	2,055
First stage		0.678								
s.d		(0.016)								
P-value		0.002								
N		2,151								
F-stat		2069.16								

Panel C: Men Interviewees

	Coding		**Problem solving**		**Likeability**		**Communication**		**Hirability**	
	ITT	2SLS	ITT	2SLS	ITT	2SLS	ITT	2SLS	ITT	2SLS
Treatment	0.162	0.225	0.218	0.302	0.093	0.129	0.199	0.276	0.168	0.234
s.d	(0.032)	(0.045)	(0.033)	(0.046)	(0.039)	(0.054)	(0.044)	(0.061)	(0.033)	(0.046)
P-value	0.000	0.000	0.000	0.000	0.019	0.016	0.000	0.000	0.000	0.000
N	8,980	8,980	8,980	8,980	8,980	8,980	8,980	8,980	8,994	8,994
First stage		0.721								
s.d		(0.016)								
P-value		0.000								
N		9,440								
F-stat		4392.79								

Notes: This table shows the main results from Experiment I (see Section 3). Both ITT and 2SLS models are shown, using the whole sample of platform users between July 8, 2017 when the automated coding measure was first introduced, and October 27, 2017 when it was generalized to all users, splitting by gender of the interviewee. For each of the five dimensions on which users are rated, the coefficient on treatment in each model is shown from left to right in the upper subpanels. The first stages are shown in the lower subpanels. Standard errors are clustered at the date level, and shown in parentheses.

Table 2: Automated Measure of Code Quality and Future Labor Market Outcomes

	Ln(first salary post graduation)		
	(1)	(2)	(3)
Female	-0.063*	-0.073*	-0.074*
	(0.036)	(0.044)	(0.043)
Non white	-0.040	-0.071	-0.070
	(0.035)	(0.046)	(0.046)
Masters Degree	0.126***	0.202***	0.200***
	(0.030)	(0.032)	(0.031)
Objective Score		0.052**	0.068**
		(0.024)	(0.032)
Objective Score × Female			-0.057
			(0.054)
City FE	Yes	Yes	Yes
Higher Education Institution FE	Yes	No	No
Observations	3,625	2,297	2,297

Notes: This table presents our analysis of labor market outcomes discussed in Section 1.7 and Appendix B. The coefficients come from Mincer-type regressions where the dependent variable is the (log) first salary post graduation using observations from participants of the platform data matched with the Revelio Lab database. Controls include the number of session on the platform and whether the participant had already graduated when they took sessions on the platform. Standard errors are clustered at the city-of-residence level, and shown in parentheses.

Table 3: Blinding Experiment — Effect Of Blinding On Gender Gaps

	Subjective coding rating		Unit test prediction		Interview prediction	
	(1)	(2)	(3)	(4)	(5)	(6)
Female code	0.027	0.023	0.192	0.198	0.025	0.023
	(0.059)	(0.059)	(0.180)	(0.182)	(0.050)	(0.050)
Non-blind code	-0.075	-0.080	-0.261	-0.252	-0.153**	-0.054
	(0.059)	(0.059)	(0.192)	(0.193)	(0.051)	(0.051)
Non-blind code×Female code	0.036	0.049	0.173	0.192	0.037	0.035
	(0.084)	(0.085)	(0.261)	(0.263)	(0.070)	(0.070)
Treatment order control	Yes	Yes	Yes	Yes	Yes	Yes
Order of scripts FE	Yes	Yes	Yes	Yes	Yes	Yes
Problem FE	Yes	Yes	Yes	Yes	Yes	Yes
Evaluator FE	No	Yes	No	Yes	No	Yes
Observations	2,323	2,292	2,323	2,292	2,704	2,704

Notes: This table provides results from Experiment II (see Section 4), testing the pre-registered hypothesis that revealing gender introduces a gender gap that penalizes women. The regression specification is as described in Equation (3). The dependent variables are the (standardized) subjective coding ratings (columns 1-2), participants' prediction of the unit tests passed by the code script (columns 3-4) and their prediction of the coder's probability of passing the interview (columns 5-6). The even columns include evaluator fixed effects. Standard errors are clustered at the evaluator level, and shown in parentheses.

Table 4: Interaction Duration & Gender Gap

	Subjective Coding Ratings						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Female	-0.119***			-0.127***		-0.132***	-0.061**
	(0.010)			(0.012)		(0.016)	(0.029)
Coding Duration	-0.041***	-0.031***		-0.056***	-0.055***	-0.082***	
	(0.004)	(0.005)		(0.005)	(0.006)	(0.008)	
Coding Duration × Female	-0.005	-0.005		-0.012	-0.004	-0.005	
	(0.010)	(0.011)		(0.012)	(0.015)	(0.016)	
Partner Coding Duration			-0.038***	-0.063***	-0.058***	-0.085***	-0.051***
			(0.006)	(0.005)	(0.007)	(0.007)	(0.005)
Partner Coding Duration x Female			-0.026*	-0.020*	-0.026*	-0.025	-0.021*
			(0.014)	(0.012)	(0.015)	(0.015)	(0.011)
Partner Obj Score			-0.033**	-0.028**	-0.052***	-0.008	0.004
			(0.014)	(0.014)	(0.014)	(0.017)	(0.013)
Partner Obj Score x Female							-0.074**
							(0.032)
Obj Score	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Problem FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Coder FE	No	Yes	Yes	No	Yes	Yes	No
Evaluator FE	No	No	No	No	No	Yes	No
Observations	51,036	46,501	28,613	32,630	28,613	23,457	32,630

Notes: This table provides results for the gender gap in subjective ratings testing for the hypothesis that longer interviews are associated with a higher gender gap. The sample includes all platform users based in the US for the period from January 2018 to May 2022. All specifications include question fixed effects, and control for coder's objective score. Columns (1)-(2) show the effect of *own* coding duration on subjective ratings and allows for differences by gender. Columns (3)-(5) show the effect of partners' coding duration on ratings. Columns 2, 3, 5 and 6 include coder fixed effects, and column 6 evaluator fixed effects. Finally, column 7 allows for differences of partner objective score by gender. Standard errors are in parentheses.

(For Online Publication)

