Non-Bayesian Statistical Discrimination*

Pol Campos-Mercade Friederike Mengel

November 9, 2022

Abstract

Models of statistical discrimination typically assume that employers make rational inference from (education) signals. However, there is a large amount of evidence showing that most people do not update their beliefs rationally. We use a model and two experiments to show that employers who are conservative, in the sense of signal neglect, discriminate more against disadvantaged groups than Bayesian employers. We find that such non-Bayesian statistical discrimination deters high-ability workers from disadvantaged groups from pursuing education, further exacerbating initial group inequalities. Excess discrimination caused by employer conservatism is especially important when signals are very informative. Out of the overall hiring gap in our data, around 40% can be attributed to Bayesian statistical discrimination, a further 40% is due to non-Bayesian statistical discrimination, and the remaining 20% is unexplained or potentially taste-based.

Keywords: Statistical discrimination, conservatism, naive employers, experiments.

JEL Classification Numbers: C92, D91, J71.

*Campos-Mercade: Department of Economics, University of Copenhagen, pcm@econ.ku.dk. Mengel: Department of Economics, University of Essex, and Department of Economics, Lund University, fr.mengel@gmail.com.
We thank Kai Barron, Alexander Coutts, Nickolas Gagnon, Boon Han Koh, Erik Mohlin, Florian Schneider, Charlie Sprenger, Roel van Veldhuizen, Erik Wengström, and seminar participants at the University of Alicante, BA conference on Social Norms (London), Berlin Behavioural Econ seminar, Bonn (briq beliefs workshop), Duke Kunshan University, GATE Lyon-St.Etienne, University of Glasgow, Gothenburg University, University of Hamburg, HSE Moscow, IMT Lucca, Indiana University, IWET workshop Santiago de Chile, University of Kent (Eastern Arc workshop), Lund University, University of Copenhagen, University of Essex, University of Pittsburgh, University College London, University of Dusseldorf, University of Toulouse, Uppsala University, Virginia Tech, and WU Vienna for helpful comments. Mengel thanks the European Research Council (ERC Starting Grant 805017) for financial support. Campos-Mercade acknowledges funding from the Danish National Research Foundation grant DNRF134 (CEBI).

1 Introduction

The labour market often treats minority workers less favorably than non-minority workers with otherwise identical characteristics (Bertrand and Duflo 2017). In economics, the most widely used approach to explain such discrimination involves models of statistical discrimination (Phelps 1972; Arrow 1973), in which employers use group identity to infer the productivity of prospective employees. These models typically assume that employers are endowed with prior beliefs about the characteristics of different groups and update these *rationally* based on observable characteristics, such as education and previous experience (e.g., Arrow 1973; Phelps 1972; Aigner and Cain 1977; Lundberg and Startz 1983; Fang and Moro 2011; Lang and Lehmann 2012).

At the same time, there is overwhelming evidence that people frequently display failures of Bayesian rationality (Benjamin 2019). One of the most frequently documented failures of Bayesian rationality is conservatism (also called signal neglect; Phillips and Edwards 1966; Peterson and Beach 1967; Edwards 1968; Mobius et al. ming; Buser et al. 2018; Coutts 2019), where people pay too much attention to the prior and not enough to new information. Conservatism might be particularly important in the context of discrimination, where priors are often linked to strongly entrenched stereotypes that in some cases have been formed over generations (Massey and Denton, 1993; Cutler et al., 1999; Telles and Ortiz, 2008).

In this paper, we use theory and two experiments to study the implications of conservatism for discrimination in the labour market.[1] We document that conservatism is indeed a major source of wrong beliefs and that it leads to excess discrimination against disadvantaged groups. Compared to the Bayesian benchmark, we observe excess discrimination especially when signals are highly informative. Such non-Bayesian statistical discrimination deters disadvantaged groups from pursuing education. Our paper shows not only that wrong beliefs matter, but that the source of wrong beliefs is crucial for our understanding of patterns of discrimination, of workers' human capital investments, and of the effectiveness of policy interventions.

A simple theoretical model of statistical discrimination makes these implications precise. In the model, there are real underlying ability differences between two groups, and we define the group that on average has lower ability as the disadvantaged group.[2] Potential employees from each group ("workers") can choose whether to invest in education, and employers make a hiring decision after observing education but not ability. Since low ability workers are less likely to succeed with education (as in, e.g., Spence 1973), situations can arise where the average ability of one group is lower, but—conditional on education—that same group has higher ability.[3] These are the situations where Bayesians and non-Bayesians will make different de-

[1] Our experiments allow us to study other failures of Bayesian rationality, such as base-rate neglect (Kahnemann and Tversky 1973; Grether 1980; Erev et al. 2008) and asymmetric updating (e.g., Eil and Rao 2011; Mobius et al. ming). We find that these other biases are empirically less relevant in our setting.

[2] We treat these ability differences as exogenous and do not consider how they might emerge. Unequal ability distributions are often observed in empirical work (Lang and Manove, 2011) and can have many reasons, including pre-market discrimination and historical inequalities.

[3] An example where we might encounter this pattern is women and computer science. There are fewer women

cisions. In particular, since conservatives neglect the education signal, they will be less likely to hire the disadvantaged group than Bayesians. We refer to this type of discrimination as *non-Bayesian statistical discrimination.*

We then study incentives to seek education across the groups. Workers from the disadvantaged group could either have fewer incentives to invest because they cannot shift employers' beliefs that they are low ability, or higher incentives because there is more value in distinguishing themselves from low ability workers. Our model highlights this trade-off and shows that which force prevails depends on whether employers update rationally. More concretely, in every equilibrium of the sequential labour market game, disadvantaged workers are (weakly) less likely to pursue education than advantaged workers of the same ability when they face conservative employers, but not when they face Bayesian employers.

We design a lab experiment that allows us to test the intuitions developed in the theory. The experiment enables us to make clean inference on whether people make rational inference from education signals and how this affects workers' decisions. Conducting a lab experiment is ideal to study these questions, as it allows us to (i) exclude channels such as taste-based discrimination, (ii) use a clean measure of ability and control how it affects employers' payoffs, and (iii) establish a causal link between employers' decisions and workers' education choices.

We find substantial evidence of conservatism, with a larger share of decisions being consistent with conservatism than with Bayesian reasoning. As a result, the disadvantaged group is hired 52% less frequently compared to what we would expect if all employers were Bayesian. We also find that—conditional on ability—workers from the disadvantaged group seek education much less frequently than others. Using a treatment variation in which we vary the proportion of conservative employers, we establish a causal link between employer naivete and a decrease in education among the disadvantaged group. This finding is important because market exit further exacerbates imbalances between the two groups, leading to a substantial welfare loss because fewer high-ability workers are hired. Moreover, the fact that many high-ability disadvantaged workers exit the market early makes it harder for employers to learn. We do indeed not find any evidence that the quality of employers' decisions improves over time.

The experiment shows that non-Bayesian statistical discrimination is empirically relevant and that it can lead to under-education and market exit by the disadvantaged group. Our second experiment focuses on understanding more precisely *how* people differ from Bayesian rationality when they update about others and whether such updating depends on their identity. Our design allows for several rounds of updating across a broad range of beliefs and in situations where tastes can matter too. Participants are shown profiles (containing age, gender, region of residence, marital status, and field of studies) of a number of different candidates and asked to indicate a prior on the likelihood that the candidate is in the top half of performers

studying computer science than men (Ceci et al. 2014) and—in line with this pattern—women are perceived as "on average worse" in coding than men (Terrell et al. 2017; see also Bohren et al. 2018). However, conditional on having programming knowledge, there is suggestive evidence that women are better coders than men (Terrell et al. 2017).

in a math and logic task. Afterwards, they receive a sequence of five informative signals and are asked after each signal to indicate a posterior belief. At the end, they decide whether to hire the candidate, which means paying a fixed wage to the candidate and earning money in case the candidate is indeed a top performer.

We find substantial conservatism in updating also in this second experiment. Since participants express on average a lower prior for female candidates, conservatism means that posteriors for women (relative to men) are lower than what they should be under Bayesian rationality. This is especially the case when the sequence of signals is very informative. For example, for five positive signals, the gender gap in posteriors is 10 times larger than what it should be with Bayesian updating. We also find gender gaps in hiring, with women on average being around 7 percentage points less likely to be hired. Back of the envelope calculations show that around 40% of these hiring gaps can be attributed to Bayesian statistical discrimination (this is, correct posteriors given the signals and participants' reported priors), a further 40% are due to non-Bayesian statistical discrimination, and the remaining 20% are unexplained or potentially taste-based.

The second experiment also allows us to explore whether there is an interaction between tastes and non-Bayesian statistical discrimination. In particular it allows us to study whether people update differently depending on the candidate's gender. We find that participants are somewhat more conservative when evaluating men compared to when they evaluate women. We also find evidence of asymmetric updating. Irrespective of who they evaluate, participants update more after seeing a negative signal than after seeing a positive signal. This effect is somewhat stronger when evaluating women than men, but the difference is not statistically significant. These results complement recent work on self-stereotyping by Coffman et al. (2021), who find that men react more to positive signals in male-typed domains when evaluating themselves.

Our paper contributes to a wide literature investigating discrimination and its sources (for reviews, see Charles and Guryan 2011, Bertrand and Duflo 2017, and Neumark 2018). Economists typically categorize discrimination as either taste-based or statistical. Taste-based discrimination assumes that discrimination arises because agents have preferences against certain groups (Becker 1957). By contrast, statistical discrimination argues that discrimination occurs because employers use group identity to make rational inference on the (unobserved) productivity of each individual (Phelps 1972; Arrow 1973). Empirically, there is convincing evidence that statistical discrimination is indeed a large and important factor in explaining group inequalities (List 2004; Autor and Scarborough 2008; Agan and Starr 2018), but there usually exists an unexplained residual that is often attributed to taste (Altonji and Pierret 2001; Knowles et al. 2001; Charles and Guryan 2008; Lippens et al. 2020).

In contrast to these explanations, this paper adds to this literature by proposing that wrong belief updating might be at the core of much of the observed discrimination. Our results highlight the importance of non-Bayesian statistical discrimination in explaining overall discrimination in labour markets. Without taking into account the possibility of non-Bayesian statistical discrimination, researchers may be tempted to classify unexplained discrimination as taste-based when it in fact can be traced back to failures of Bayesian rationality. Distinguishing Bayesian and non-Bayesian statistical discrimination is also important because policy recommendations can differ depending on the source of discrimination. For example, affirmative action policies can often backfire under Bayesian statistical discrimination (Coate and Loury 1993; Moro and Norman 2003; Fang and Norman 2006). A policy that improves access to education for disadvantaged groups, for example, might make educated workers from these groups less attractive for Bayesian employers. However, such policies might be effective against non-Bayesian statistical discrimination: First, they can address the issue of "under-education" of the disadvantaged group by inducing more people to seek education who would be doing so if employers were rational. Second, by reducing market exit from high-ability workers of disadvantaged groups, such policies can allow employers to learn to make improved inference from information (Beaman et al., 2009; Niederle et al., 2013).