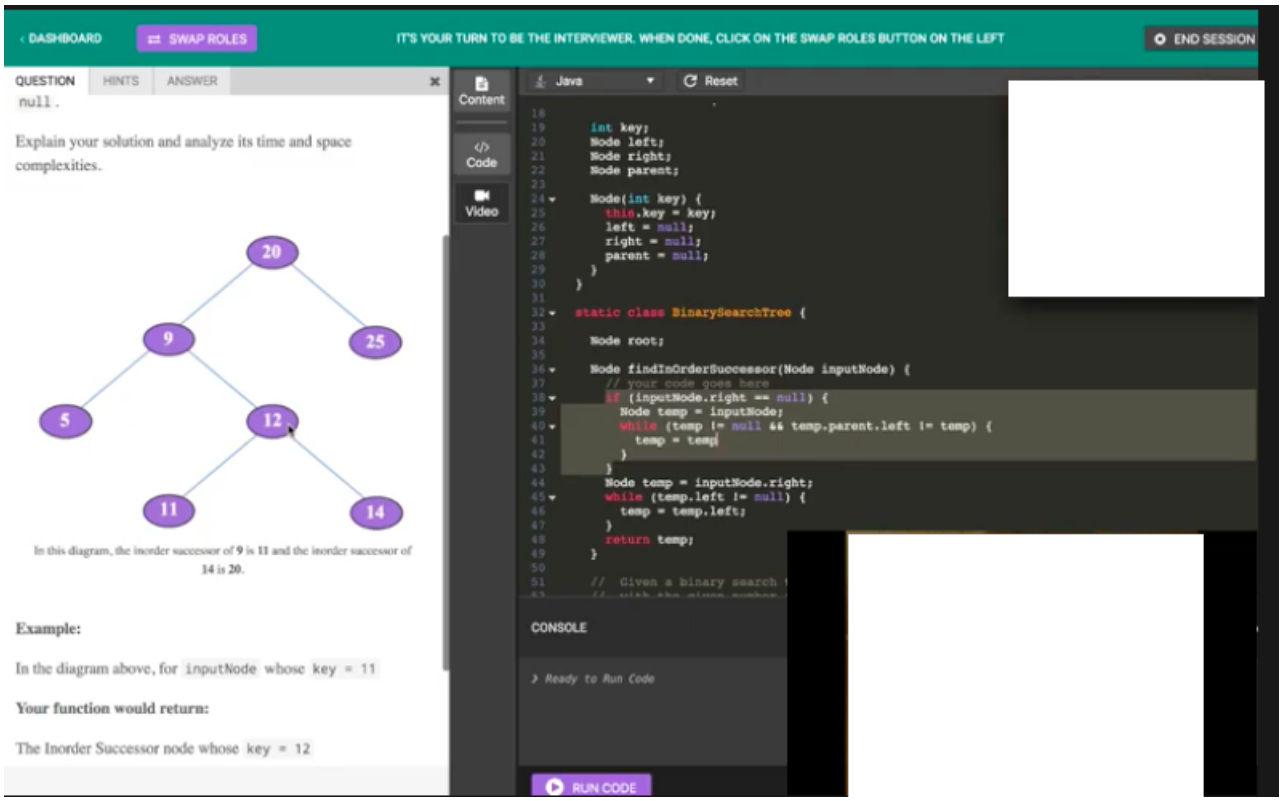
Appendix to

Decoding Gender Bias in Interviews

Abdelrahman Amer, Ashley C. Craig and Clémentine Van Effenterre

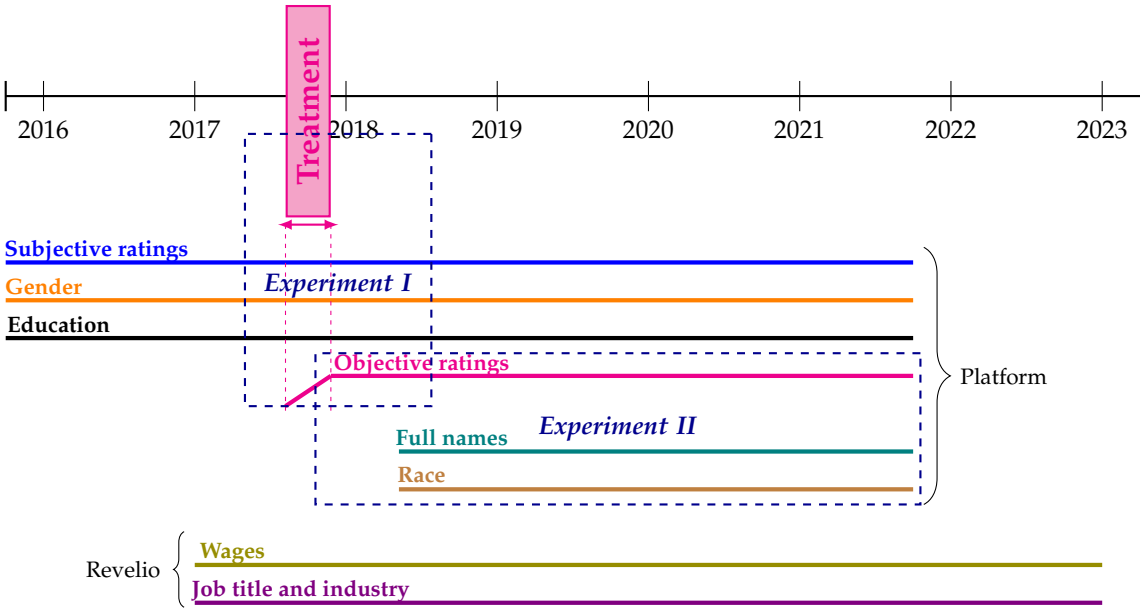
Appendix A Institutional details

Figure A1: Environment of the Platform



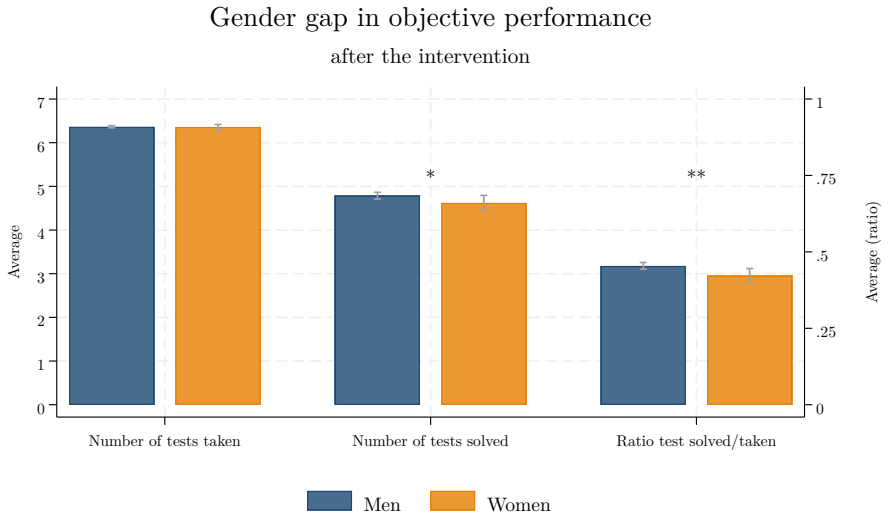
Notes: This figure shows the platform layout for a mock interview. The white squares indicate where the videos of the candidate and the interviewer are displayed.

Figure A2: Summary of Data Availability



Notes: This diagram shows the data infrastructure we use to build Experiment I and II and the validation exercise using labor market outcomes from Revelio Lab. Experiment I is described and analyzed in Section 3. Experiment II is described and analyzed in Section 4. The Revelio data are described in Section 1.7, with further discussion in Appendix B.

Figure A3: Gender Gap In Objective Performance After The Intervention



Notes: This figure presents the level of objective performance for men and women after the intervention in terms of number of tests taken, number of tests solved or failed (right y-axis), and the share of unit tests passed (right y-axis). The sample includes all platform users who activated the objective performance measure from July 2017 to April 2018.

Table A1: Descriptive Statistics Pre-Intervention

Number of sessions	30,466
Number of interviewees	12,960
Number of interviewers	12,707
Number of problems	31
Share of female interviewees	16.46
Share of female interviewers	16.44

Panel A: All

Variable	**Mean**	**Std. Dev.**	**Min.**	**Max.**	**N**
Country: USA	0.716	0.451	0	1	60,513
Interviewee's deg.: computer science	0.661	0.473	0	1	60,483
Interviewee without working experience	0.267	0.442	0	1	60,508
Interviewee with a graduate degree	0.45	0.497	0	1	60,513
Interviewee Preparation Level	2.897	0.798	1	5	60,307

Panel B: Women

Variable	**Mean**	**Std. Dev.**	**Min.**	**Max.**	**N**
Country: USA	0.796	0.403	0	1	9,959
Interviewee's degree : computer science	0.652	0.476	0	1	9,959
Interviewee without working experience	0.309	0.462	0	1	9,957
Interviewee with a graduate degree	0.514	0.5	0	1	9,959
Interviewee Preparation Level	2.779	0.786	1	5	9,940

Panel C: Men

Variable	**Mean**	**Std. Dev.**	**Min.**	**Max.**	**N**
Country: USA	0.701	0.458	0	1	50,554
Interviewee's deg.: computer science	0.662	0.473	0	1	50,524
Interviewee without working experience	0.259	0.438	0	1	50,551
Interviewee with a graduate degree	0.437	0.496	0	1	50,554
Interviewee Preparation Level	2.92	0.799	1	5	50,367

Notes: This table shows descriptive statistics for the sample of interviews we analyze in Section 1.4, from December 2015 to July 2017, before the introduction of objective code quality measures. The top panel shows key aggregate statistics. The lower three panels present summary statistics for interviewee characteristics overall, for men and for women respectively.

Table A2: Gender Gap in Subjective Ratings Pre-Intervention

	Coding				
	(1)	(2)	(3)	(4)	(5)
Interviewee female	-0.127***	-0.121***	-0.121***	-0.121***	-0.118***
	(0.016)	(0.016)	(0.016)	(0.018)	(0.019)
Observations	26,306	25,952	25,952	25,932	25,952
	Problem Solving				
	(1)	(2)	(3)	(4)	(5)
Interviewee female	-0.126***	-0.110***	-0.110***	-0.111***	-0.117***
	(0.016)	(0.016)	(0.016)	(0.018)	(0.018)
Observations	26,306	25,952	25,952	25,932	25,952
	Likability				
	(1)	(2)	(3)	(4)	(5)
Interviewee female	-0.042***	-0.042***	-0.042***	-0.043**	-0.045**
	(0.015)	(0.015)	(0.015)	(0.017)	(0.018)
Observations	26,306	25,952	25,952	25,932	25,952
	Communication				
	(1)	(2)	(3)	(4)	(5)
Interviewee female	-0.000	0.000	-0.000	-0.001	0.006
	(0.016)	(0.016)	(0.016)	(0.019)	(0.019)
Observations	26,306	25,952	25,952	25,932	25,952
	Hireability				
	(1)	(2)	(3)	(4)	(5)
Interviewee female	-0.104***	-0.101***	-0.101***	-0.102***	-0.095***
	(0.016)	(0.016)	(0.016)	(0.019)	(0.019)
Observations	26,264	25,911	25,911	25,911	25,911
Interviewee's controls	No	Yes	Yes	Yes	Yes
Interviewer's controls	No	Yes	Yes	Yes	Yes
Problem FE	No	No	No	Yes	No
Date FE	No	No	No	No	Yes

Notes: This table shows the estimation of the gender gap in subjective ratings pre-intervention from December 2015 to July 2017, using a linear regression model in which we progressively add controls (see Section 1.4). 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 problem fixed effects. In columns 5, we add date-of-interview fixed effects.

Table A3: Revelio & Platform Characteristics

	Revelio			Platform		
	N	Mean	SD	N	mean	SD
Panel A. Pre Intervention						
Share Female	118,978	0.23	0.42	6,786	0.19	0.39
Highest Degree Bachelor	118,978	0.70	0.46	6,786	0.57	0.49
Highest Degree Masters	118,978	0.29	0.45	6,786	0.36	0.48
Panel B. Post-Intervention						
Share Female	482,114	0.23	0.42	27,557	0.25	0.43
Share Non-white	482,114	0.46	0.50	27,557	0.61	0.49
Highest Degree Bachelor	482,114	0.75	0.43	27,557	0.50	0.50
Highest Degree Masters	482,114	0.25	0.43	27,557	0.42	0.49

Notes: This table presents demographic summary statistics for the CS graduating cohorts of 2016-2017 using Revelio database, and comparing it with demographics of the Platform users before (Panel A) and after the intervention (Panel B) in Experiment I.

Table A4: Gender Gap Reweighted

	Subjective Coding Ratings			
	Pre-intervention		Post-Intervention	
	Unweighted	Reweighted	Unweighted	Reweighted
	(1)	(2)	(3)	(4)
Female	-0.133***	-0.129***	-0.160***	-0.182***
	(0.015)	(0.016)	(0.007)	(0.008)
Problem FE	Yes	Yes	Yes	Yes
Evaluator FE	Yes	Yes	Yes	Yes
Observations	29,269	29,269	140,024	140,024

Notes: This table presents results for the gender gap in subjective coding ratings after reweighting observations on the Platform to be representative of characteristics of CS graduates on Revelio. In the pre-intervention period, we use the 2016 and 2017 graduate cohorts for reweighting. In the post-intervention period, we use the 2018 to 2022 cohorts. Columns (1) and (3) present unweighted results, in the pre- and post-intervention periods. Columns (2) and (4) are the reweighted results for the pre- and post-intervention periods respectively. Weights are obtained using the inverse probability of being on the platform. We use a probit regression in which we include the sociodemographic variables present both in Revelio and in the platform datasets.

Table A5: Gender Gap in Subjective Ratings Controlling for Objective Coding Quality Measure

	Coding					
	(1)	(2)	(3)	(4)	(5)	(6)
Interviewee female	-0.106***	-0.081***	-0.081***	-0.082***	-0.090***	-0.085***
	(0.018)	(0.017)	(0.017)	(0.017)	(0.017)	(0.020)
Objective performance		0.485***	0.485***	0.486***	0.543***	0.509***
		(0.0141)	(0.0141)	(0.0141)	(0.0239)	(0.0173)
Observations	19,559	19,559	19,559	19,559	18,139	19,559
	Problem Solving					
	(1)	(2)	(3)	(4)	(5)	(6)
Interviewee female	-0.139***	-0.110***	-0.110***	-0.111***	-0.116***	-0.121***
	(0.018)	(0.017)	(0.017)	(0.017)	(0.016)	(0.020)
Objective performance		0.565***	0.565***	0.566***	0.673***	0.598***
		(0.014)	(0.014)	(0.014)	(0.035)	(0.017)
Observations	19,559	19,559	19,559	19,559	18,139	19,559
	Likability					
	(1)	(2)	(3)	(4)	(5)	(6)
Interviewee female	-0.032*	-0.016	-0.016	-0.018	-0.012	-0.004
	(0.020)	(0.019)	(0.019)	(0.019)	(0.023)	(0.022)
Objective performance		0.317***	0.317***	0.318***	0.371***	0.322***
		(0.016)	(0.016)	(0.016)	(0.023)	(0.020)
Observations	19,559	19,559	19,559	19,559	18,139	19,559
	Communication					
	(1)	(2)	(3)	(4)	(5)	(6)
Interviewee female	-0.000	0.015	0.015	0.014	0.017	0.022
	(0.018)	(0.018)	(0.018)	(0.018)	(0.019)	(0.020)
Objective performance		0.303***	0.303***	0.304***	0.338***	0.271***
		(0.015)	(0.015)	(0.015)	(0.021)	(0.018)
Observations	19,559	19,559	19,559	19,559	18,139	19,559
	Hireability					
	(1)	(2)	(3)	(4)	(5)	(6)
Interviewee female	-0.053***	-0.040**	-0.040**	-0.041**	-0.039*	-0.029
	(0.019)	(0.018)	(0.018)	(0.018)	(0.019)	(0.020)
Objective performance		0.350***	0.350***	0.351***	0.422***	0.374***
		(0.015)	(0.015)	(0.015)	(0.022)	(0.019)
Observations	18,132	18,132	18,132	18,132	18,132	18,132
Interviewee's controls	No	No	Yes	Yes	Yes	Yes
Interviewer's controls	No	No	Yes	Yes	Yes	Yes
Problem FE	No	No	No	No	Yes	No
Date FE	No	No	No	No	No	Yes

Notes: This table shows the estimation of the gender gap in subjective ratings after July 2017, controlling for the objective coding quality measure (columns 2 to 6), using a linear regression model in which we progressively add controls. In column 3, 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 4 to 6, we control for the gender of the interviewer. In columns 5, we add problem fixed effects. In columns 6, we add date-of-interview fixed effects.

Appendix B Labor Market Data

In this Appendix, we describe how we link our data to labor market outcomes from Revelio labs, and analyze the merged dataset. The Revelio data contain information from publicly available LinkedIn profiles, and job posting boards. These data contain close to the universe of Computer Science (CS) graduates in the US labor market, and their job spells. We also observe an estimate of their salaries imputed using job posting data, H1B-visa records and the Current Population Survey.[A.1]

One concern with such data is that there may be some degree of sample selection. For example, only high achieving graduates might have profiles. However, we have two reasons to believe that this is less of a problem in our setting than others. First, participants on the platform are actively seeking employment in a CS related position, making an online presence highly desirable if not unavoidable. Second, the US produces around 60,000 computer science baccalaureates annually, and there are about this many such degrees in the Revelio data from 2016 to 2026.[A.2]

From the set of interviewees on the platform, we select those residing in the US who have a Bachelor's or Master's degree. We then match this sample to the universe of individuals in the Revelio data who attained a CS-related degree from a US institution. We use only exact matches based on their first and last name, and degree type. Observations matched to multiple Revelio profiles are dropped.[A.3] The final sample consists of 5,126 matched CS graduates from 2016 to 2023. We have unit test data for about 50 percent of this sample.

We use a Mincer-type wage regression of log earnings on individuals' unit test scores, their characteristics, year-of-graduation and city fixed effects. The main outcome is the first salary after graduation, although we also look at average salary after graduation. Results are presented in Table 2. Column (1) shows that there is a 6.3 percent residual gender gap for computer science graduates in their first salary after graduation. In column (2), we add the average objective measure of coding quality across all sessions on the platform, the number of past sessions on the platform and whether the participant had graduated at the time of their interview session.[A.4] We

[A.1] More detail regarding the Revelio data database is available www.reveliolabs.com.

[A.2] See Loyalka et al. (2019) for a cross-country analysis of CS university graduates.

[A.3] This follows the same matching method adopted by Abramitzky et al. (2012), Abramitzky et al. (2014) and Abramitzky and Boustan (2017).

[A.4] To reduce noise, we also tried re-weighting the regression for the number of sessions each user had

find a positive and statistically significant coefficient (0.052, SD=0.024) for the standardized objective score measure, which implies that going from the 25th to the 75th percentile of standardized score is associated with a wage increase of 4.5 percent.

Finally, we note that there is suggestive evidence of heterogenous returns of skills by gender in column 3, with little return of the objective measure of coding performance for women. However, the estimate for women is imprecise.

Appendix C Additional Theoretical Results

Our guiding model in Section 2 assumes that both prior distributions and the noise in interviewer signals are normally distributed. This is a commonly adopted assumption, which makes the model highly tractable. However, our results in Section 3 indicate that the additional signal that the new unit tests that were introduced provided a signal that is closer to binary (see Figure 2). For completeness, this appendix works through this case mathematically. The core insights are preserved.

C.1 Model Setup

Men and women have true performance distributions as follows.

$$y_i \sim \mathcal{F}_m \tag{A.1}$$

$$y_i \sim \mathcal{F}_f \tag{A.2}$$

These distributions are arbitrary here, but the distributions could be normal as in the baseline model in Section 2.

A test provides additional information about performance. Specifically, the outcome of the test is $t_i \in \{0, 1\}$, aligning with the unit tests introduced in Experiment I. Specifically let's assume that $t_i = 1$ if $y_i > y^*$ and $t_i = 0$ otherwise. Below, we will normalize $y^* = 0$, as the units of performance make no differnce to the results.

C.2 Belief Formation With Correct Priors

Conditional on passing the test ($t_i = 1$), expected productivity is:

$$E\left(y_i \mid t_i = 1\right) = E\left(y_i \mid y_i > 0\right) = \frac{\int_0^\infty y_i d\mathcal{F}_g}{\Pr\left(y_i > 0, g\right)} \tag{A.3}$$

on the platform. The results are qualitatively similar.

Alternatively, if $t_i = 0$:

$$E\left(y_i \mid t_i = 0\right) = E\left(y_i \mid y_i \leq 0\right) = \frac{\int_{-\infty}^{0} y_i d\mathcal{F}_g}{\Pr\left(y_i \leq 0, g\right)} \tag{A.4}$$

The unconditional expectation is therefore:

$$\begin{aligned} E\left(y_i\right) &= \Pr\left(y_i > 0, g\right) \times \frac{\int_{0}^{\infty} y_i d\mathcal{F}_g}{\Pr\left(y_i > 0, g\right)} + \Pr\left(y_i \leq 0, g\right) \times \frac{\int_{-\infty}^{0} y_i d\mathcal{F}_g}{\Pr\left(y_i \leq 0, g\right)} \\ &= \int_{-\infty}^{\infty} y_i d\mathcal{F}_g \end{aligned} \tag{A.5}$$

C.3 Belief Formation With Incorrect Priors

Next, suppose that the interviewer's priors are incorrect. They prior distributions are:

$$y_i \sim \widetilde{\mathcal{F}}_m \tag{A.6}$$

$$y_i \sim \widetilde{\mathcal{F}}_f \tag{A.7}$$

In this case, the ex ante expectation of an interviewer's expectation of productivity following the signal for someone of gender g is as follows.

$$\begin{aligned} \widetilde{E}\left(y_i\right) &= \Pr\left(y_i > 0, g\right) \times \frac{\int_{0}^{\infty} y_i d\widetilde{\mathcal{F}}_g}{\widetilde{\Pr}\left(y_i > 0, g\right)} + \Pr\left(y_i \leq 0, g\right) \times \frac{\int_{-\infty}^{0} y_i d\widetilde{\mathcal{F}}_g}{\widetilde{\Pr}\left(y_i \leq 0, g\right)} \\ &= \frac{\Pr\left(y_i > 0, g\right)}{\widetilde{\Pr}(y_i > 0, g)} \times \int_{0}^{\infty} y_i d\widetilde{\mathcal{F}}_g + \frac{\Pr(y_i \leq 0, g)}{\widetilde{\Pr}\left(y_i \leq 0, g\right)} \times \int_{0}^{\infty} y_i d\widetilde{\mathcal{F}}_g \\ &= \int_{-\infty}^{\infty} y_i d\widetilde{\mathcal{F}}_g + \left[\Pr(y_i > 0, g) - \widetilde{\Pr}(y_i > 0, g)\right] \times \left[\widetilde{E}(y_i \mid y_i > 0) - \widetilde{E}(y_i \mid y_i \leq 0)\right] \end{aligned} \tag{A.8}$$

Here, $\Pr\left(y_i > 0, g\right)$ and $\widetilde{\Pr}\left(y_i > 0, g\right)$ are the true and expected probabilities that $y_i > 0$ respectively; with parallel notation for $\Pr\left(y_i \leq 0, g\right)$ and $\widetilde{\Pr}\left(y_i \leq 0, g\right)$.

C.4 Effect of New Information

These equations allow us to quantify the expected movement of beliefs in response to new information. Specifically, we would expect the following shift.

$$\underbrace{\left[\Pr\left(y_i > 0, g\right) - \widetilde{\Pr}\left(y_i > 0\right), g\right]}_{(a)\text{Incorrect probability}} \times \underbrace{\left[\widetilde{E}\left(y_i \mid y_i > 0\right) - \widetilde{E}\left(y_i \mid y_i \leq 0\right)\right]}_{(b)\text{Perceived informativeness}} \tag{A.9}$$

The first term is the difference between the actual probability of this gender passing the test, compared to the the true probability of passing the test. If people thought women were unlikely to pass, but they were actually more likely to pass, the average belief increases upon test observance. The second term is the perceived informativeness of the test, as measured by the expected difference in productivity for people of this gender who pass it, compared to those who do not.

If perceived informativeness is the same across genders, then the impact of information on the gender gap is proportional to the following difference in differences.

$$[\Pr(y_i > 0, m) - \Pr(y_i > 0, f)] - [\widetilde{\Pr}(y_i > 0, m) - \widetilde{\Pr}(y_i > 0, f)] \tag{A.10}$$

This result is very similar to our results with our baseline model in Section 2. Specifically both equation 5 and equation A.10 show that the expected change in beliefs is proportional to the difference between the true and (prior) perveived gender gaps in performance as measured by the signal. Thus, if the evaluator is too pessimistic about female coders releative to men, then the term being subtracted is large relative to the term added on the left, and the whole expression is negative. The gender gap shrinks on average with the new information provided by the unit tests.

Appendix D Experiment I: Additional Results

D.1 Explaining a Persistent Gender Gap

Our results indicate that gender gaps did not decrease with more information. While this may be due to statistical chance, it suggests that evaluators may be unduly pessimistic about men relative to women. Experiment I could not shed more direct light on prior beliefs, but we later collected information about beliefs in Experiment II. As we discuss in Section 4, we do find evidence that is consistent with evaluators discounting slightly the performance of men relative to women, compared to the true gender gap in performance as measured by the unit tests.