Our paper also contributes to an emerging literature that studies whether labour market discrimination can stem from wrong or inaccurate beliefs (Bohren et al., 2021). Mobius and Rosenblat (2006) find that subjects pay a wage beauty premium because they wrongly believe that attractive people are more productive. In a setting in which subjects choose whether to hire a woman or a man, Reuben et al. (2014) document a strong bias against hiring women that is only partly attenuated by objective information about past performance. In a similar setting, Barron et al. (2020) find belief-based discrimination against women even when they are equally qualified as men. Finally, Bohren et al. (2021) document discrimination against Americans and women partly based on wrong stereotypes. They also show that, when people can hold wrong beliefs, it is no longer possible to identify taste-based or statistical discrimination using methods of inference common in the literature. Esponda et al. (2022) find that contrast-biased beliefs are an important factor in generating discrimination and Ruzzier and Woo (2022) examine the consequences of inaccurate beliefs and confirmation bias for discrimination.[4]

While this literature documents that discrimination may stem from wrong beliefs, it does not study *why* such wrong beliefs emerge. Our paper contributes to this work by showing that non-Bayesian updating may generate the wrong beliefs that lead to discrimination. We further show that both the patterns of discrimination as well as the policy implications can differ depending on why beliefs are inaccurate. With conservatism, we would expect "excess discrimination" (compared to the Bayesian case) especially in cases where signals are very informative. By contrast, if beliefs are inaccurate simply because people hold wrong priors, "excess discrimination" is more likely when signals are uninformative (and hence priors are more im-

[4]Relatedly, there are few papers that show discrimination based on wrong beliefs in non-labour market settings. In a trust game, Fershtman and Gneezy (2001) show that subjects give less when the trustee is a person from a different ethnicity, but this is because they have mistaken ethnic stereotypes. Albrecht et al. (2013) find that subjects believe that members of a group that performs worse are less likely to be high performers, even when group identity is completely irrelevant for the evaluation. Finally, Arnold et al. (2018) argue that judges in the US discriminate against black defendants not because they are racially prejudiced, but because they make racially biased prediction errors.

portant). In the latter case, merely providing high quality information should eliminate biases, but in the former case providing information alone might not be successful, as people might not make accurate inference from the information provided. We also go beyond the literature by causally showing, both theoretically and empirically, that non-Bayesian discrimination can discourage workers from the disadvantaged group to pursue education, further exacerbating the group inequalities.

Finally, we also contribute to a small literature that studies how cognitive limitations affect stereotypes and discrimination. Bertrand et al. (2005) review previous evidence in economics and psychology and argue that discrimination may stem from unintended implicit attitudes. Bartoš et al. (2016) show that the time that people use in screening an application depends on whether the applicant is from a minority group, which is consistent with a model of endogenous allocation of costly attention. Bordalo et al. (2016) propose and test a model in which the representativeness heuristic (Kahneman and Tversky 1972) leads to wrong stereotypes (see also Benjamin et al. 2016 for a study on the economic implications of the "nonbelief in the Law of Large Numbers"). Delavande and Zafar (2018) show that anti-american attitudes persist after information provision because updating is not rational. We contribute to this literature by showing that conservatism (Phillips and Edwards 1966), one of the most well-studied cognitive biases, could be a key determinant of discrimination. More concretely, we show that conservatism can lead to "excess discrimination" against disadvantaged groups, which in turn makes members of these groups less likely to invest in their human capital.

The paper is organized as follows. In Section 2 we outline our theoretical model. Section 3 contains the design and results of the lab experiment and Section 4 contains the design and results of the online experiment. Section 5 concludes.

2 Theory: non-Bayesian Statistical Discrimination

In this section, we outline our model of statistical discrimination and explain how conservatism can affect discrimination. The aim of the model is to capture the essence of non-Bayesian statistical discrimination in a simple and intuitive way. It is not meant to be a major theoretical contribution in itself.

We consider a labour market with a large number of workers. Each worker is characterized by three parameters: (i) an ability level $a \in \{l, m, h\}$ (low, medium or high), (ii) a level of education $\theta \in \{e, ne\}$ (educated or not educated), and (iii) a group identity $i \in \{r, g\}$ (red or green). The proportion of workers of group identity i who have ability a is denoted by i_a. We allow ability distributions to be ex ante unequal, as is typically observed in empirical work (Lang and Manove, 2011). Specifically, without loss of generality we assume that the red workers are a *disadvantaged group* that, for whatever reason, has a less favorable distribution of abilities than green workers. In particular, we assume that red workers have on average lower ability than green workers (as assumed in e.g. Phelps 1972).

Workers decide whether to pursue education or not. As is typical in statistical discrimination models (Fang and Moro, 2011) whether a worker is successful in obtaining education depends on whether the worker pursues education and on the worker's ability. Specifically, if a worker of ability a pursues education, the probability of being successful is given exogenously by the following probabilities,

$$p_a = \begin{cases} 0 & \text{if } a = l \\ p_m & \text{if } a = m \\ 1 & \text{if } a = h \end{cases},$$

where $p_m \in (0, 1)$.

Employers do not know the ability of each worker, but they know the prior ability distribution of each identity. They also observe workers' education level θ and can use it to make inference on their (unobserved) ability. For simplicity, we assume that each employer chooses between one worker of identity red and one of identity green. They earn $X_{\theta,a}$ if the hired worker is of education θ and ability a. Employer payoffs satisfy (i) $(X_{e,a} - X_{ne,a}) = \alpha_a$, $\forall a$ with $\alpha_a \in \mathbb{R}^+$, i.e. conditional on ability employers prefer educated to non-educated workers, and (ii) $(X_{\theta,h} - X_{\theta,m}) = (X_{\theta,m} - X_{\theta,l}) = \beta_\theta$, $\forall \theta$ with $\beta_\theta \in \mathbb{R}^+$, i.e. conditional on education employers prefer higher levels of ability, with the ability premium constant across levels of ability. Payoffs for the worker are w if they are hired and 0 otherwise.

Discrimination. Employers discriminate if, given the same observable level of education of two candidates, they have a strict preference for one identity over the other.

Discrimination can be rational if employers are Bayesian or irrational if there are failures of Bayesian rationality. We do not consider taste-based discrimination in this section.

2.1 Employer Decisions

To develop the intuition for the results we first assume that all workers pursue education.

Bayesian employers. For a Bayesian employer, the probability that an educated worker of identity $i \in \{g, r\}$ has ability $a \in \{l, m, h\}$ is

$$P_B(a|e, i) = \frac{p_a i_a}{p_l i_l + p_m i_m + p_h i_h}$$

and their expected payoff of hiring an educated worker of identity i is

$$\pi_B(e, i) = \frac{p_l i_l X_{e,l} + p_m i_m X_{e,m} + p_h i_h X_{e,h}}{p_l i_l + p_m i_m + p_h i_h}.$$

A Bayesian employer will hence discriminate against red educated workers whenever $\pi_B(e, g) > \pi_B(e, r)$, or

$$\frac{g_h}{g_m} > \frac{r_h}{r_m}. \tag{B}$$

Naive Employers (Conservatives). Naive employers suffer from signal neglect. Hence, for these workers $P_N(a|e,i) = P_N(a|i) = i_a$, meaning that they do not account for the information contained in the education signal.[5] Naive employers will discriminate against red educated workers whenever $\pi_N(e,g) > \pi_N(e,r)$, or

$$g_m - r_m > 2(r_h - g_h). \tag{N}$$

This inequality is identical to the condition ensuring that green workers have on average higher ability than red workers (see Appendix A). Hence naive employers always discriminate against the disadvantaged group. This leads to the following proposition.

Proposition 1*. Suppose all workers attempt education. If* ***B*** *is satisfied, then Bayesian and Naive employers both discriminate against red workers. If not, then only Naive employers discriminate against red workers.*

Proof. Appendix A. □

The proposition shows that Naive employers will discriminate against red both when it is rational to discriminate but also in some cases in which it is not rational to do so. Hence, compared to the rational (Bayesian) case, naive employers discriminate against red workers too often. By contrast, they never discriminate against green workers when Bayesians would not.[6]

2.2 Full Characterization of Equilibria

We now study whether and how employer naivete can affect workers' decisions to educate. To do so, we consider the full two-stage game where workers decide in stage 1 whether to pursue education or not $\eta \in \{E, \neg E\}$. If they do not pursue education, they remain uneducated. If they pursue education, they become educated with probability p_a. Pursuing education costs c regardless of whether it is successful or not. In stage 2, each employer is randomly matched with one red and one green worker and chooses who to hire. To simplify the analysis, and in line with the experiment, we will assume $X_{ne,a} < w \; \forall a$, such that employers are only interested in hiring educated workers.

The first observation to note is that since education is costly and low ability workers have no chance of succeeding, they never pursue education. We hence denote by $(\eta_m^r, \eta_h^r; \eta_m^g, \eta_h^g)$ the four-tuple of the medium and high types' education decisions for each identity. We assume that $w - c > 0$, ruling out the case that education is not worthwhile for anyone.

Bayesian employer. When all employers are Bayesian, there are six possible pure strategy equilibria: two symmetric equilibria in which workers of the same ability make the same

[5] In the paragraph "Robustness and Extensions" we discuss, among other extensions, the case where employers are partially naive.

[6] Note that, under base-rate neglect, employers ignore the differences in base-rates and instead only react to the education signal. Hence, if both workers have the same education signal, such employers are indifferent and do not discriminate against either identity.

education choices, $\{\neg E, E; \neg E, E\}$ and $\{E, E; E, E\}$; two equilibria in which green workers pursue education more than red workers, $\{\neg E, \neg E; \neg E, E\}$ and $\{\neg E, E; E, E\}$; and two equilibria in which red workers pursue education more than green workers, $\{\neg E, E; \neg E, \neg E\}$ and $\{E, E; \neg E, E\}$. In equilibrium, uneducated workers are never hired and the choice between educated workers depends on the ability distribution *among educated workers* of each identity (as shown in Appendix Table B.3).

Naive employers There are five possible pure strategy equilibria when employers are Naive: two symmetric equilibria in which workers of the same ability make the same education decisions, $\{\neg E, E; \neg E, E\}$ and $\{E, E; E, E\}$; and three asymmetric equilibria in which green workers educate more than red workers, $\{\neg E, \neg E; \neg E, E\}$, $\{\neg E, \neg E; E, E\}$ and $\{\neg E, E; E, E\}$. Appendix Table B.3 shows the parameter conditions that support each of these equilibria. There are no equilibria in which red workers are more likely to educate. The reason is that, since—conditional on education—naive employers always prefer green workers, there are always higher incentives for green workers than for red workers to pursue education.

Proposition 2. *If the employer is Naive, then green workers pursue education weakly more often than red workers in all equilibria and strictly more often in some.*

Proof. Appendix B.1. □

Hence "excess discrimination" by naive employers translates into "under-education" by the discriminated group. The logic of this proposition extends to mixed populations. Consider a situation in which workers face Naive employers with probability γ and Bayesian employers with probability $1 - \gamma$. We assume that workers know the proportion of employers who are Naive and Bayesian, but do not know the type of the employer that they are matched with. The following proposition holds.

Proposition 3. *If the share of Naive employers in a labour market γ is sufficiently large, then green workers pursue education weakly more often than red workers in all equilibria and strictly more often in some.*

Proof. Appendix B.3. □

Robustness and Extensions. The basic intuition behind our results is very clear: since naive employers fail to correctly interpret the information contained in education signals, they base their decisions on beliefs that are "too pessimistic" about the quality of red workers. This basic intuition still holds under many modeling alternatives we could consider. In Appendix A, we show that Proposition 1—with adapted qualifiers—also generalizes to more general payoff schemes (Appendix A.2) as well as to continuous ability distributions (Appendix A.3). Appendix B contains the proofs of Propositions 2 and 3 and shows that—qualitatively—both results also hold when more general payoff schemes are considered (Appendix B.2).[7] Extending

[7]When including more general payoff structures for the employers, since condition **N** is not always satisfied anymore, there can emerge equilibria with Naive employers in which red workers are more likely to educate than green workers of the same ability. Appendix B.2 shows the results of numerical simulations where we study how common the different equilibria for different payoff structures are. In sum, we find that both equilibria in which

the model to continuous (e.g. normally distributed) education levels is straightforward as long as education decisions are considered exogenous (Proposition 1), but raises conceptual questions when education decisions are endogenous. Last, it should be noted that our naive agents are "fully naive" in the sense that they ignore the education signal entirely. It is, of course, possible to relax this assumption. If we assume that naive agents take the signal (fully) into account with a certain probability $1-\gamma$ and fully ignore it with probability γ, we generate a model that is isomorphic to the model with the mixed population considered in Proposition 3. Alternatively we could assume that naive agents account for the signal but not fully, e.g. by assuming that their posterior is a convex combination of $P_B(a|e,i)$ and $P_N(a|e,i)$. This does not generate a fully isomorphic model to the one underlying Proposition 3, but it does generate very similar incentives for workers.

We now proceed with our empirical analysis which features two experiments. Experiment 1 tests the model above and is mainly focused on education decisions. It was designed to (i) study whether "excess discrimination" is empirically relevant and (ii) identify a causal link between "excess discrimination" and under-education by workers of the disadvantaged group. In Experiment 2, by contrast, education decisions are shut down. Here, the main focus is to study how people differ from Bayesian rationality when they update their beliefs about others, and to explore whether such updating depends on the other's identity.