We can also evaluate other possibilities, one of which is that the unit tests were more informative for men than women.[A.5] To see why this could conceivably explain our results, consider an extension of the model in Section 2. Rather than the weight

[A.5] Beyond these two explanations, the differential impact could be due to a non-linear mapping between beliefs and ratings, or to statistical chance.

on the signal being the same for men and women ($s_m = s_f$), let the signal be more informative for one gender. In this case, the gender gap given signal realization θ_i is:

$$\text{Gender Gap} \mid \theta_i = \overbrace{s_m \mu_m^* + (1 - s_m)\, \mu_m}^{\textbf{Male Belief}} - \overbrace{\left[s_f \mu_f^* + \left(1 - s_f\right) \mu_f \right]}^{\textbf{Female Belief}} \tag{A.11}$$

where $s_g = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_\varepsilon^2} \in (0, 1)$ is the weight placed on the signal for gender $g \in \{m, f\}$. The impact of more information on the gender gap is then:

$$d\text{Gap} = ds_m \underbrace{\left(\mu_m^* - \mu_m\right)}_{\textbf{Male Pessimism}} - ds_f \underbrace{\left(\mu_f^* - \mu_f\right)}_{\textbf{Female Pessimism}} \tag{A.12}$$

where ds_g is the marginal impact of information on s_g.

This highlights the two reasons why the gender gap could persist with more information. First, $\mu_m^* - \mu_m$ may larger than $\mu_f^* - \mu_f$, which would imply that evaluators are unduly pessimistic about men compared to women, relative to the true performance.

Second, the impact on the signal may be larger for men than for women, ($ds_m > ds_f$). This could occur for example if men are assigned problems which are more informative. However Table D2 shows that men and women face similar problems. This is true in terms of difficulty, as measured by average performance of others on those problems. It is also true for problems with different cross-sectional variances in performance, which could indicate that some tests are more discerning than others.

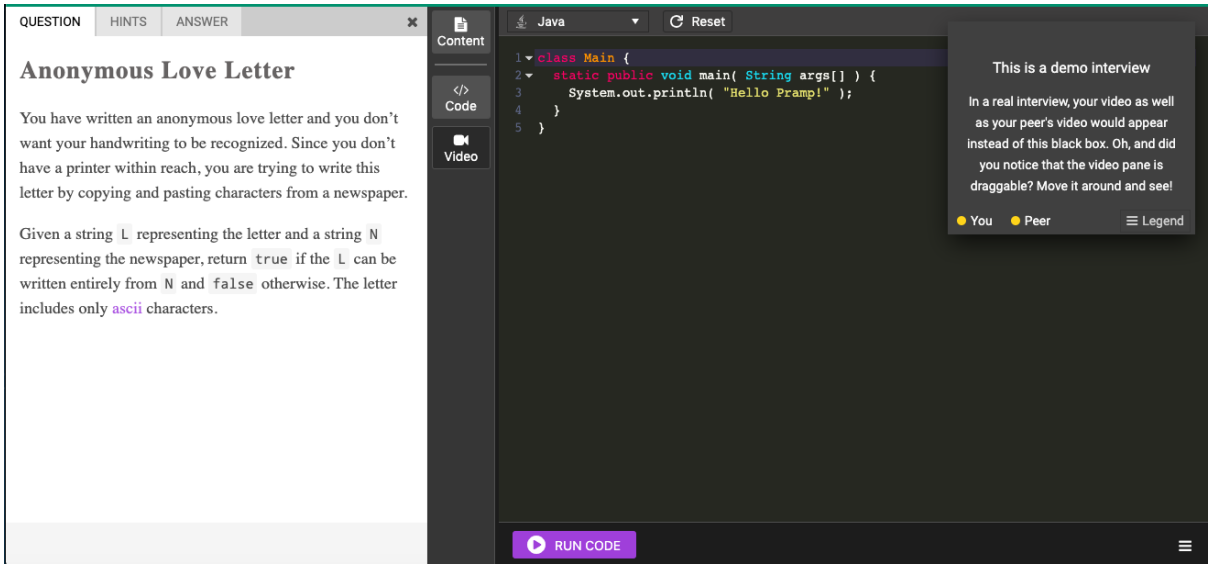
D.2 Complier Characteristics.

We show observable characteristics of compliers in Table D4.[A.6] Characteristics are similar between treated and untreated compliers. Column (5) presents characteristics for never-takers. The comparisons in Table D4 reveal that the representation of most subgroups among compliers is similar to the overall sample, although compliers do have slightly less experience. However, the gender gap in activation translates into under-representation of women among the compliers.

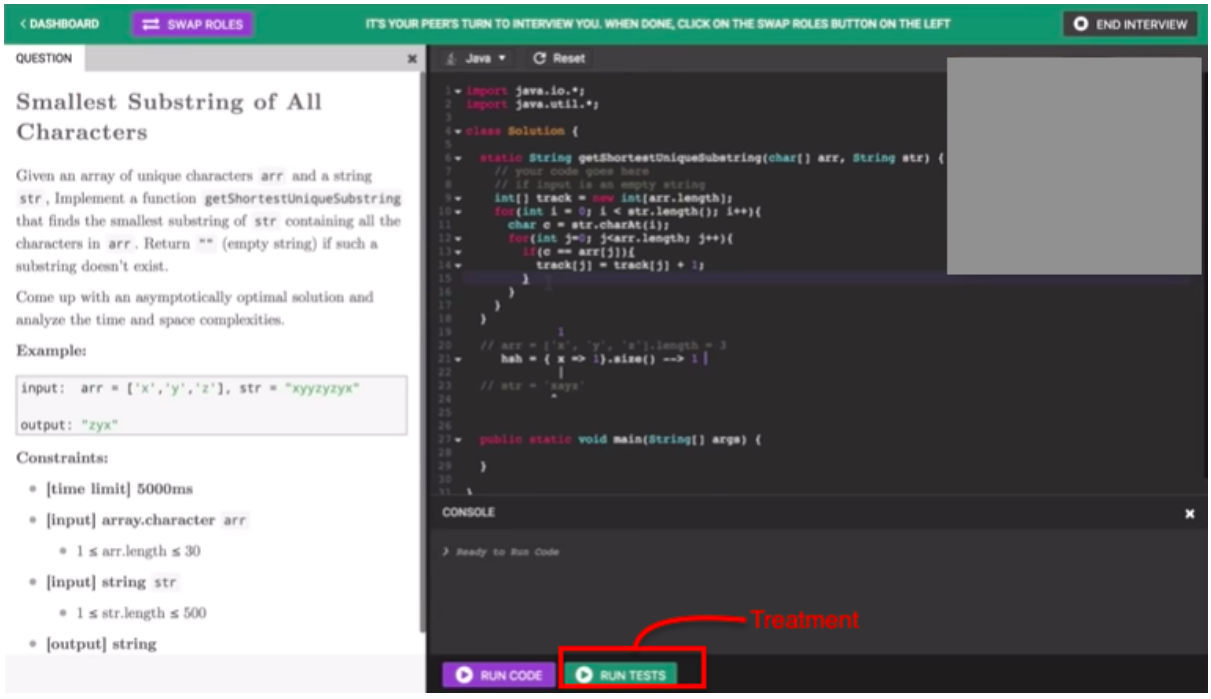
[A.6] Following Abadie (2003), these characteristics are recovered by calculating the fraction of compliers in different subsamples. The results come an IV procedure where the dependent variable is $X_i D_i$ (Column 4) and $X_i(1 - D_i)$, using T_i as an instrument for D_i.

D.3 Additional Figures and Tables

Figure D1: Environment of the Platform (Treatment vs. Control)



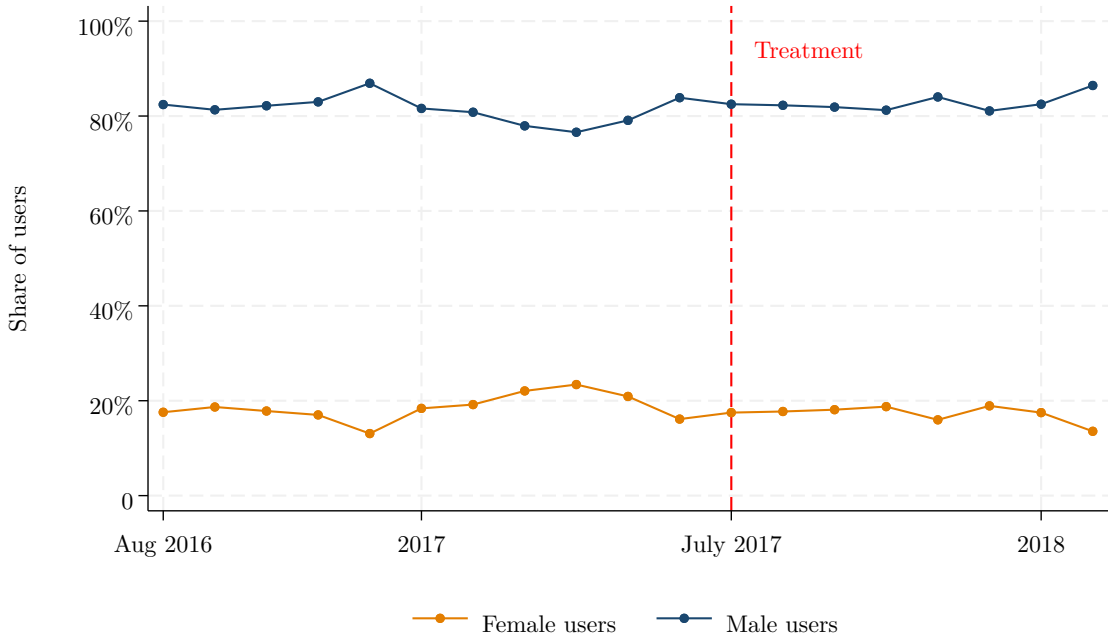
(a) Control



(b) Treatment

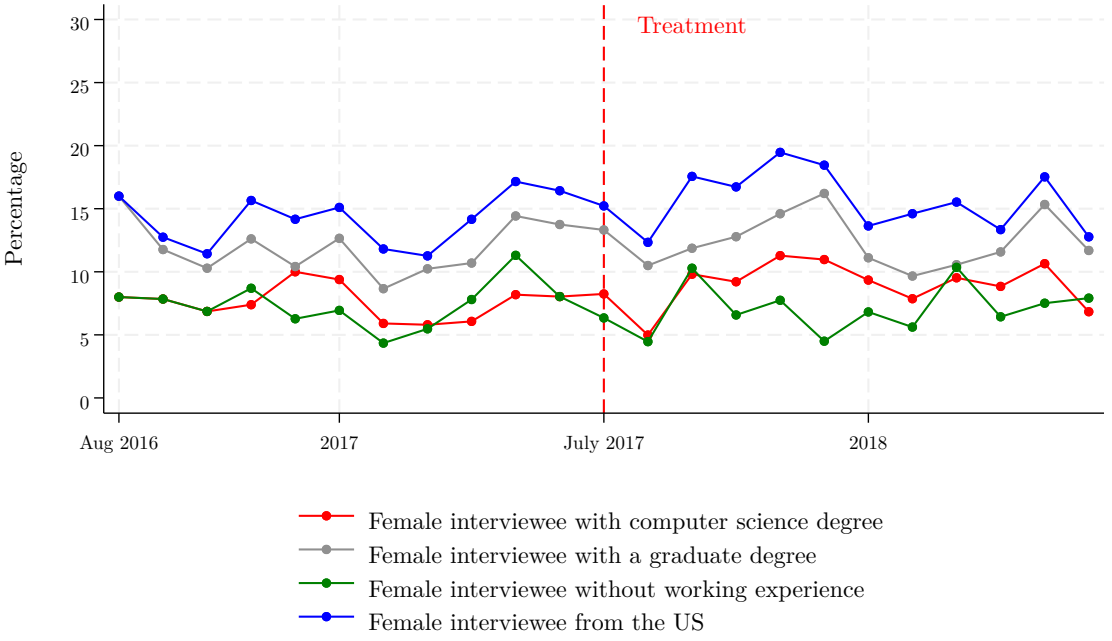
Notes: This figure shows the platform layout for a mock interview. Panel (a) shows the control condition, where the code can be run but there are no build in "unit tests" to verify code quality. Panel (b) shows the treatment condition, in which a button is added to run the diagnostic tests.