3 Experiment I

This section contains the design and the results of our first experiment. The experiment was designed to remain close to theory and address two questions. First, do people in the role of employers discriminate against the disadvantaged group "too much" compared to Bayesians in the setting outlined above? And second, if they do, is this learned by workers and does it lead to under-education of the disadvantaged group? The second question is the main question this experiment was designed to address. We implement the experiment in the lab, which allows us to closely mimic the setting described in Section 2, to shut down taste-based discrimination, and to cleanly identify the causal effect of employer naivete on workers' education decisions. Our second experiment discussed in Section 4 will then focus on updating and "excess discrimination" in more detail.

3.1 Design

Participants in the lab were randomly assigned either the role of worker or the role of employer and—if they were a worker—they were also randomly assigned (i) an ability level $a \in \{l, m, h\}$ (low, medium, or high) and (ii) an identity $i \in \{g, r\}$ (green or red). To avoid taste-based discrimination based on political or undesirable connotations of these colours, in the lab we

red workers educate more and symmetric equilibria are *much* more common with Bayesian employers than with Naive employers.

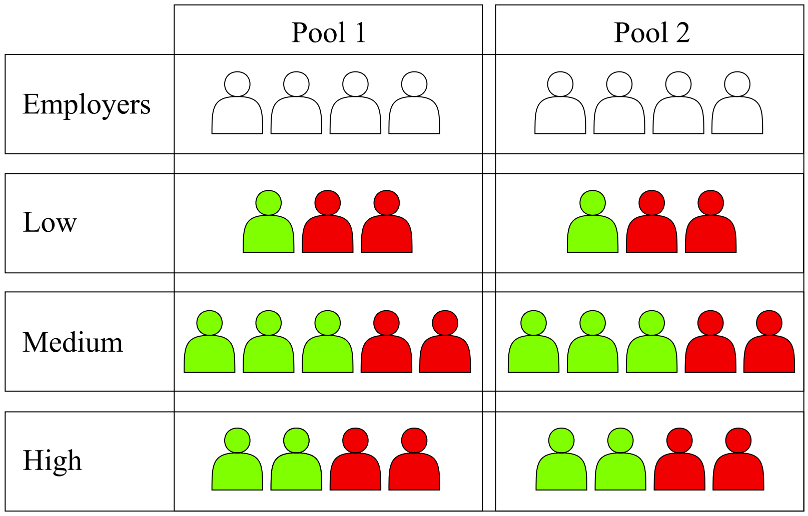


Figure 1. Role assignment in each session.

used the colours *yellow* and *orange*.[8]

Figure 1 describes the role assignment in each session. Each session had 32 participants, out of whom 24 were randomly selected to be workers and 8 to be employers. Out of the 24 workers, half were randomly placed in the green group and the other half in the red group. Out of the 12 green workers, four had high ability, six had medium ability, and two had low ability. For the red workers, four each had high, medium, and low ability. Neither the workers nor the employers knew the total number of workers of each ability and colour in the session. The role, colour, and the ability of each participant remained fixed throughout the experiment.[9]

The workers and employers were divided into two *pools* of 4 employers and 12 workers each, with symmetric ability distributions. Participants played 60 rounds. In each round and within each pool, four workers of each colour were randomly drawn to play that round, with the condition that the four green workers are on average better than the red workers.[10] This means that two green and two red workers in each pool remained unmatched and received a default payment of 4 GBP for that round. This feature of the design is important since it allows the ability distribution of the matched workers to randomly vary in each round. For example, in one round a group could consist of two low, one medium, and one high ability worker, while in another round it could consist of two low and two high ability workers. As shown in Figure 2, all employers and workers could see the drawn ability distribution in each round.

[8]For consistency with the theory section, we use red and green throughout the paper. Since typically green has better connotations than red, we hope that using these colours will help the reader recall that the green workers are the advantaged group.

[9]The experiment hence uses artificial identities as in much of the literature on in-group bias (Chen and Li, 2009; Chen and Mengel, 2016). Note, though, that as employers are not assigned a colour, different identity-groups never interact in our experiment. As a consequence there is no scope for in-group bias in our experiment.

[10]We impose this condition as we are interested—in line with the theory—in situations where green workers are better on average. We considered randomly switching the labels (green and red) across rounds, but decided that this is not desirable as we wanted to allow participants to learn across rounds.

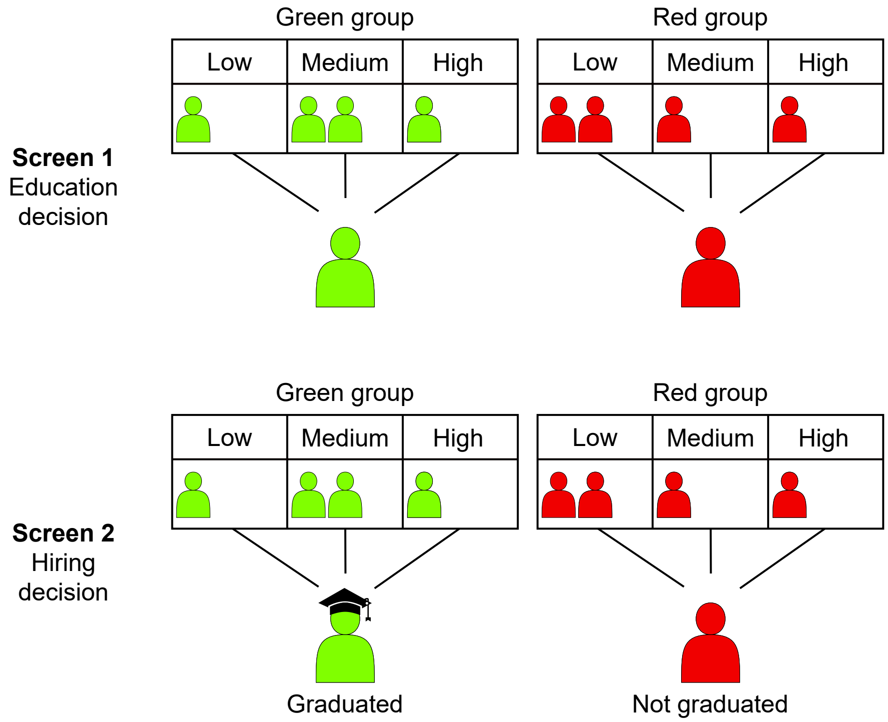


Figure 2. Example of an ability distribution as presented to participants.

Note: On Screen 1, participants saw the ability distribution, were reminded on the same screen of their role and, if they were workers, of their identity and ability level. The screen also summarized the payoff parameters of the game. Workers were asked whether they wanted to pursue education. On Screen 2, participants were in addition able to see whether each of the two workers in their match was successful with education ("graduated"). Employers were asked to make their hiring decision.

Education and Hiring Decisions. In each round, the matched workers and employers took the following decisions. First, the workers were endowed with 4 GBP and decided whether to pursue education (workers who were not drawn just kept their 4 GBP as their round payment). The cost of pursuing education was 1 GBP and $p_m = 0.8$. Hence, a worker of medium ability was successful in pursuing education with probability 0.8. As described in the theory section, $p_l = 0$ and $p_h = 1$. Second, employers made a hiring decision. They chose between hiring the red worker, the green worker, and not hiring at all. When making their decision, they observed whether the red worker and the green worker, respectively, were successful with education as well as the ability distribution among the 8 matched workers in that round. Figure 2 shows how this information was presented to employers and workers in the experiment. Both workers and employers observed both screens, but in Screen 1 only the workers made a decision and in Screen 2 only the employers made a decision.

In each round, employers received 20, 15, or 10 GBP if they hired an educated worker of high, medium, or low ability, i.e. $X_{E,h} = 20, X_{E,m} = 15, X_{E,l} = 10$. If they hired a worker who was not successful with education they received 0 GBP, and if they decided not to hire they received 8 GBP. Hired workers received a wage of 8 GBP. A worker who was not hired received 0 GBP. Given these parameters, employers should not hire at all if neither worker is educated and hire the green (red) worker if only the green (red) worker is educated. More importantly, the ability distributions across all rounds were such that, if medium and high-ability workers always pursued education, then when both workers are educated Bayesian employers would hire the green (red) worker exactly 50 percent of the time, while naive employers would always hire the green worker.

At the end of each round, employers were informed about their payoffs and—if they hired someone—about the ability level of the worker hired. Workers were informed about their payoffs and whether they were hired.

New Pools after Round 30. To study the causal effect of employer conservatism on education decisions, we reassigned pools after the first thirty rounds of the experiment. To do so, the computer counted the number of times that each employer hired a red worker as a proxy for naive behavior (e.g., hiring red when a Bayesian would have hired green). It then reassigned randomly the four most Bayesian employers to one pool (the "Bayesian pool") and the four most naive (conservative) employers to another pool ("the Naive pool"). Workers were told on the screen that the set of employers might have changed, but they did not know how the employers had been reassigned. Since the pool assignment was done randomly, any differences in behavior across pools is due to whether workers are more likely to face naive or bayesian employers.

Questionnaire. At the end of the experiment, we elicited workers' beliefs regarding the proportion of employers who hired a red worker and the proportion of medium-ability green and red workers who pursued education. They were paid 2 GBP for each question if their guess was within 10% of the correct answer. Additionally, all subjects filled out six cognitive ability questions and self-reported risk aversion. At the end, we also gathered demographics information: sex, age, country of origin, ethnicity, field of studies, and self-reported social class. Appendix C contains information on these variables.

Procedure. The final payment was the sum of one randomly selected round and the incentive payments from the questionnaire. The experiment lasted about one hour and thirty minutes and participants were paid on average 16.67 GBP. The experiment was conducted at the Essex Lab at the University of Essex in September-October 2019. In total, 320 participants participated in 10 sessions of 32 participants each. Participants read the instructions and were asked to correctly answer 12 understanding questions before the experiment began. We received ethical approval from the University of Essex Social Sciences subcommittee under Annex B in April 2019. The experiment was pre-registered at the AEA registry in September 2019 (AEARCTR-0004652). We used hroot (Bock et al. 2014) to recruit participants and ztree (Fischbacher 2007) to code the experiment.

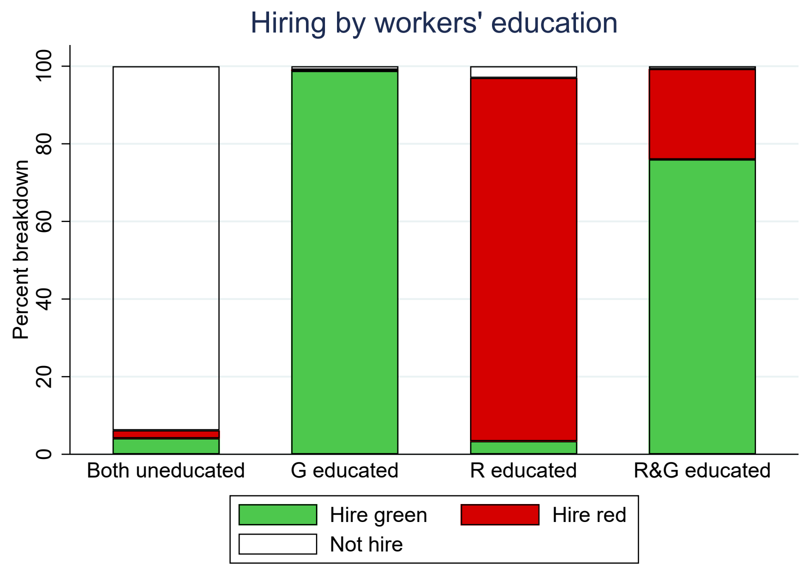


Figure 3. Summary of Employer Decisions.

Note: Percent of the time a green worker, a red worker, or no one is hired depending on whether neither worker, only the green worker, only the red worker, or both workers are educated.

3.2 Results

3.2.1 Employer decisions

This section studies whether employers' behavior is in line with conservatism, i.e. whether employers do indeed not sufficiently account for education signals. We start by looking at overall hiring decisions. Figure 3 shows employer decisions depending on which of the two workers was successful in obtaining education, across all sessions and rounds. We find that when neither worker is educated, the vast majority of employers (94%) do not hire. When only the green worker is educated, almost all employers (98%) hire the green worker. And when only the red worker is educated, a similarly large proportion (94%) hire the red worker. This shows that employers understand well the incentives provided in the experiment. The most interesting bar is the rightmost one, which shows who is hired when both workers are educated. The figure shows that in this case the green worker is hired 76% of the time and the red worker only 23% of the time. Recall that, according to our set of ability distributions, if all workers pursued education then Bayesian employers should have hired each type of workers exactly 50% of the time. Since naive employers would always hire green, this suggests that a considerable proportion of employers do indeed show conservatism. Note that misspecified beliefs cannot explain the low frequency with which red workers are hired. To choose a red worker with only 23 percent frequency, Bayesians would have to believe that medium ability red workers are more likely to pursue education than high ability red workers *and* more likely than green workers. This belief is both unreasonable (diametrically opposed to red workers' incentives) and not supported by the data. In fact medium ability red workers are by far the

Table 1. Employer decisions in Experiment 1

	% hiring red				% high-ability		Cost of Mistake
	Bayesian	Lab	p-value	N	Red	Green	in GBP
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Overall	50	23	0.0190	1124	64	48	1.37
G $\succ_B$ R	0	7	0.2530	464	21	53	1.60
R $\sim_B$ G	50	13	0.0375	103	38	38	0.00
R $\succ_B$ G	100	38	0.0034	557	72	43	1.45

Note: Percent of Employers hiring red workers *conditional on both workers being educated.* G $\succ_B$ R, R $\sim_B$ G, and R $\succ_B$ G capture distributions for which Bayesian employers prefer green workers, are indifferent, and prefer red workers, respectively. Column (1) shows Bayesian decision-makers (theoretical prediction) and column (2) lab averages overall and by preferences of a Bayesian. The p-value is of a post-regression test on whether lab behaviour equals the Bayesian prediction in column (1). (The regression regresses the share of red workers hired on a constant, accounting for auto-correlation at the individual level and session fixed effects). N is the number of observations. Columns (5) and (6) show the percentage of hired red and green workers who are high-ability and column (7) shows the corresponding cost of a mistake, i.e. of making a choice inconsistent with Bayesian updating.

least likely group to pursue education (apart from low ability workers, see Section 3.2.2).[11]

Table 1 focuses on employer decisions in the most interesting case: when both workers were successful in obtaining education. The table shows that participants hire red workers much less frequently than green workers overall, even when a Bayesian decision-maker would strictly prefer a red worker (bottom row of Table 1). In this case, 72% of red workers have high ability as opposed to only 43% of green workers. The cost of making a mistake is 1.45 GBP, which corresponds to about 9% of their total earnings. Yet, in 62% of their decisions, employers make this mistake. Note further that—in all cases—a larger share of decisions is consistent with full signal neglect (conservatism) than it is with Bayesian reasoning. Overall, this analysis indicates that there is a substantial degree of conservatism in employer decisions.

We also ask whether these mistakes self-correct over time. Specifically, we ask how participants' propensity to hire red workers in situations where a Bayesian would hire a green worker changes over time (i.e. over the 60 rounds of the experiment). We find no evidence of a time trend. The coefficient on the round variable is -0.0004 ($p = 0.926$) in a simple linear

[11]We elaborate a bit more on this point. The 50% Bayesian benchmark assumes that 1) employers face each possible ability distribution in the experiment the same number of times, and 2) medium and high-ability workers always pursue education. However, since the analysis conditions on situations where both workers were actually educated, and since not all medium and high-ability workers always pursued education, in practice the real benchmark may be different from 50%. If we check the actual ability distributions that employers faced in the experiment when both workers were educated, and assume that employers have correct beliefs about the proportion of workers of each type who pursue education in each game, then the Bayesian benchmark becomes 46%. This means that, when both workers are educated, Bayesian employers with rational (correct) expectations would have hired green workers 46% of the times, so even less often. To exceed 50% (let alone reach 76%) with misspecified beliefs Bayesians would need to rely beliefs that are (i) almost the exact opposite of what we see empirically and (ii) diametrically opposed to red workers incentives by assuming that the type that has lower probability of success will attempt education substantially more often than the high type. Hence, the 50% benchmark that we use is an upper bound.

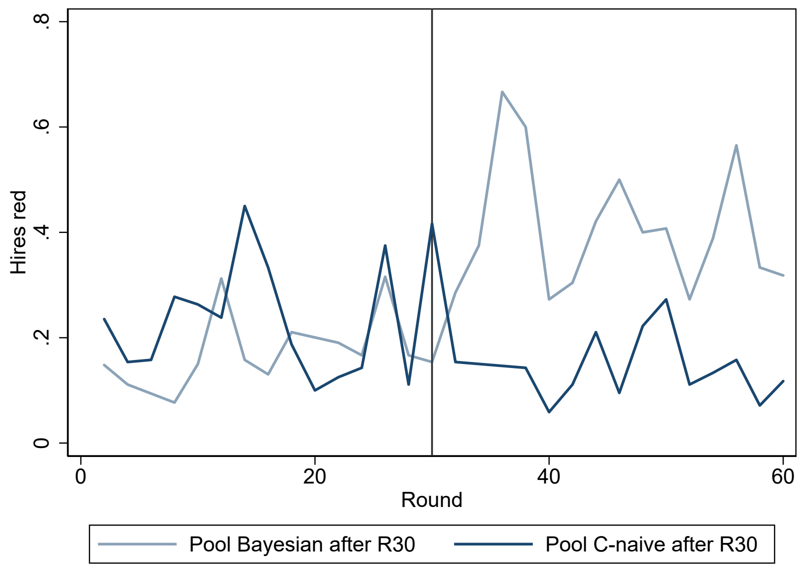


Figure 4. Share of employers hiring a red worker

Note: Share of employers hiring the red worker across the sixty rounds of the experiment *conditional on both workers being educated*. Light blue line is Pool 1 which after round 30 contains the most Bayesian employers. Dark blue line is Pool 2 which after round 30 contains the most naive employers.

regression. This suggests that these mistakes are persistent and do not simply go away as people learn to make decisions in these environments.

Last, we provide some evidence on how employers that are allocated to the "Bayesian pool" differ from those who are allocated to the "Naive pool" after round 30. Across the first 30 rounds, employers hire the red worker on average 9 times (median 9). This number ranges between 1 and 21 times. There is hence substantial heterogeneity at the individual level in terms of how close decisions are to the Bayesian benchmark. Those who end up in the "Bayesian pool" after round 30 hired a red worker on average 13 times (median 11) and those who end up in the "Naive pool" on average 7 times (median 6).[12]

Figure 4 shows that those who are assigned to the Bayesian pool after round 30 are also more likely to hire red workers subsequently. This shows that employers are consistently Bayesian or Naive. It also means that red workers assigned to the "Naive pool" after round 30 face more discrimination than those placed in the "Bayesian pool". We next ask whether this has an effect on workers' education decisions.

[12]We also checked whether employers in the Bayesian pool differ from participants in the Naive pool in other systematic ways. If, for example, employers in the Bayesian pool showed higher willingness to engage in cognitive reflection in our post-experimental task (Frederick, 2005) than employers in the Naive pool, then this could suggest that one of the reasons that people show naive behaviour is that they are trying to avoid possibly higher cognitive costs associated with the Bayesian decision. We do not find substantial differences neither in cognitive reflection between the two groups nor in terms of the demographics we elicited.

Table 2. Education choice of medium and high ability workers

	Green workers				Red workers			
	All	All	m	h	All	All	m	h
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Round>30	0.037***	0.040***	0.046**	0.031	-0.023	-0.025	-0.045	0.005
	(0.013)	(0.014)	(0.019)	(0.019)	(0.033)	(0.033)	(0.040)	(0.056)
Round>30 × Naive Pool	0.021	0.027	0.031	0.021	-0.134***	-0.136**	-0.215***	-0.013
	(0.015)	(0.017)	(0.025)	(0.022)	(0.051)	(0.053)	(0.074)	(0.064)
Observations	3764	3764	2206	1558	2380	2380	1452	928
Mean outcome	0.946	0.946	0.940	0.954	0.645	0.645	0.590	0.732
Fixed Effects		X	X	X		X	X	X

Note: Education choice of medium and high-ability workers regressed on a dummy indicating *Round>30* and the interaction with a dummy indicating that the worker was randomly assigned to the Naive pool. Columns (1)-(4) consider green workers and columns (5)-(8) consider red workers. Columns (3) and (7) consider medium-ability workers, columns (4) and (8) consider high-ability workers, and columns (1), (2), (5), and (6) consider both workers of medium and high ability. *Fixed Effects* includes individual fixed effects. The main variable of interest is *Round>30× Bayesian Pool*, which captures whether workers educated differently if they were assigned to the Bayesian Pool as opposed to the Naive Pool. Column (6) contains the main pre-registered specification (AEARCTR-0004652). Standard errors are clustered at the subject level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

3.2.2 Worker decisions

This section studies whether the disadvantaged group educates "too little" and whether this can be linked to employers' conservatism. We first have a look at overall rates of education. Across the 60 rounds, green high-ability workers choose education 95$\%$ of the time, compared with 73% of the time for red high-ability workers. For medium-ability workers, the comparison is 94% for green versus 59% for red workers. A proportions test reveals that both of these comparisons are highly statistically significant ($p < 0.001$). Hence, red workers clearly educate much less than green workers.

To understand whether these differences can indeed be linked to employer naivete, and in particular to conservatism, our main pre-registered test asks how behaviour changes after round 30 for those workers now matched with the most Bayesian employers as compared to those matched with the most conservative ones.[13] Table 2 reports OLS estimates where we regress education choice of medium and high-ability workers on a dummy indicating Round>30, and the interaction with a dummy indicating that the worker was randomly assigned to the Naive pool. Our main hypothesis was that this interaction should be negative for red workers, i.e. that red workers educate less when assigned to naive employers as opposed to Bayesian ones.

Table 2 shows that this is indeed the case. Red workers placed in the Naive pool are on average 13.6 percentage points less likely to choose education than those in the Bayesian pool ($p = 0.013$ in column 6, which is the main pre-registered specification). This represents ap-

[13] Note that in order for this test to work we don't actually need employers to "be" Bayesian or conservative. It suffices that they hire red workers with different frequencies as predicted by the theory. Experiment II will study updating in more detail.

proximately a 23% decrease over the baseline of around 60 percent. Hence, we find a clear link between the education decisions of the disadvantaged group and employer naivete. Interestingly, this effect is mainly driven by medium-ability red workers, who pursue education 21.5 percentage points less often in the Naive pool ($p = 0.006$).[14] High-ability red workers pursue education 1.3 percentage points less often in the Naive pool, but this effect is far from statistically significant ($p = 0.823$).

Green workers seem to be less affected by their pool. Column (2) in Table 2 shows that green workers are 2.7 percentage points more likely to pursue education when they are in a Naive pool, although the estimate is not statistically significant ($p = 0.128$). A potential explanation for this smaller effect is the already high baseline education rate among green workers (95%). Most of these workers probably learned early on that if they pursued education they were very likely to be hired and few reconsidered their decision after round 30.

In sum, our lab experiment shows that—in line with conservatism—employers hire disadvantaged workers less often than a Bayesian would. As a result, disadvantaged groups realize that education may not be worthwhile for them and they pursue education less often.

4 Experiment II

Our lab experiment identifies a causal link between excess discrimination due to non-Bayesian updating and under-education by workers of a disadvantaged group. While Experiment I was focused on establishing this causal link, Experiment II focuses in much more detail on the belief updating process and on how conservatism in updating leads to excess discrimination. Experiment II involves an online experiment with 515 participants who have to evaluate the performance of male and female candidates (with otherwise equal characteristics) in a math and logic test. After the evaluation, participants decide whether to hire each candidate or not. This setting allows us to study *how* people differ from Bayesian rationality when they update beliefs about others and whether such updating affects their hiring decisions. It also allows us to explore whether people update differently based on the candidate's identity.