Figure D2: Share of male and female users over time



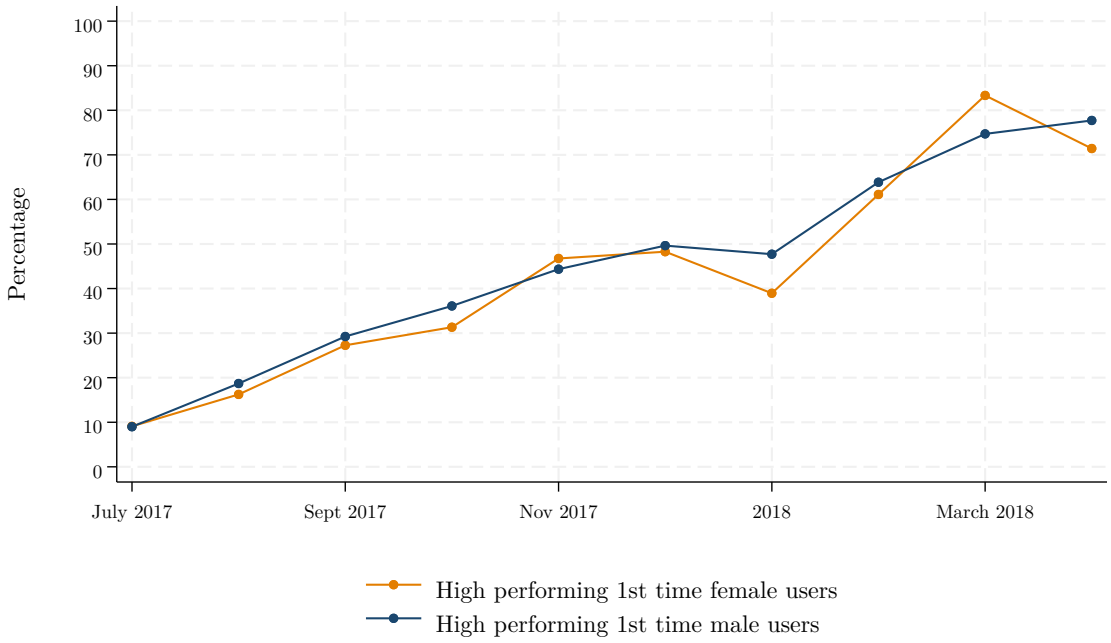
Notes: This figure shows the evolution of the shares of female and male users on the platform before and after the unit tests began to be introduced. The vertical red line shows when the introduction started.

Figure D3: Evolution of First-Time Female Users' Characteristics



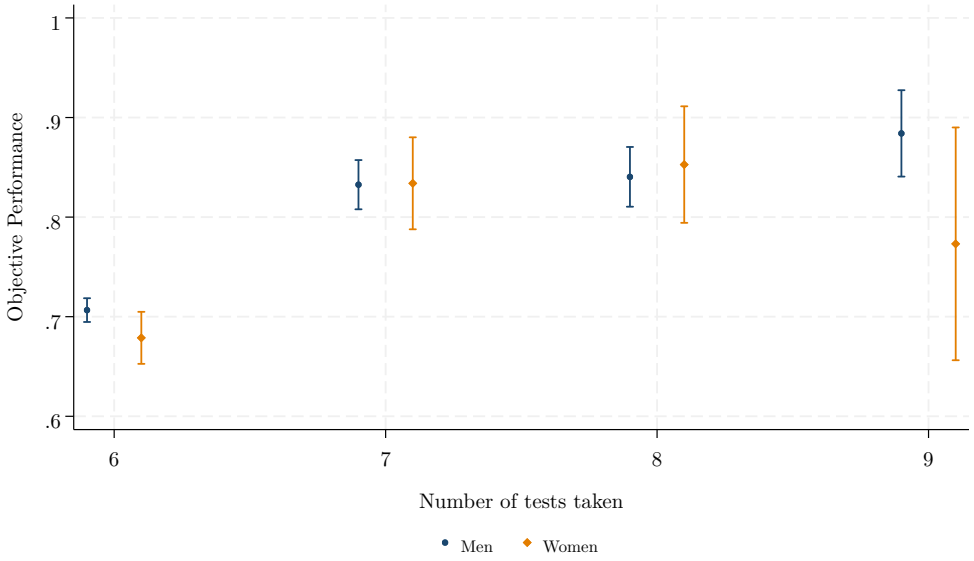
Notes: The figure presents the evolution of first-time female users' characteristics averaged by month around the date that the unit tests began to be introduced. The vertical red line shows when the introduction started.

Figure D4: Share of High-Performing First-Time Female and Male Users



Notes: The figure presents the evolution of the share of high-performing first-time female and male users by month after the unit tests began to be introduced. High-performing users are defined as those passing all unit tests taken for a given problem.

Figure D5: 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. The sample includes all platform users who activated the objective performance measure from July 2017 to April 2018.

Table D1: Robustness Checks for Experiment I

	Coding	Problem solving	Likeability	Communication	Hireability
Panel A: Baseline					
Treatment	0.166***	0.222***	0.099**	0.197***	0.178***
S.E	0.032	0.032	0.039	0.044	0.033
Treatment*Woman	-0.099	-0.056	-0.074	0.006	-0.045
S.E	0.066	0.061	0.084	0.069	0.076
N	11029	11029	11029	11029	11049
Panel B: with Month FE					
Treatment	0.140***	0.212***	0.079**	0.161***	0.150***
S.E	0.029	0.029	0.036	0.042	0.030
Treatment*Woman	-0.109*	-0.067	-0.066	0.013	-0.044
S.E	0.064	0.059	0.082	0.067	0.074
N	11029	11029	11029	11029	11049
Panel C: with Controls					
Treatment	0.168***	0.226***	0.104***	0.199***	0.180***
S.E	0.032	0.032	0.038	0.044	0.033
Treatment*Woman	-0.093	-0.061	-0.074	0.003	-0.044
S.E	0.066	0.060	0.084	0.070	0.076
N	11029	11029	11029	11029	11049
Panel D: no Date FE					
Treatment	0.160***	0.221***	0.100***	0.167***	0.149***
S.E	0.028	0.028	0.033	0.041	0.029
Treatment*Woman	-0.106	-0.066	-0.067	0.014	-0.044
S.E	0.064	0.059	0.082	0.067	0.074
N	11029	11029	11029	11029	11049
Panel E: Including pre-treatment period					
Treatment	0.146***	0.213***	0.082**	0.197***	0.162***
S.E	0.031	0.031	0.034	0.040	0.028
Treatment*Woman	0.011	-0.009	0.025	0.007	0.041*
S.E	0.023	0.024	0.023	0.021	0.024
N	54077	54077	54077	54077	51533
Panel F: Controlling for Propensity Score Matching					
Treatment	0.165***	0.221***	0.099**	0.195***	0.177***
S.E	0.032	0.033	0.039	0.044	0.033
Treatment*Woman	-0.099	-0.055	-0.073	0.008	-0.045
S.E	0.066	0.061	0.084	0.068	0.076
N	11029	11029	11029	11029	11049
Panel G: with Individual FE					
Treatment	-0.005	0.082**	0.028	0.079*	0.060
S.E	0.036	0.033	0.044	0.047	0.037
Treatment*Woman	-0.031	-0.026	-0.169*	0.023	-0.036
S.E	0.092	0.090	0.097	0.111	0.093
N	9797	9797	9797	9797	9816

Notes: This table shows results a series of robustness checks using the whole sample of platform users between July 8, 2017 when the automated coding measure was first introduced, and October 27, 2017 when it was generalized to all users. 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 control for propensity score matching. In Panel G, we control for interviewee fixed effects. Standard errors are clustered at the date level.

Table D2: Problems' and Evaluators' Characteristics

	Problem Difficulty	Variation of the Performance	Harsh Evaluator	
	(1)	(2)	(3)	(4)
Interviewee female	-0.003	0.006	0.005	0.005
	(0.008)	(0.008)	(0.010)	(0.010)
Interviewer Gender	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes
Problem FE	No	No	No	Yes
N	26,667	26,667	22,582	19,635

Notes: This table shows the coefficient on gender from regressions with dependent variables of problem difficulty, within-problem variation in performance, and whether or not the evaluator was historically harsh as measured by whether the ratings they chose in the past were lower than the median.

Table D3: Balancing Test – Whole Sample

Variables	Control	ITT	Difference	P-value
Interviewee female	0.179	0.187	0.007	0.549
Interviewer female	0.178	0.187	0.008	0.504
Gender interviewer missing	0.049	0.048	-0.001	0.873
Country: USA	0.686	0.684	-0.002	0.923
Interviewee's deg.: computer science	0.645	0.653	0.008	0.635
Interviewer's deg.: computer science	0.643	0.653	0.009	0.578
Interviewer's deg.: postgraduate	0.437	0.431	-0.006	0.700
Interviewee's deg.: postgraduate	0.441	0.430	-0.012	0.498
Interviewee's years of experience	2.943	3.087	0.144	0.224
Interviewer's years of experience	2.958	3.090	0.132	0.271
N	1,587	10,004		
Test of joint significance	*F*-stat: 1.100 (*p*-value: 0.377)			

Notes: This table shows descriptive statistics for the control and ITT samples for Experiment I (see Section 3), along with p-values which test whether differences are significant.