4.1 Design

Candidates. We selected candidates from 93 participants from the Essex Lab subject pool at the University of Essex in March 2020. These 93 participants had completed a *math and logic test* consisting of 25 questions and filled in a demographics survey, including their first name, age, gender, region, marital status, and field of study at the university.[15] They all had given

[14]Since workers were randomly assigned to either of the pools, these effects can solely be attributed to the type of employers that workers expect to face. However, one may wonder whether workers update their beliefs immediately or whether they need several periods to learn. Figure D.1 shows the proportion of workers choosing to pursue education in each period. The figure shows that workers need several periods to update their beliefs, and the pool differences only become clear after period 38.

[15]Among the 93 candidates a majority (67%) were men, fell into the 18-25 years age bracket (87%) and most studied Social Sciences, Computer Science, Biology or Math. Most of them were single and lived in the East of

us consent to use their first name, test score, and some basic demographics in future online studies. We subsequently picked 80 out of these candidates at random and created four groups of 20 candidates each. We coded the performance of each candidate as being either in the top half or in the bottom half of the group.

We then asked a sample of 217 subjects on Prolific, a UK-based online survey panel, to tell us their perceptions associated with the names of 32 of the 80 candidates from the Essex Lab sample in terms of country of origin, ethnic group, religion, social class, intelligence and likability.[16] The reason to do this was to avoid selecting candidates for the main experiment who are perceived very differently in terms of characteristics other than gender. Each participant provided perceptions only about sixteen candidates. Which sixteen candidates they assessed was determined randomly. Participants in this part of the experiment were paid a fixed amount of 1.50 GBP for a survey that took around 10 minutes.

We picked four pairs of one male and one female candidate with equal characteristics (this means that for each female with a given age, region, marital status, field of studies, and whether she had been a top performer, we picked a male with the exact same characteristics) and similar perceptions.[17]

Experiment. At the beginning of the experiment, we explained the context of the math and logic test and showed participants three sample questions from the test. Since our framework requires that evaluators have different priors for the two identity groups, we also showed them some information about the characteristics of "top performers". In particular, we showed them the percentage of the (original eighty) candidates who were top performers by their "field of studies" at the university, depending on whether English was their native language, and by gender. The salient element was gender, with 65 percent of top performers being men.[18] We expected that this information would induce heterogenous priors. Since everything else in the experiment is symmetric across genders, differing priors are crucial to allow us to test whether non-Bayesian statistical discrimination leads to bigger differences in posteriors than Bayesian updating would.

Participants in the experiment then made decisions on four randomly selected out of the eight candidates. Participants were told that they will evaluate four out of a total of eighty candidates, but were not given any information on how these four are selected. For each can-

England.

[16]We picked these 32 candidates such that they were relatively similar in terms of self-reported characteristics and performance. We, for example, dropped candidates who were much older than the rest, who had very bad performance, or who had a very rare field of study.

[17]The candidates were Julia and Liam for group 1, Becky and Joseph for group 2, Anna and Alan for group 3, and Megan and Matthew for group 4. The prolific sample rated all candidates to be likely born in the UK, white, and Christian. They assessed some names to sound more working class than others, but on average they perceived them similarly across genders. Intelligence was perceived to be similar across genders (5.71 for women and 5.69 for men) as well as likability (6.09 for women and 6.01 for men). Appendix Table C.2 summarizes these perceptions.

[18]Appendix Figure D.2 shows how this information was presented to participants.

CV of Julia	
Age	18-21
Gender	Female
Region	East of England
Marital status	Single
Field of studies	Social Sciences
Signal 1	Positive
Signal 2	Negative
Signal 3	Positive
Signal 4	Positive
Signal 5	Positive

Figure 5. Example of candidate information after five signals have been received.

didate, they first saw the candidate's age, gender, region of residence, marital status, and field of study. They then gave us a prior belief on the probability with which they believed that the candidate was a top performer. We use the cross-over rule (or stochastic BDM) to incentivize belief elicitation (Karni, 2009; Burfurd and Wilkening, 2018). Specifically, if a participant's stated belief exceeds a randomly drawn number $X \in \{0, ..., 100\}$, then they get paid 5 GBP if the candidate assessed is a top performer and 0 GBP otherwise. If their stated belief is below X, then they get 5 GBP with probability $X\%$ and 0 GBP otherwise. Participants were told simply that it is in their best interest to state their true beliefs, but they were able to click on a link to obtain precise information on the payment mechanism.

We build on Mobius et al. (ming)'s experimental paradigm to study updating. After reporting their prior, participants received a signal. The signal was positive with 70% probability if the candidate was a top performer and with 30% probability if the candidate was *not* a top performer. After receiving the signal, they gave us their updated belief (posterior) using the cross-over rule as described above. They received four more signals and each time gave us a posterior, which means we observe five rounds of updating in total.

Participants evaluated candidates consecutively: they observed one candidate, indicated their prior, indicated their five posteriors after observing the signals, and only then they moved to evaluating the following candidate. Figure 5 shows an example of candidate information provided after five signals had been received.

After participants had evaluated the four candidates, they were surprised with a new part of the experiment (of which they had no information before-hand) in which they had to decide whether to hire each of the candidates they reviewed (presented in random order). Participants were endowed with 2.5 GBP and asked whether they wanted to hire each of the candidates. If they hired the candidate, they paid 2.5 GBP to the candidate and received 5 GBP if the candidate was a top performer and 0 GBP otherwise. If they did not hire the candidate, they kept their

2.5 GBP and the candidate got nothing. We chose to actually pay the candidates hired in order to make the hiring decision consequential for the candidates and to give more room for employer tastes to potentially affect the decision. This will allow us to see if non-Bayesian statistical discrimination is still important once room is given to tastes or social preferences to potentially affect the decision.[19]

This design allows us to distinguish statistical discrimination in hiring based on differences in the prior (and subsequent Bayesian updating) from non-Bayesian statistical discrimination which is caused by updating errors. The hiring decision also allows us to see whether there is taste-based discrimination, which we will define as discrimination that cannot be explained by differences in the posterior. Last, by studying whether updating errors differ across the gender of candidates, we can study whether there is a possible interaction between tastes and non-Bayesian statistical discrimination.

At the end of the experiment, we elicited a number of demographic characteristics in a post-experimental questionnaire. We also asked participants about their perception of their own intelligence and their perceptions regarding gender discrimination.

Procedures. Our experiment was coded in Qualtrics and fielded using the platform Prolific.[20] We restricted participants to be UK residents, as our candidates are UK residents and because we elicited name perceptions and stereotypes from a sample of UK residents. We fielded the experiment in six different waves across two days in September 2020. We fielded different waves in order to avoid selection effects based on time of day. Participants were paid a 1 GBP fixed fee for participating and in addition received the payment for one randomly selected round (either a belief or hiring round) in the experiment. The reason for only paying one randomly selected round (rather than e.g. all rounds) is to avoid giving participants' incentives to hedge across different decisions. The experiment received ethical approval from the University of Essex Faculty of Social Sciences subcommittee with number ETH1920-1029.

Sample. The sample size consists of 515 participants in the role of evaluators. Table 3 shows some characteristics of our sample. While women are somewhat over-represented among our respondents, we have good variation in age, self-reported social class, and education levels. Most of our respondents self-classify as White British and almost all have English as a first language.[21]

[19] There is some evidence, for example, for gender differences in dictator giving (Ben-Ner et al., 2004).

[20] One advantage of using this platform is that it allows us to access a more diverse sample of participants than a typical university lab sample. Recently there has been some discussion on whether online platforms tend to solicit socially desirable responses (Kellar and Hall, 2022). We agree that this could be a potential issue and would lead to an underestimate of the amount of discrimination in our study.

[21] The sample size for the three parts (math and logic test, name perceptions and main experiment) were determined as follows. The sample of 93 names comes from the fact that 93 participants from the first part gave us consent to use their first name and score in the subsequent parts. It is hence endogenously determined from a bigger group of participants who did the math and logic test. The second and third part were conducted on prolific and we had requested 200 and 500 participants, respectively, from the platform. The slightly higher sam-

Table 3. Sample characteristics

Demographics		*Ethnicity*		*Education*	
Female	0.583	White British	0.823	Post-graduate	0.173
Age	35.04	Other White	0.035	University	0.490
English 1st language	0.959	Black	0.037	Higher (A-levels, BTEC..)	0.258
Working Class	0.463	East Asian	0.010	Secondary School	0.079
Middle Class	0.530	South Asian	0.039	Primary school	0.020

4.2 Results

We now discuss our main results. We first ask whether there are departures from Bayesian rationality in updating. We then study hiring decisions and to which extent discrimination in hiring can be traced back to Bayesian, non-Bayesian, or taste-based discrimination.

4.2.1 Updating

We first have a brief look at differences in priors. On average, subjects initially believe that the probability that men are top performers is 52.05%, while their belief is 42.93% for women. Appendix Figure D.3 shows the distribution of prior beliefs. The figure shows that priors for men are clearly higher than those for women in the sense of first-order stochastic dominance. The figure also shows that there is a substantial degree of heterogeneity in priors. For both candidate genders priors cover almost the entire range from 0-100.

To study updating, we follow the approach by Mobius et al. (ming) (see also Grether 1992 El-Gamal and Grether 1992, Buser et al. 2018, Coutts 2019, or Augenblick and Rabin 2021) and write Bayes rule in terms of a logistic function as follows

$$\ln(\frac{p_i^t}{1-p_i^t}) = \delta_i \ln(\frac{p_i^{t-1}}{1-p_i^{t-1}}) + \beta_{i,l}\lambda_l \mathbf{1}_{i,neg}^{t-1} + \beta_{i,h}\lambda_h \mathbf{1}_{i,pos}^{t-1} + \epsilon_i,$$

where p_i^t is participant i's stated belief in round t, $\mathbf{1}_{i,neg}^{t-1}$ is a dummy indicating that i received a negative signal in round $t-1$, and $\mathbf{1}_{i,pos}^{t-1}$ is the corresponding dummy indicating that the signal was positive. λ_h and λ_l are the log likelihood ratios of a positive (h) or negative (l) signal, respectively. In our experiment $\lambda_h = -\lambda_l = \ln(\frac{7}{3})$. The parameter δ_i relates to a property called invariance (see Mobius et al. ming and Augenblick and Rabin 2021). Of key interest for us are the parameters $\beta_{i,l}$ and $\beta_{i,h}$, which capture the extent of updating after a negative or a positive signal, respectively. For a Bayesian we should have $\beta_{i,l} = \beta_{i,h} = 1$, and there is evidence of conservatism when $\beta_{i,l}, \beta_{i,h} < 1$.

Table 4 shows the results of estimating this equation. Columns (1)-(3) use all data and in columns (4)-(6) we drop candidate evaluations in which participants updated in the wrong direction at least once, in the sense that their posterior decreased after receiving a positive signal or vice versa. Mistakes in which participants update in the wrong direction could seriously

ple sizes are due to over-recruitment by prolific (which can happen e.g. if some responses time out or if more than expected participants start the survey within a given time interval). We did not do a power analysis to determine the planned sample sizes of 200 and 500. This is due to the fact that, for both parts, we did not have one specific hypothesis in mind that we wanted to test neither for the name perceptions nor for the implications of conservatism.