Table D4: Baseline Characteristics of Compliers and Never-Takers

	First Stage	**Sample mean**	**Compliers**		**Never-takers**
	(1)	(2)	(3)	(4)	(5)
			Treated	Untreated	
Interviewee female	0.678***	0.186	0.177	0.166	0.212
	(0.015)		(0.007)	(0.016)	(0.008)
Country: USA	0.718***	0.684	0.681	0.684	0.693
	(0.010)		(0.008)	(0.021)	(0.010)
Interviewee's deg.: computer science	0.709***	0.652	0.660	0.649	0.663
	(0.011)		(0.008)	(0.021)	(0.009)
Interviewee's deg.: postgraduate	0.726***	0.431	0.434	0.450	0.424
	(0.011)		(0.008)	(0.021)	(0.009)
Interviewee's years of experience	0.736***	3.067	3.061	2.859	3.225
	(0.021)		(0.045)	(0.159)	(0.062)
Interviewee Preparation Level (self-declared on 1-5 scale)	0.621***	2.880	2.928	2.768	2.816
	(0.049)		(0.013)	(0.034)	(0.017)

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 D5: Gender Gap in Coding Ratings and Interviewer's Experience

	Subjective Coding Ratings				
	(1)	(2)	(3)	(4)	(5)
Interviewee female	-0.081***	-0.081***	-0.084***	-0.076***	-0.088***
	(0.018)	(0.018)	(0.021)	(0.021)	(0.000)
Interviewer's total # of sessions	Yes				
Interviewer's # of past sessions		Yes			
Interviewer's total # of female interviewees			Yes		
Past top female performer				Yes	
Interviewer's work experience $\geq$ 2 years					Yes
Objective performance	Yes	Yes	Yes	Yes	Yes
Interviewer gender	Yes	Yes	Yes	Yes	Yes
Interviewee's sociodemographic controls	Yes	Yes	Yes	Yes	Yes
Interviewer's sociodemographic controls	Yes	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes	Yes
Problem FE	Yes	Yes	Yes	Yes	Yes
Observations	19,551	19,551	14,677	13,541	18,138

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 E Experiment II: Additional Results

E.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' undergraduate and graduate programs. Our recruitment email disclosed that we were studying how evaluators judge the performance of software developers, but did not mention gender.

Testing the salience of treatment In the piloting phase of the experiment, we asked a random sample of online participants ("evaluators") 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 some "evaluators" did not 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 evaluation tasks. We asked them to predict the performance of three different hypothetical coders. We cross-randomized the first name (alternating gender) and the skill level for each vignette. The vignetted are constructed as follows:

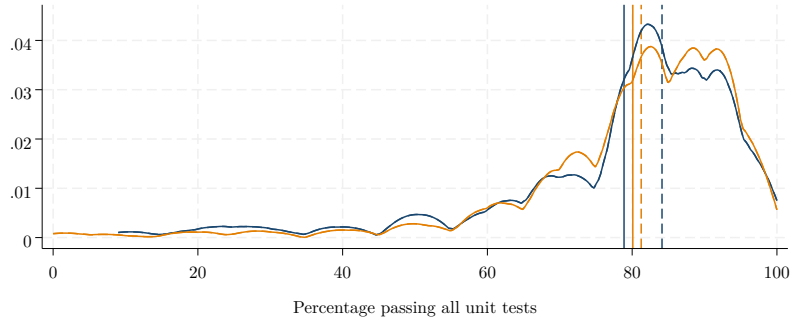
> *82% of the codes you will potentially see resulted in a perfect score and passed all the unit tests. We ask your opinion about the potential performance of different hypothetical coders. If your guess is within 5% of the truth, we will send you an additional reward!*
>
> *"[First Name] holds [Skills]. According to you, what is the percent chance that [First Name]'s code passed all the unit tests?"*

The names and skills shown in the vignettes are as follows.

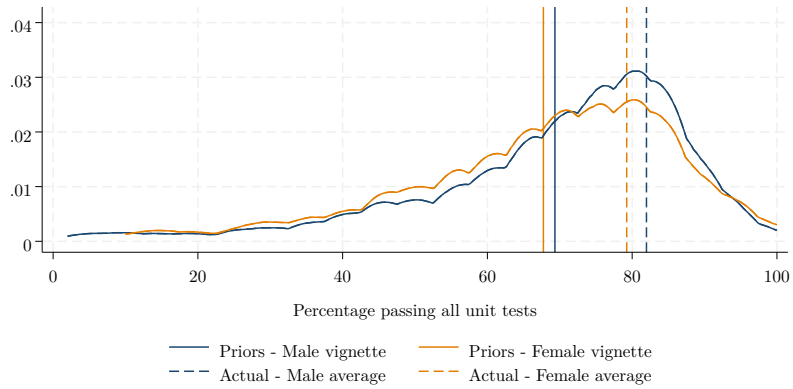
Skills	First names
a M.Sc in computer science and has 2 years of work experience	Katie/Tom
a Ph.D. in mathematics and has no industry experience	Alexa/Mickael
a B.Sc. degree in computer science	Corinne/Matt

Our results regarding prior beliefs using the resulting data are discussed briefly in Section 4. The accompanying figures follow below.

Figure E1: Respondents' Priors Beliefs About Performance by Gender



(a) M.Sc in computer science
2 years of work experience



(b) B.Sc. degree in computer science

Notes: This figure shows the distributions of respondents' prior beliefs by gender and skill level of the vignette. The continuous lines represent the mean prior for each gender. The dash lines represent the actual performance for each gender calculated from the sample of codes from the experimental sample. In the overall sample of codes, 82 percent of users pass all unit tests.

(a) Interface-Non-Blind Male

(b) Interface- Blind Male

Figure E2: Exp II Code block examples written by a male coder as presented in non-blind and blind conditions.

(a) Interface- Non-Blind Female

(b) Interface-Blind Female

Figure E3: Exp II Code block examples written by a female coder as presented in non-blind and blind conditions.

E.2 Descriptive Statistics: Sample of Code Blocks

Table E1: Descriptive Statistics — Follow-up Experiment— January 2018-May 2022

	Raw Data	Clean Data	Experimental Data
Number of session-participant pairs	482,390	178,717	38,322
Number of unique participants	97,614	30,633	10,380
Number of unique problems	39	39	38
Share non-missing unit score	0.42	0.56	1.00
Share of Python scripts	0.30	0.37	0.43
Share of Java scripts	0.35	0.35	0.45
Share of C++ scripts	0.17	0.09	0.12
Share Female			0.18
Share Nonwhite			0.62
Share Full Score			0.82

Notes: This table presents basic characteristics for the code blocks in the sample used in Experiment II (see Sections 1.2 and 4). 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 the United States.

Table E2: Descriptive Statistics — Coding Blocks

	Mean	Std. Dev.
Female Users	0.500	0.501
Objective score	0.744	0.314
Passed all unit tests	0.500	0.501
Subjective Rating	3.379	0.713
Num. lines	47.14	13.70
C++	0.088	0.283
Java	0.544	0.499
Python	0.368	0.483
Master degree or more	0.520	0.500
Major in CS	0.827	0.379
Years of FT work experience	3.055	3.143
N	456	

Notes: This table provides summary statistics for the final set of code blocks on which Exp II was conducted.

E.3 Descriptive Statistics: Evaluators

Table E3: Descriptive Statistics — Participants

	Mean	Std. Dev.	N
Gender			
Female	0.278	0.448	565
Male	0.658	0.475	565
Non-binary / third gender	0.03	0.171	565
Prefer not to say	0.03	0.171	565
Prefer to self-describe	0.004	0.059	565
Recoded race			
White	0.164	0.371	603
South Asian	0.216	0.412	603
Chinese	0.526	0.5	603
Black	0.005	0.07	603
Latinx	0.018	0.134	603
Other	0.071	0.258	603
Unknown	0.158	0.365	716
Current situation			
Currently a student	0.828	0.377	705
Completed at least one degree	0.166	0.372	705
Didn't complete a degree	0.006	0.075	705
Highest degree completed			
Associates or technical degree	0.004	0.065	704
Bachelor's degree	0.736	0.441	704
High School diploma or GED	0.021	0.145	704
MA, MSc or MEng	0.151	0.358	704
PhD	0.047	0.212	704
Some college, but no degree	0.034	0.182	704
Prefer not to say	0.007	0.084	704
Experience with Python			
Basic	0.221	0.415	707
Intermediate	0.448	0.498	707
Advanced	0.331	0.471	707
Experience with Java			
Basic	0.536	0.499	676
Intermediate	0.361	0.481	676
Advanced	0.104	0.305	676
Experience with C++			
Basic	0.643	0.479	673
Intermediate	0.272	0.445	673
Advanced	0.085	0.279	673
Preferred language			
C++	0.089	0.285	716
Java	0.141	0.348	716
Python	0.77	0.421	716

Notes: This table shows descriptive statistics participants in Experiment II (see Section 4).

Table E4: Treatment-Control Balance — Whole Sample

	Non-blind to Blind (1)	Blind to Non-blind (2)	Difference (3)	p-value of diff. (4)
Female	0.278	0.278	-0.000	0.992
Male	0.662	0.655	-0.008	0.850
White respondent	0.158	0.170	0.011	0.714
South Asian	0.205	0.227	0.022	0.510
Chinese	0.554	0.497	-0.057	0.161
Black	0.007	0.003	-0.003	0.569
Latinx	0.020	0.017	-0.003	0.776
Other	0.056	0.087	0.030	0.149
Unknown	0.146	0.169	0.024	0.387
Currently a student	0.827	0.830	0.003	0.927
Completed at least one degree	0.164	0.168	0.003	0.908
Didn't complete a degree	0.008	0.003	-0.006	0.303
Bachelor's degree	0.708	0.764	0.056	0.090
MA, MSc or MEng	0.170	0.131	-0.039	0.144
PhD	0.059	0.034	-0.025	0.115
C++	0.082	0.097	0.015	0.479
Java	0.161	0.122	-0.039	0.137
Python	0.758	0.781	0.024	0.455
Observations	1,420	1,444		

Notes: This table presents balancing checks for the whole sample. The p-values are obtained from a linear regression on each covariate with strata fixed effect. Standard errors are clustered at the evaluator level.

Table E5: Blinding Experiment — Main Results (Reweighted)

	Coding subjective rating		Unit tests prediction		Interview prediction	
	(1)	(2)	(3)	(4)	(5)	(6)
Female code	0.090	0.089	0.311	0.312	0.088	0.087
	(0.074)	(0.070)	(0.217)	(0.213)	(0.065)	(0.061)
Non-blind code	-0.040	-0.034	-0.338	-0.307	-0.121*	-0.044
	(0.071)	(0.070)	(0.241)	(0.242)	(0.066)	(0.066)
Non-blind code × Female code	0.006	-0.001	0.237	0.211	0.000	-0.004
	(0.100)	(0.099)	(0.319)	(0.318)	(0.089)	(0.088)
Treatment order control	Yes	Yes	Yes	Yes	Yes	Yes
Order of scripts FE	Yes	Yes	Yes	Yes	Yes	Yes
Problem FE	Yes	Yes	Yes	Yes	Yes	Yes
Evaluator FE	No	Yes	No	Yes	No	Yes
Observations	2,314	2,284	2,314	2,284	2,664	2,664

Notes: This table provides results from Experiment II (see Section 4), testing the pre-registered hypothesis that revealing gender introduces a gender gap that penalizes women. The regression specification is as described in Equation (3).The even columns include evaluator fixed effects, while the odd columns do not. Standard errors are clustered at the evaluator level. Results are weighted by gender and education composition of users on the platform. Weights are equal the inverse predicted probability of being in the experiment relative to the Platform.