Table 4. Updating in Experiment II

	All data			No mistakes		
	(1) All	(2) Male	(3) Female	(4) All	(5) Male	(6) Female
δ	0.843	0.845	0.841	0.907	0.921	0.895
	(0.012)	(0.015)	(0.015)	(0.010)	(0.013)	(0.013)
β_l	0.603	0.586	0.621	0.757	0.738	0.781
	(0.019)	(0.023)	(0.023)	(0.020)	(0.022)	(0.024)
β_h	0.567	0.555	0.578	0.694	0.669	0.713
	(0.017)	(0.021)	(0.021)	(0.018)	(0.022)	(0.022)
$p(\delta = 1)$	0.000	0.000	0.000	0.000	0.000	0.000
$p(\beta_l = 1)$	0.000	0.000	0.000	0.000	0.000	0.000
$p(\beta_h = 1)$	0.000	0.000	0.000	0.000	0.000	0.000
$p(\beta_l = \beta_h)$	0.027	0.159	0.053	0.000	0.001	0.003

Note: Standard errors are clustered at the participant level.* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

bias the results. We hence follow the literature and restrict our main analysis to the candidate evaluations where there are no mistakes (see, e.g., Mobius et al. ming, Coutts 2019, Barron 2020, Erkal et al. 2021, Kogan et al. 2021).[22]

The results clearly show conservatism, rejecting both hypotheses $\beta_l = 1$ and $\beta_h = 1$ ($p < 0.001$). The deviations from Bayesian posteriors are substantial with participants losing on average ≈ 0.60 GBP by not indicating a Bayesian posterior and there is considerable heterogeneity in this number. We also find some evidence of asymmetric updating with participants paying more attention to negative than positive signals, i.e. $\beta_l > \beta_h$.[23] Columns (2), (3), (5), and (6) show the results when we focus only on updating for male and female candidates, respectively. Comparing these estimates allows us to study whether updating is different depending on the candidate's gender. We find that participants are slightly more conservative (smaller β_l, β_h) when evaluating men than when evaluating women (p-value$= 0.094$ when comparing β_l between men and women, and p-value$= 0.064$ when comparing β_h between men and women). The difference $\beta_l - \beta_h$ is, however, virtually identical across candidate genders. Participants do not seem to update more "positively" or "negatively" depending on the candidate's gender on average. This is in line with the idea that, once stereotypes are held constant, gender *per se* does not matter that much (Coffman et al., 2020).[24]

Figure 6 gives us a first idea of the implications of such conservatism. The solid lines show

[22] In total, 10.52% of the updating decisions go in the wrong direction and 17.5% of the candidate evaluations have at least one mistake, i.e. update in the wrong direction (these numbers are in line with what is typically found in the literature, see e.g. Mobius et al. ming, Coutts 2019, Barron 2020, Erkal et al. 2021, and Kogan et al. 2021). Importantly, we do not find differences in wrong updating for female or male candidates, indicating that in our case wrong updating is likely to be noise.

[23] Appendix Figure D.4 shows that deviations from Bayesian rationality occur across the entire range of priors.

[24] When stereotypes are not held constant, however, there has been evidence that people, specifically physicians, update beliefs about female surgeons more negatively than for male surgeons (Sarsons, 2021).

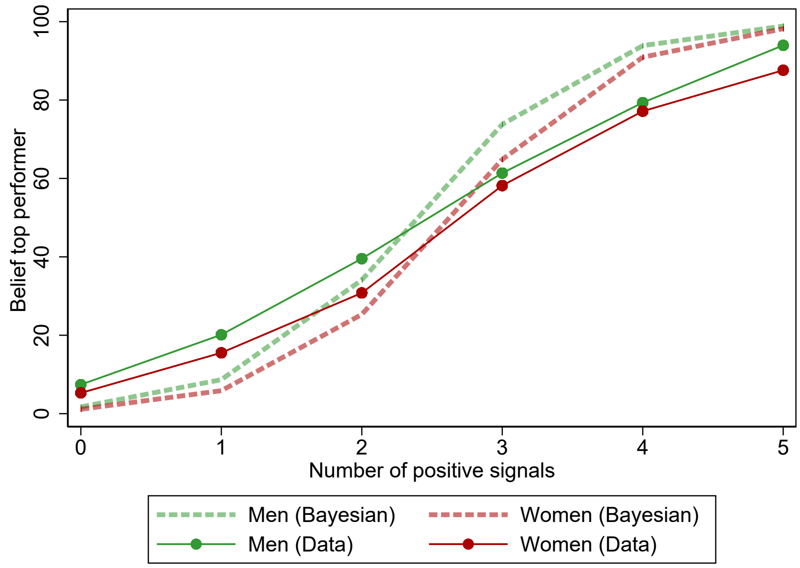


Figure 6. Mean beliefs that a candidate is a top performer after five rounds of updating depending on the number of positive signals.

the average posterior—i.e., belief after five rounds of updating—for male and female candidates depending on the number of positive signals received. These are contrasted with the Bayesian posterior for each case—i.e., the posterior that a Bayesian participant would have after all the observed signals, given the reported priors. As we would expect with conservatism, the implications of non-Bayesian statistical discrimination are largest when the sequence of signals is very informative (e.g., five positive or five negative signals). In this case, there is almost no difference in Bayesian posteriors, as the information contained in the signals largely eliminates the difference in priors. However, in line with conservatism, we see that there remains a substantial difference in participants' posteriors.[25]

Table 5 uses OLS regressions to study whether participants' posteriors do indeed exaggerate the difference between men and women compared to the Bayesian benchmark. We regress participants' actual posteriors after five rounds of updating on the Bayesian posterior and a dummy indicating whether the candidate is female. The Bayesian posterior is calculated for each participant based on the participant's subjective prior assuming correct updating after each signal. Columns (A) and (B) show that there remains a gender gap in posteriors which cannot be explained by rational statistical discrimination. Columns (0)-(5) then show the same regression where we split the sample depending on the number of positive signals received (between 0 and 5). In line with the intuition developed above, the unexplained gender gap (that cannot be explained by Bayesian updating) is especially large in Columns (0) and (5), when signals are very informative. For five positive signals, the average posterior is $\approx 91\%$ for men and $\approx 85\%$ for women. By contrast, the average Bayesian posterior in this case is $\approx 98.8\%$ for men and $\approx 98.2\%$ for women. When signals are relatively uninformative, as in Columns

[25]It should be noted that these averages hide a substantial degree of heterogeneity with the difference in posteriors even larger among the most conservative participants.

Table 5. Belief updating in Experiment II

	All data		Number of positive signals					
	(A)	(B)	(0)	(1)	(2)	(3)	(4)	(5)
Bayesian posterior	0.517***	0.516***	-0.031	0.198	0.639***	0.561***	0.981***	1.405**
	(0.030)	(0.031)	(0.076)	(0.158)	(0.063)	(0.079)	(0.192)	(0.544)
Female	-1.367*	-1.375*	-4.047*	-5.498**	-1.504	1.157	1.983	-5.894**
	(0.824)	(0.835)	(2.235)	(2.329)	(1.900)	(1.860)	(1.750)	(2.828)
Observations	1400	1358	108	233	293	314	298	112
Mean outcome	49.477	49.704	5.972	18.047	35.113	59.592	78.560	91.402
All controls		X	X	X	X	X	X	X

Note: Belief updating based on the candidate's gender and the correct Bayesian posterior. Column (A) controls for the number of positive signals. *All controls* controls for all the subjects' answers in the final questionnaire. More specifically, it includes fixed effects for subjects': gender, age (in 15-year bands), ethnicity, nationality, self-reported social class, whether their first language is English, self-assessed intelligence, and perceptions regarding gender equality. They also control for wave fixed effects. Standard errors clustered at the participant level in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

(2) and (3), the unexplained gender gap is much smaller and not statistically significant.

4.2.2 Hiring

We now study how these updating mistakes translate into hiring decisions. Figure 7 shows the proportion of candidates hired depending on the number of positive signals (with its corresponding 95% confidence interval). The figure shows that candidates with fewer than three positive signals are rarely hired, candidates with three positive signals are hired about half of the time, and candidates with four or five positive signals are hired frequently. We also see that men are hired more frequently than women for any number of signals. In line with the intuition developed above (Table 5), the gender gap in hiring is biggest for candidates with five positive signals. Here, 96% of men are hired as opposed to only 81% of women.

Table 6 studies the gender gap in hiring more formally. Columns (1)-(2) show that on average women are about 0.069 p.p. (18%) less likely to be hired than men. In Columns (3)-(4) we control for prior beliefs. Doing so does not reduce the gender gap substantially and the coefficient on prior beliefs is small and statistically not significant. Controlling for participants' posteriors (Columns (5)-(6)), however, substantially reduces the gender gap (test $\beta(2) = \beta(6)$, $p = 0.018$).[26] In fact, the gender gap is no longer statistically different from zero once posteriors are controlled for. This shows that beliefs matter and that at most a small share of the overall gap is taste-based (unexplainable by beliefs). Last, in Columns (7)-(8) we include both participants' actual posterior and the Bayesian posterior. This reduces the gender gap further. A back of the envelope calculation using these regression results allows us to decompose the overall hiring gap. This exercise suggests that around $22\%(=\frac{-0.015}{-0.069})$ of the overall gap is un-

[26] Figure D.5 in the Online Appendix shows the proportion of participants who hire male and female candidates by each prior and posterior.

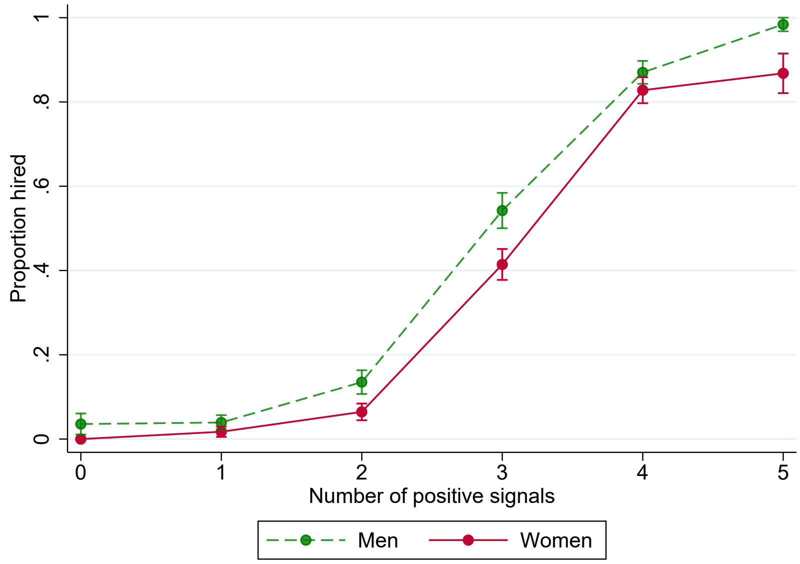


Figure 7. Hiring decisions in Experiment II

explained discrimination (potentially taste-based), 39% is Bayesian statistical discrimination and another 39% is non-Bayesian statistical discrimination.[27]

These findings also highlight that the mechanism by which people arrive at "wrong beliefs" matters. While there are gender differences in initial beliefs, those cannot explain the hiring gap. Most priors lie in a range (20-75%) where candidates would not be hired anyhow and the variation presented within this range is largely unrelated to candidate quality.[28] The hiring gap arises because there is insufficient updating for those with many positive signals, i.e. for those who might be considered "good enough" to be hired in the experiment (see Table 5). Indeed, when we re-run regression (3) restricted to those with at least four positive signals, we find that the coefficient on the prior becomes highly statistically significant ($p < 0.001$) and the gender gap reduces to about -0.016 p.p. Hence, while differences in priors are clearly important in generating differences in posteriors, these differences by themselves are not always sufficient to generate a gender gap. The reason is that not every difference in the prior distribution will generate a difference in posteriors *in the relevant range* where employers consider whether to hire or not. While the relevant variation in this setting is among those with very positive signals, it is important to notice that in other contexts it may be among those with average or negative signals. To be able to predict heterogeneous outcomes, it is hence essential to

[27]We investigated several dimensions of heterogeneity. Splitting the sample by education, for example, shows that the gender gap is somewhat bigger (0.073 p.p.) for those with less education than for those with higher education (0.059 p.p.). However, the decomposition in terms of the share explained by Bayesian and non-Bayesian statistical discrimination remains broadly stable. Interestingly, while the hiring gap is similar for female respondents (0.069 p.p.) and for male respondents (0.099 p.p.), once we control for posteriors the gap vanishes for females (0.015 p.p.), while it remains relatively high for males (0.067 p.p.). While this result is relatively noisy and we do not have power for further exploration, it could point towards tastes playing a larger role for male respondents.

[28]Using CRRA utility x^{γ} with risk aversion (concavity) parameter $\gamma = 0.25$ ($\gamma = 0.5$), beliefs should exceed 70% (84%) for candidates to be hired.

Table 6. Hiring in Experiment II

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Female (β)	-0.071***	-0.069***	-0.062**	-0.057**	-0.026	-0.023	-0.018	-0.015
	(0.025)	(0.026)	(0.026)	(0.027)	(0.019)	(0.019)	(0.018)	(0.019)
Prior			0.001	0.001				
			(0.001)	(0.001)				
Posterior					0.011***	0.011***	0.005***	0.005***
					(0.000)	(0.000)	(0.001)	(0.001)
Bayesian posterior							0.005***	0.005***
							(0.001)	(0.001)
p-value			0.717	0.672	0.019	0.018	0.004	0.004
Observations	1400	1358	1400	1358	1400	1358	1400	1358
Mean outcome	0.398	0.402	0.398	0.402	0.398	0.402	0.398	0.402
All controls		X		X		X		X

Note: Hiring decisions across all rounds based on each participant's gender and beliefs. *All controls* controls for all the subjects' answers in the final questionnaire. More specifically, it includes fixed effects on subjects': gender, age (in 15-year bands), ethnicity, nationality, self-reported social class, whether their first language is English, self-assessed intelligence, and perceptions regarding gender equality. They also control for wave fixed effects. The p-value row compares the β coefficient in each column with the β coefficient in the first two columns. More concretely, p-value in column (x) tests whether $\beta(x) = \beta(1)$ for odd columns and $\beta(x) = \beta(2)$ for even columns. Standard errors clustered at the participant level in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

understand the mechanism by which people arrive at wrong posteriors.

5 Conclusions

This paper studies the implications of statistical discrimination when employers may not update their beliefs rationally. Using theory and two different experiments, we show that employers who suffer from conservatism discriminate against disadvantaged groups more often than Bayesian employers, especially when signals are highly informative. Such non-Bayesian discrimination then reduces the willingness of high-ability workers from the disadvantaged group to pursue education, since they expect that their education signal will not be sufficiently considered by prospective employers.

Understanding the source of discrimination is important because the policy implications of taste-based, Bayesian, and non-Bayesian statistical discrimination are very different. For example, providing employers with information (e.g. about the characteristics of graduates from different groups) might be effective against rational statistical discrimination, but the same information may not help reduce non-Bayesian statistical discrimination as it might not be properly interpreted by naive employers. In fact, numerous studies show that holding better information is often not associated with diminished stereotypes or discrimination (e.g., Johnston and Macrae 1994, Crisp et al. 2005, Dumesnil and Verger 2009, Delavande and Zafar 2018). Oreopoulos (2011), for example, finds that listing language fluency on a CV does not lead to less discrimination against immigrants in hiring, despite the fact that recruiters justify their behaviour by concerns about language skills.[29] Delavande and Zafar (2018) find that anti-american attitudes persist as respondents do not update sufficiently after receiving new information.

A second example is that of affirmative action. While some affirmative action policies might backfire under Bayesian statistical discrimination (Coate and Loury 1993; Moro and Norman 2003; Fang and Norman 2006), they are likely to be effective against non-Bayesian statistical discrimination. For example, a policy that improves access to education for disadvantaged groups (e.g. via subsidies or admission quotas) may make disadvantaged workers less attractive for Bayesian employers since it makes their education signal less informative. By contrast, such policies might be effective against non-Bayesian statistical discrimination. First, they can address the issue of "under-education" of the disadvantaged group by inducing more people to seek education who should be doing so and would be doing so if employers were Bayesian. Second, by reducing market exit, such policies can allow employers to learn to make improved inference from information (Beaman et al., 2009; Niederle et al., 2013).

Clearly, discrimination can have different sources which might be present at the same time and sometimes interact. While conservatism seems to be the main mechanism leading to wrong beliefs in our experiments, different mechanisms may also matter in other contexts. Stereotypes may lead to inaccurate beliefs when one characteristic is clearly over-represented in a group (Bordalo et al., 2016; Bohren et al., 2021). Limited attention may lead to different information acquisition for different identities (Bartoš et al. 2016). There is also experimental evidence of other biases, such as cursedness, selection neglect, failures of contingent reasoning, or analogy based reasoning (Grimm and Mengel, 2012; Martinez-Marquina et al., 2019; Barron et al., 2021), though these have not yet been linked to discrimination. We conjectured that in the context of discrimination, where priors are often linked to strongly entrenched stereotypes that in some cases have been formed over generations (Massey and Denton, 1993; Cutler et al., 1999; Telles and Ortiz, 2008), conservatism in updating might be a particularly important bias. And indeed we have seen that it plays an important role in generating discrimination. Future research could study when we should expect other mechanisms to matter as well and whether and when there is any interplay between them. Ultimately, we hope that this work will inform us about how to best design policy interventions to reduce discrimination.

[29]It should also be noted, though, that employer naivete is not the only conceivable reason why employers do not react sufficiently to information. In practice it is likely that a number of factors contribute to this under-reaction.

References

Agan, A. and Starr, S. (2018). Ban the box, criminal records, and racial discrimination: A field experiment. *The Quarterly Journal of Economics*, 133(1):191–235.

Aigner, D. J. and Cain, G. G. (1977). Statistical theories of discrimination in labor markets. *Industrial and Labor Relations Review*, 30(2):175–187.

Albrecht, K., Von Essen, E., Parys, J., and Szech, N. (2013). Updating, self-confidence, and discrimination. *European Economic Review*, 60:144–169.

Altonji, J. and Pierret, C. (2001). Employer learning and statistical discrimination. *The Quarterly Journal of Economics*, 116(1):313–350.

Arnold, D., Dobbie, W., and Yang, C. S. (2018). Racial bias in bail decisions. *The Quarterly Journal of Economics*, 133(4):1885–1932.

Arrow, K. (1973). The theory of discrimination. In *O. Ashenfelter and A. Rees (Eds.), Discrimination in Labor Markets*. Princeton University Press.

Augenblick, N. and Rabin, M. (2021). Belief Movement, Uncertainty Reduction, and Rational Updating*. *The Quarterly Journal of Economics*, 136(2):933–985.

Autor, D. H. and Scarborough, D. (2008). Does job testing harm minority workers? evidence from retail establishments. *The Quarterly Journal of Economics*, 123(1):219–277.

Barron, K. (2020). Belief updating: does the 'good-news, bad-news' asymmetry extend to purely financial domains? *Experimental Economics*, pages 1–28.

Barron, K., Ditlmann, R., Gehrig, S., and Schweighofer-Kodritsch, S. (2020). Explicit and implicit belief-based gender discrimination: A hiring experiment. Technical report, WZB Discussion Paper.

Barron, K., Huck, S., and Jehiel, P. (2021). Everyday econometricians: Selection neglect and overoptimism when learning from others. *working paper*.

Bartoš, V., Bauer, M., Chytilová, J., and Matějka, F. (2016). Attention discrimination: Theory and field experiments with monitoring information acquisition. *American Economic Review*, 106(6):1437–75.

Beaman, L., Chattopadhyay, R., Duflo, E., Pande, R., and Topalova, P. (2009). Powerful women: Does exposure reduce bias? *The Quarterly Journal of Economics*, 124(4):1497–1540.

Becker, G. S. (1957). *The economics of discrimination*. University of Chicago press.

Ben-Ner, A., Kong, F., and Putterman, L. (2004). Share and share alike? gender-pairing, personality, and cognitive ability as determinants of giving. *Journal of Economic Psychology*, 25(5):581–589.

Benjamin, D. J. (2019). Errors in probabilistic reasoning and judgment biases. In *Handbook of Behavioral Economics: Applications and Foundations 1*, volume 2, pages 69–186. Elsevier.

Benjamin, D. J., Rabin, M., and Raymond, C. (2016). A model of nonbelief in the law of large numbers. *Journal of the European Economic Association*, 14(2):515–544.

Bertrand, M., Chugh, D., and Mullainathan, S. (2005). Implicit discrimination. *American Economic Review*, 95(2):94–98.

Bertrand, M. and Duflo, E. (2017). Field experiments on discrimination. In *Handbook of economic field experiments*, volume 1, pages 309–393. Elsevier.

Bock, O., Baetge, I., and Nicklisch, A. (2014). hroot: Hamburg registration and organization online tool. *European Economic Review*, 71:117–120.

Bohren, A., Haggag, K., Imas, A., and Pope, D. (2021). Inaccurate statistical discrimination: An identification problem. *NBER working paper*, 25935.

Bohren, A., Imas, A., and Rosenberg, M. (2018). The dynamics of discrimination: Theory and evidence. *American Economic Review*, 109:3395–3436.

Bordalo, P., Coffman, K., Gennaioli, N., and Shleifer, A. (2016). Stereotypes. *The Quarterly Journal of Economics*, 131(4):1753–1794.

Burfurd, I. and Wilkening, T. (2018). Experimental guidance for eliciting beliefs with the stochastic becker–degroot–marschak mechanism. *Journal of the Economic Science Association*, 4(1):15–28.

Buser, T., Gerhards, L., and van der Weele, J. (2018). Responsiveness to feedback as a personal trait. *Journal of Risk and Uncertainty*, 56:165–192.

Ceci, S. J., Ginther, D. K., Kahn, S., and Williams, W. M. (2014). Women in academic science: A changing landscape. *Psychological Science in the Public Interest*, 15(3):75–141.

Charles, K. K. and Guryan, J. (2008). Prejudice and wages: an empirical assessment of becker's the economics of discrimination. *Journal of political economy*, 116(5):773–809.

Charles, K. K. and Guryan, J. (2011). Studying discrimination: Fundamental challenges and recent progress. *Annu. Rev. Econ.*, 3(1):479–511.

Chen, Y. and Li, S. (2009). Group identity and social preferences. *American Economic Review*, 99(1):431–457.

Chen, Y. and Mengel, F. (2016). Social identity and discrimination: Introduction to the special issue. *European Economic Review*, 90:1–3.

Coate, S. and Loury, G. C. (1993). Will affirmative-action policies eliminate negative stereotypes? *American Economic Review*, pages 1220–1240.

Coffman, K., Collis, M., and Kulkarni, L. (2021). Stereotypes and belief updating. *mimeo*.

Coffman, K., Exley, C., and Niederle, M. (2020). The role of beliefs in driving gender discrimination. *Management Science*, in press.

Coutts, A. (2019). Good news and bad news are still news: Experimental evidence on belief updating. *Experimental Economics*, 22(2):369–395.

Crisp, A., Gelder, M., Goddard, E., and Meltzer, H. (2005). Stigmatization of people with mental illnesses: a follow-up study within the changing minds campaign of the royal college of psychiatrists. *World psychiatry*, 4(2):106.

Cutler, D., Glaeser, E., and Vigdor, J. (1999). The rise and decline of the american ghetto. *Journal of Political Economy*, 107:455–506.

Delavande, A. and Zafar, B. (2018). Information and anti-american attitudes. *Journal of Economic Behavior and Organization*.

Dumesnil, H. and Verger, P. (2009). Public awareness campaigns about depression and suicide: a review. *Psychiatric Services*, 60(9):1203–1213.

Edwards, W. (1968). Conservatism in human information processing. *Formal representation of human judgment*.

Eil, D. and Rao, J. M. (2011). The good news-bad news effect: asymmetric processing of objective information about yourself. *American Economic Journal: Microeconomics*, 3(2):114–38.

El-Gamal, M. A. and Grether, D. (1992). Are people bayesian? uncovering behavioural strategies. *Journal of the American Statistical Association*.

Erev, I., Shimonowitch, D., Schurr, A., and Hertwig, R. (2008). Base rates: how to make the intuitive mind appreciate or neglect them. In Plessner, H., Betsch, C., and Betsch, T., editors, *Intuition in Judgement and Decision-Making*, pages 135–148. Lawrence Erlbaum, NJ.

Erkal, N., Gangadharan, L., and Koh, B. H. (2021). By chance or by choice? biased attribution of others' outcomes when social preferences matter. *Biased Attribution of Others' Outcomes when Social Preferences Matter (March 8, 2021).*

Esponda, I., Oprea, R., and Yuksel, S. (2022). Contrast-biased evaluation. *mimeo.*

Fang, H. and Moro, A. (2011). Theories of statistical discrimination and affirmative action: A survey. In *Handbook of Social Economics*, volume 1, chapter 5. Elsevier B.V.

Fang, H. and Norman, P. (2006). Government-mandated discriminatory policies: theory and evidence. *International Economic Review*, 47(2):361–389.

Fershtman, C. and Gneezy, U. (2001). Discrimination in a segmented society: An experimental approach. *The Quarterly Journal of Economics*, 116(1):351–377.

Fischbacher, U. (2007). z-tree: Zurich toolbox for ready-made economic experiments. *Experimental economics*, 10(2):171–178.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, 19(4):25–42.