Table E6: Current Students' Gender Gap

	Subjective Coding Ratings			
	Overall		Students Only	
Female	-0.094***	-0.106***	-0.142***	-0.131***
	(0.007)	(0.008)	(0.023)	(0.046)
Objective Score	Yes	Yes	Yes	Yes
Problem FE	Yes	Yes	Yes	Yes
Evaluator FE	No	Yes	No	Yes
Observations	89,716	75,681	8,758	3,382

Notes: This table provides results for the gender gap in subjective ratings in the overall online platform sample in columns (1)-(2), and the gender gap amongst current students only in columns (3)-(4), for the sample from 2018 to 2022. Current students are those who are studying towards a Bachelor degree and have zero years of full-time experience at the time of using the platform.

Table E7: Effect Of Blinding On Gender Gaps — Quality Sample

	Subjective coding rating		Unit test prediction		Interview prediction	
Female code	-0.030	-0.022	0.072	0.081	0.022	0.024
	(0.066)	(0.065)	(0.207)	(0.207)	(0.060)	(0.059)
Non-blind code	-0.120	-0.116	-0.364	-0.357	-0.112	-0.073
	(0.066)	(0.067)	(0.219)	(0.220)	(0.062)	(0.062)
Non-blind code×Female code	0.105	0.107	0.290	0.335	0.072	0.073
	(0.094)	(0.095)	(0.299)	(0.299)	(0.086)	(0.086)
Treatment order control	Yes	Yes	Yes	Yes	Yes	Yes
Order of scripts FE	Yes	Yes	Yes	Yes	Yes	Yes
Problem FE	Yes	Yes	Yes	Yes	Yes	Yes
Evaluator FE	No	Yes	No	Yes	No	Yes
Observations	1,852	1,835	1,852	1,835	1,946	1,946

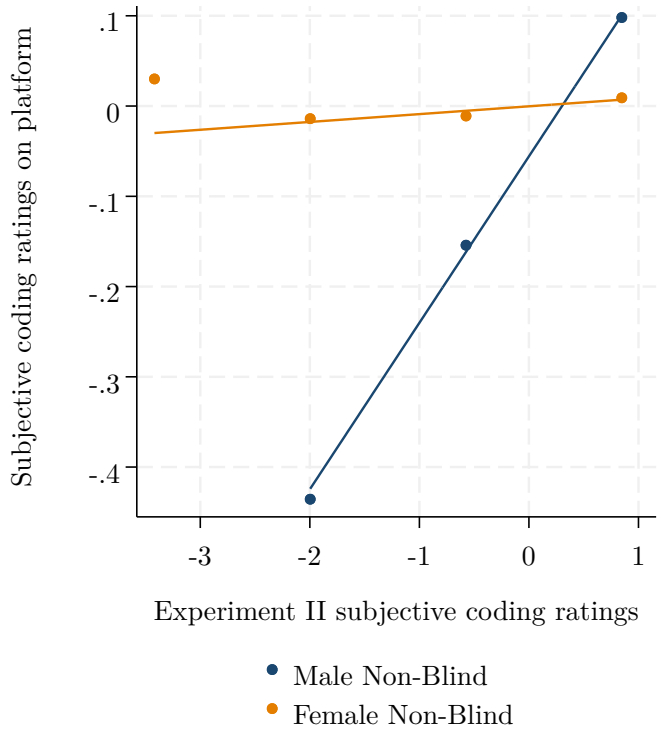
Notes: This table provides results from Experiment II (see Section 4), testing the pre-registered hypothesis that revealing gender introduces a gender gap that penalizes women for the quality sample, for the quality sample, namely restricting to participants who passed the first attention check question, and excluding respondents whose survey completion time falls within the bottom 10th (less than 8 minutes) and top 90th percentiles (4 hours or more). The regression specification is as described in Equation (3). The dependent variables are the (standardized) subjective coding ratings (columns 1-2), participants' prediction of the unit tests passed by the code script (columns 3-4) and their prediction of the coder's probability of passing the interview (columns 5-6). The even columns include evaluator fixed effects. Standard errors are clustered at the evaluator level.

Table E8: Blinding Experiment — Excluding Primed Participants

	Subjective coding rating		Unit test prediction		Interview prediction	
	(1)	(2)	(3)	(4)	(5)	(6)
Female code	0.006	-0.000	0.139	0.149	0.023	0.023
	(0.061)	(0.061)	(0.186)	(0.188)	(0.052)	(0.052)
Non-blind code	-0.111	-0.116	-0.321	-0.310	-0.173**	-0.068
	(0.060)	(0.061)	(0.194)	(0.196)	(0.053)	(0.053)
Non-blind code×Female code	0.070	0.083	0.276	0.289	0.046	0.038
	(0.086)	(0.087)	(0.266)	(0.269)	(0.072)	(0.072)
Treatment order control	Yes	Yes	Yes	Yes	Yes	Yes
Order of scripts FE	Yes	Yes	Yes	Yes	Yes	Yes
Problem FE	Yes	Yes	Yes	Yes	Yes	Yes
Evaluator FE	No	Yes	No	Yes	No	Yes
Observations	2,183	2,152	2,183	2,152	2,564	2,564

Notes: This table provides results from Experiment II (see Section 4), testing the pre-registered hypothesis that revealing gender introduces a gender gap that penalizes women on the sample of code scripts seen first. In this analysis, we restrict to the first script evaluated by each participant. The regression specification is as described in Equation (3). The dependent variables are the (standardized) subjective coding ratings (column 1), participants' prediction of the unit tests passed by the code script (column 2) and their prediction of the coder's probability of passing the interview (column 3). Standard errors are clustered at the evaluator level, and shown in parentheses.

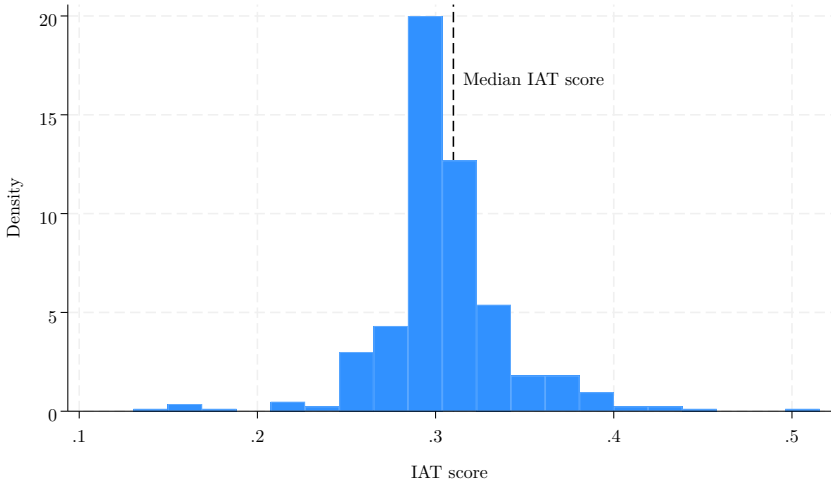
Figure E4: Comparability of Coding Ratings between Experiments I and II



Notes: This figure shows the average relationship between ratings in Experiment II in the non-blind treatment, and ratings on the platform for the same code block, for men and women.

Appendix F Implicit Bias Results

Figure F1: Distribution of IAT Scores



Notes: This figure presents the distribution of IAT scores of evaluators' metropolitan statistical areas (MSA) of graduation in our sample described in Section 1.7. The dash line indicates the US median.

Table F1: Gender Gap By Evaluator IAT

	Subjective Coding Ratings		
	Low Bias	High Bias	All
Female	-0.085***	-0.151***	-0.082***
	(0.027)	(0.050)	(0.027)
Female x High IAT dummy			-0.083*
			(0.050)
High Score	0.475***	0.579***	0.498***
	0.028	(0.053)	(0.025)
Observations	5,730	1,672	7,402

Notes: This tables shows the gender gap in (standardized) subjective ratings for two groups. Column (1) presents the gender gap when evaluators graduated from a higher education institution located in an MSA with below-median IAT score (i.e less prejudice against women in science). Column (2) presents results when evaluators graduated from a higher education institution located in an MSA with above-median IAT score (i.e more prejudice against women in science). Column (3) tests for statistical differences in the gender gap between both groups. Objective score is controlled for in all specifications. Evaluators' institutions are obtained from LinkedIn data as described in Section 1.7. IAT scores are from the Gender-Science IAT module for the years 2018 and 2019 of the Harvard Implicit Project.

F.1 Closing the Gender Gap

To gauge the importance of the gender gaps we see, we provide a back-of-the-envelope calculation of the impact on future job market outcomes of closing them. To do this, we estimate the relationship between subjective ratings on employment at a top tech company. We then combine this relationship with the size of the gender gap we see to estimate the impact that closing the skill assessment gaps. We note that this rough

calculation requires the strong assumption that the cross-section relationship between subjective ratings and employment ratings is a good approximation of the causal impact of receiving better ratings in these interviews, or similar ones that candidates later encounter when they apply for these jobs.

Our first step to run a linear regression on the platform sample matched with Revelio data, in which the dependent variable is a dummy variable equal to one if the individual has ever been employed in the big six company within two years after obtaining their first CS related degree, and the independent variable is the standardized subjective rating. We restrict to a men to avoid comparing outcomes of men and women, which may be subject to bias at later stages. We control for graduation-year fixed effects, whether the individual has a master degree, and their student status when using the platform. The regression is weighted by the number of sessions on the platform to account for multiple ratings per platform user.

The estimates from this regression suggest that a one-standard deviation increase in subjective ratings is associated with a 5.2 percentage-point increase in the probability of being employed in these firms within two years of and indicidual obtaining their first computer science related degree. We multiply this by the gender gap in subjective ratings (0.12) and divide by baseline share of women employed in software engineering positions in these companies two years after graduation (27 percent): $(0.12 * 5.2)/27 = 2.3$. This indicates that closing the gap in subjective ratings would increase female employment at these top firms by 2.3 percent.

Appendix G Video Analysis

G.1 Gemini's Prompts

Prompt Candidate *This video features an interviewer and an interviewee. Identify the interviewee, who is being asked to complete a task. Evaluate that interviewee and provide a rating on each of the following dimensions as a decimal.*

1. Confidence (higher is more confident).
2. Upticks in pitch at the end of sentences (higher is more upticks).
3. Hesitation before answering (higher is more hesitation).
4. Propensity to apologize for things (higher is more).
5. Share of speech by interviewee rather than the interviewer (higher is more interviewee speech).
6. Friendliness (higher is more friendly).
7. Propensity to make errors (higher is more errors).

8. Propensity to ask clarifying questions (higher is more clarifying questions).

Report these scores, separated by semicolons. For example, a valid response could be: 8.32;0.28;8.88;9.32;3.45;8.87;5.87,9.51;9.77;5.44
Do not report any other text, or formatting. Just numbers in that form.

Prompt Interviewer *This video features an interviewer and an interviewee. Identify the interviewee, who is being asked to complete a task. Evaluate that interviewer and provide a rating on each of the following dimensions as a decimal.*

1. How often does the interviewer interrupt the interviewee? (Higher is more often)
2. Does the interviewer actively listen to the interviewee? (Higher is more active listening)
3. Is the interviewer impatient? (Higher is more impatient)
4. Is the interviewer respectful? (Higher is more respectful)
5. Is the interviewer condescending? (Higher is more respectful)
6. Is the interviewer harsh or excessively critical of the interviewee? (Higher is harsher / more critical)
7. Does the interviewer explain the problem clearly? (Higher is more clearly explained)
8. Does the interviewer build effective rapport with the interviewee? (Higher is more effective rapport)

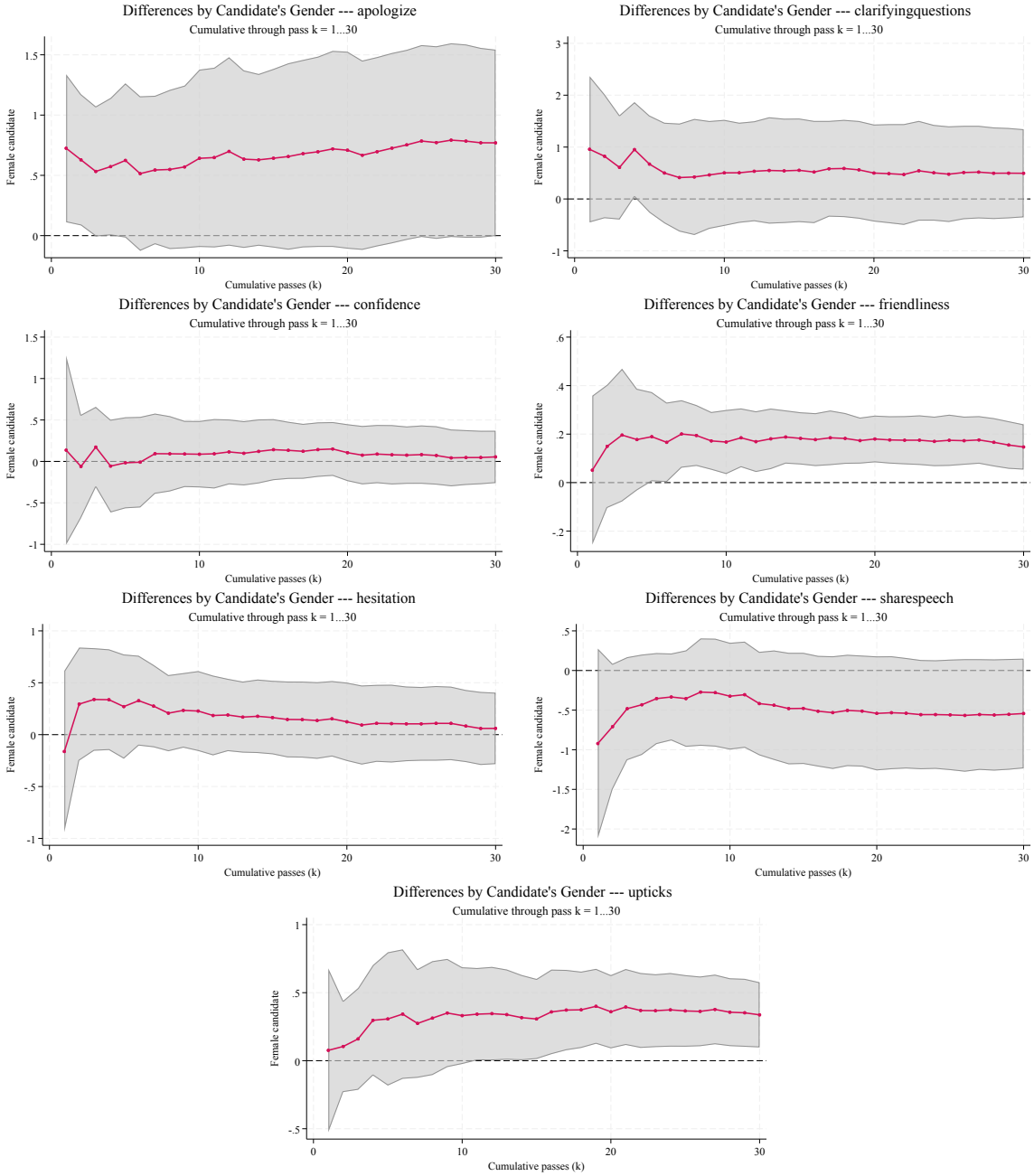
Report these scores, separated by semicolons. For example, a valid response could be: 8.32;0.28;8.88;9.32;3.45;8.87;5.87,9.51;9.77;5.44
Do not report any other text, or formatting. Just numbers in that form.

Table G1: Summary Statistics — Video Analysis

Variable	Mean	Std. Dev.	Min.	Max.	N
Panel A. URL Level					
Female Candidate	0.065	0.246	0	1	186
Subjective Coding Rating	0	1	-3.327	0.944	187
Panel B. Candidate Analysis from Gemini					
Apologize	0	1	-0.955	5.744	5,562
Clarifying Question	0	1	-1.824	2.129	5,562
Confidence	0	1	-4.448	2.712	5,562
Friendliness	0	1	-4.523	3.695	5,562
Hesitation	0	1	-2.745	3.725	5,562
Share Speech	0	1	-2.733	2.077	5,562
Upticks	0	1	-1.019	7.289	5,562
Panel C. Interviewer Analysis from Gemini					
Condescending	0	1	-1.102	9.718	5,580
Explain	0	1	-6.242	1.363	5,580
Harsh	0	1	-0.86	7.183	5,580
Impatient	0	1	-1.166	5.683	5,580
Interrupt	0	1	-0.805	3.62	5,580
Listen	0	1	-4.464	1.202	5,580
Rapport	0	1	-5.562	1.512	5,579
Respect	0	1	-6.613	0.906	5,580

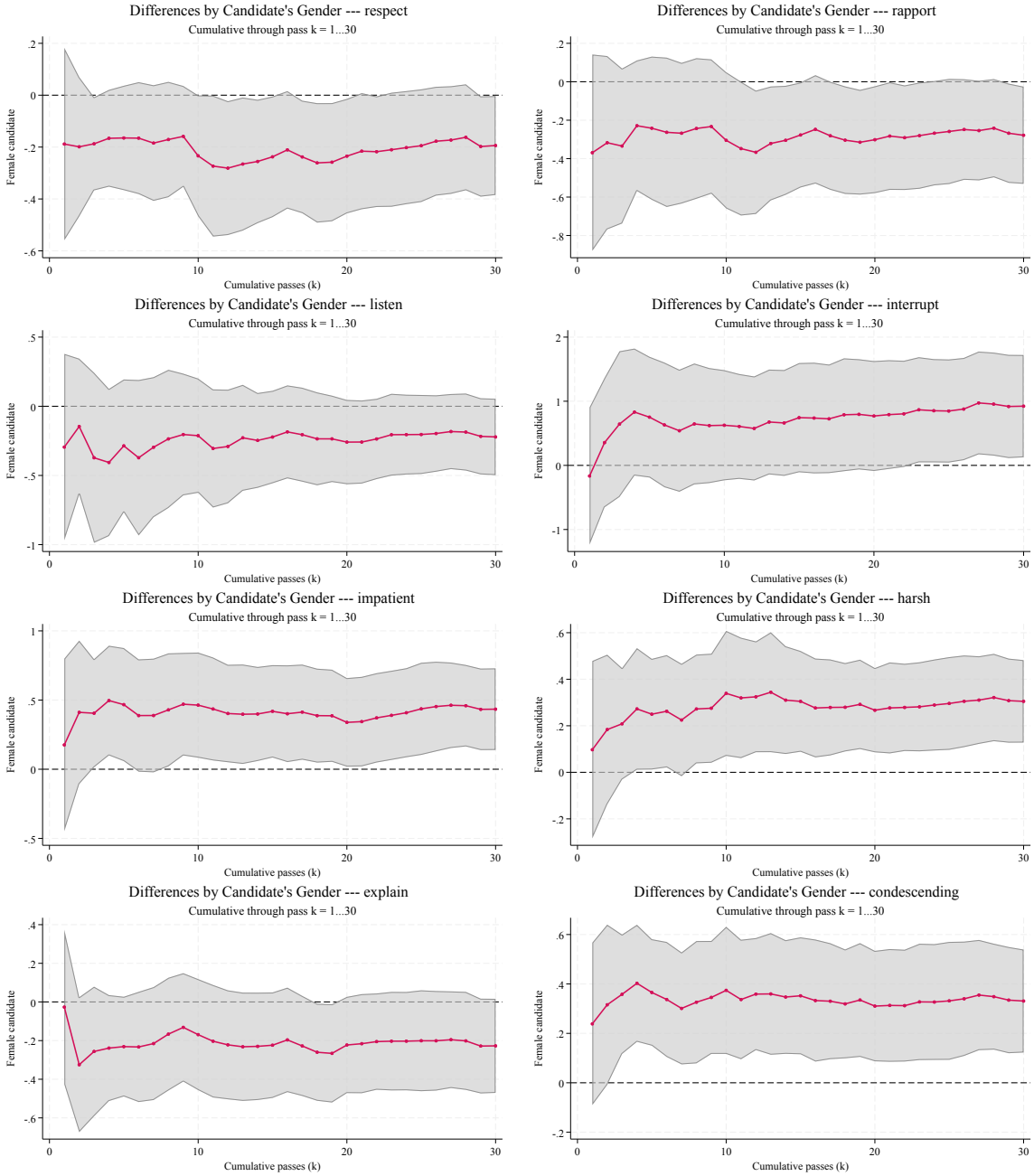
Notes: This table presents summary statistics from the video analysis. Panel A reports information at the video/URL level, based on metadata coded directly from YouTube. Panel B displays quantitative evaluations of candidates' behaviors observed in the videos, assessed using Google's Gemini Flash AI (version 2.0), with each evaluation repeated thirty times. Panel C shows quantitative assessments of interviewers' behaviors observed in the videos skills, assessed using Google's Gemini Flash AI (version 2.0), with each evaluation repeated thirty times.

Figure G1: Effect of the Gender of the Candidate on Behaviors — By Number of Iterations



Notes: This figure presents the effect of the gender of the candidate on the candidates' behaviors based on the number of iterations of Gemini's analysis of the video of the interview.

Figure G2: Effect of Gender of the Candidate on Interviewers' Behaviors — By Number of Iterations



Notes: This figure presents the effect of the gender of the candidate on the behaviors of the interviewers based on the number of iterations of Gemini's analysis of the video of the interview.