Grether, D. (1992). Testing bayes rule and the representativeness heuristic: some experimental evidence. *Journal of Economic Behavior and Organization*, 17(1):31–57.

Grether, D. M. (1980). Bayes rule as a descriptive model: The representativeness heuristic. *The Quarterly journal of economics*, 95(3):537–557.

Grimm, V. and Mengel, F. (2012). An experiment on learning in a multiple games environment. *Journal of Economic Theory*.

Johnston, L. C. and Macrae, C. N. (1994). Changing social stereotypes: The case of the information seeker. *European Journal of Social Psychology*, 24(5):581–592.

Kahneman, D. and Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive psychology*, 3(3):430–454.

Kahnemann, D. and Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, 80(4).

Karni, E. (2009). A mechanism for eliciting probabilities. *Econometrica*, 77(2):603–606.

Kellar, S. and Hall, E. V. (2022). Measuring racial discrimination remotely: A contemporary review of unobtrusive measures. *Perspectives on Psychological Sciences*, 17(5):1404–1430.

Knowles, J., Persico, N., and Todd, P. (2001). Racial bias in motor vehicle searches: Theory and evidence. *Journal of Political Economy*, 109(1):203–229.

Kogan, S., Schneider, F., and Weber, R. A. (2021). Self-serving biases in beliefs about collective outcomes. *University of Zurich, Department of Economics, Working Paper*, (379).

Lang, K. and Lehmann, J.-Y. K. (2012). Racial discrimination in the labor market: Theory and empirics. *Journal of Economic Literature*, 50(4):959–1006.

Lang, K. and Manove, M. (2011). Education and labor market discrimination. *American Economic Review*, 101:1467–1496.

Lippens, L., Baert, S., Ghekiere, A., Verhaeghe, P.-P., and Derous, E. (2020). Is labour market discrimination against ethnic minorities better explained by taste or statistics? a systematic review of the empirical evidence. *IZA Discussion paper*, (13523).

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

Lundberg, S. J. and Startz, R. (1983). Private discrimination and social intervention in competitive labor market. *American Economic Review*, 73(3):340–347.

Martinez-Marquina, A., Niederle, M., and Vespa, E. (2019). Failures in contingent reasoning: The role of uncertainty. *American Economic Review*, 109(10):3437–3474.

Massey, D. A. and Denton, N. (1993). *American Apartheid: Segregation and the Making of the Underclass*. Harvard University Press, Cambridge.

Mobius, M., Niederle, M., Niehaus, P., and Rosenblat, T. (forthcoming). Managing self-confidence. *Management Science*.

Mobius, M. M. and Rosenblat, T. S. (2006). Why beauty matters. *American Economic Review*, 96(1):222–235.

Moro, A. and Norman, P. (2003). Affirmative action in a competitive economy. *Journal of Public Economics*, 87(3-4):567–594.

Neumark, D. (2018). Experimental research on labor market discrimination. *Journal of Economic Literature*, 56(3):799–866.

Niederle, M., Segal, C., and Vesterlund, L. (2013). How costly is diversity? affirmative action in light of gender differences in competitiveness. *Management Science*.

Oreopoulos, P. (2011). Why do skilled immigrants struggle in the labor market? a field experiment with thirteen thousand resumes. *American Economic Journal: Economic Policy*, 3(4):148–171.

Peterson, C. R. and Beach, L. R. (1967). Man as an intuitive statistician. *Psychological bulletin*, 68(1):29.

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

Phillips, L. D. and Edwards, W. (1966). Conservatism in a simple probability inference task. *Journal of experimental psychology*, 72(3):346.

Reuben, E., Sapienza, P., and Zingales, L. (2014). How stereotypes impair women's careers in science. *Proceedings of the National Academy of Sciences*, 111(12):4403–4408.

Ruzzier, C. and Woo, M. (2022). Discrimination with inaccurate beliefs and confirmation bias. *Working Paper, Universidad de San Andres*.

Sarsons, H. (2021). Interpreting signals in the labor market: Evidence from medical referrals. *working paper*.

Spence, M. (1973). Job market signaling. *The Quarterly Journal of Economics*, 87(3):355–374.

Telles, E. and Ortiz, V. (2008). *Generations of Exclusion: Mexican Americans, Assimilation and Race*. Russel Sage Foundation.

Terrell, J., Kofink, A., Middleton, J., Rainear, C., Murphy-Hill, E., Parnin, C., and Stallings, J. (2017). Gender differences and bias in open source: Pull request acceptance of women versus men. *PeerJ Computer Science*, 3:e111.

Online Appendix "Irrational Statistical Discrimination"

Pol Campos-Mercade and Friederike Mengel

Contents

A Theoretical Results: Exogenous Case

A.1 Proof of Proposition 1

For the naive employer we have that $\pi_N(e,g) = g_l X_{e,l} + g_m X_{e,m} + g_h X_{e,h}$ and $\pi_N(e,r) = r_l X_{e,l} + r_m X_{e,m} + r_h X_{e,h}$. Given the assumption on payoffs we can replace $X_{e,m}$ by $X_{e,l} + \beta_e$ and $X_{e,h}$ by $X_{e,l} + 2\beta_e$. After these substitutions we get that $\pi_N(e,g) > \pi_N(e,r)$ whenever $g_m - r_m > 2(r_h - g_h)$, which is condition **(N)**. We also assumed that, as in Phelps (1972), the red group has on average lower ability with all ability levels carrying equal marginal weight (i.e. $(h-m) = (m-l)$). Noting that the average ability level of group i is given by $i_l l + i_m(l + (m-l)) + i_h(l + 2(m-l))$, we see that this assumption is equivalent to $g_m - r_m \geq 2(r_h - g_h)$ which is the same condition as condition **(N)**. The naive employer hence always discriminates against the red group.

For the Bayesian employer we derived that their expected payoff of hiring an educated worker of identity i is

$$\pi_B(e,i) = \frac{p_l i_l X_{e,l} + p_m i_m X_{e,m} + p_h i_h X_{e,h}}{p_l i_l + p_m i_m + p_h i_h}$$

or

$$\pi_B(e,i) = \frac{p_m i_m X_{e,m} + i_h(X_{e,m} + \beta_e)}{p_m i_m + i_h}.$$

Hence $\pi_B(e,r) > \pi_B(e,g)$ is equivalent to $X_{e,m} + \beta_e \frac{g_h}{p_m g_m + g_h} > X_{e,m} + \beta_e \frac{r_h}{p_m r_m + r_h}$ which is equivalent to condition **(B)**.

A.2 More general payoff schemes

We now consider a more general payoff scheme in which the employer's payoffs do not necessarily increase linearly in worker's ability. More concretely, we define $\alpha_1 \equiv X_{e,m} - X_{e,l}$ and $\alpha_2 \equiv X_{e,h} - X_{e,m}$, where $\alpha_i \in \mathbb{R}_+ \; \forall i \in \{1,2\}$ and $\beta_1 \equiv X_{ne,m} - X_{ne,l}$ and $\beta_2 \equiv X_{ne,h} - X_{ne,m}$, where $\beta_i \in \mathbb{R}_+ \; \forall i \in \{1,2\}$.

We show the results for the case of hiring educated workers. The analysis for non-educated workers is equivalent and yields similar results, as we will discuss below.

Table A.1 represents in which situations Naive and Bayesian employers prefer green or red workers for different assumptions on the ability distributions among the two groups. Note that the case $g_m < r_m$ and $g_h < r_h$ is ruled out by the assumption that green workers are on average better.

With this setting, we can prove that when $\alpha_1 \leq \alpha_2$ (i.e., the employer prefers to go from medium to high ability than from low to medium ability), Proposition 1 holds. We will discuss the case when $\alpha_1 > \alpha_2$ below.

Proposition A.1. *When $\alpha_1 \leq \alpha_2$, if bayesians discriminate against educated red workers, Naive employers will discriminate against educated red workers too.*

Proof. We will prove the proposition studying each of the parameter combinations considered in Table A.1.

1. *Case 1.* If $g_h > r_h$ and $g_m > r_m$ bayesians prefer green for some parameter combinations, while naives always prefer green.

Table A.1. Employer preferences for educated workers

	$g_h > r_h$	$g_h < r_h$
$g_m > r_m$	N: Green B: ?	N: ? B: Red
$g_m < r_m$	N: ? B: Green	- -

Note: This table represents what kind of educated workers naive and Bayesian employers prefer for different parameter combinations. "?" means that the choice depends on additional parameters.

2. *Case 2.* If $g_h < r_h$ and $g_m > r_m$ bayesians always prefer red and naives sometimes prefer red and sometimes green. So bayesians never prefer green workers.

3. *Case 3.* If $g_h > r_h$ and $g_m < r_m$ bayesians prefer green whenever $\frac{g_h}{g_m} > \frac{r_h}{r_m}$ and naives prefer green if $\frac{\alpha_1}{\alpha_1+\alpha_2} < \frac{r_h-g_h}{g_m-r_m}$. Note that $\frac{r_h-g_h}{g_m-r_m} = \frac{g_h-r_h}{r_m-g_m} \geq \frac{1}{2}$ since $g_m - r_m \geq 2(r_h - g_h) \iff r_m - g_m \leq 2(g_h - r_h)$ (that ensures that reds are on average worse than greens) must hold. Since we have assumed that $\alpha_1 \leq \alpha_2$, then $\frac{\alpha_1}{\alpha_1+\alpha_2} \leq \frac{1}{2}$ and hence naives will always prefer green workers (like bayesians do). $\square$

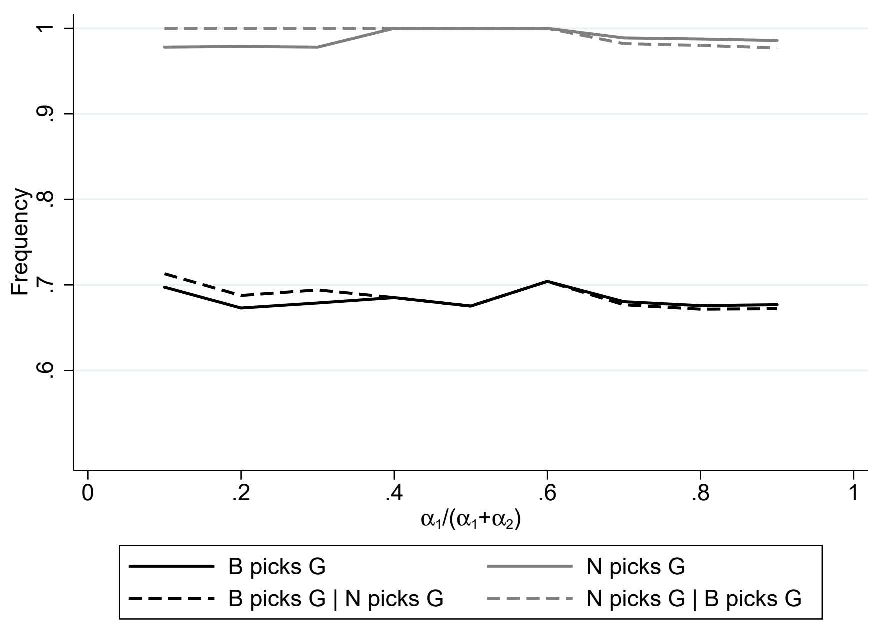


Figure A.1. Numerical simulations where Bayesian (B) and Naive (N) employers choose whether to hire green (G) or red (R) workers.

To understand whether the essence of our results (i.e., that Naive employers discriminate more against the disadvantaged group than Bayesian employers) also holds when $\alpha_1 > \alpha_2$, we perform numerical simulations. In these simulations, we assume that four green and four red workers are drawn with random ability (we pick four to mimic the lab experiment, but the same intuition holds regardless of the size of the urn), such that on average the green workers have higher ability than the red workers. These workers form the baseline distribution that employers know. We then study employers' hiring decision from very low $\frac{\alpha_1}{\alpha_1+\alpha_2}$ (implying increasing marginal returns to ability) to very high $\frac{\alpha_1}{\alpha_1+\alpha_2}$ (implying decreasing marginal returns to ability). When $\frac{\alpha_1}{\alpha_1+\alpha_2} \leq \frac{1}{2}$, we have the case above in which whenever naive employers hire red educated workers, Bayesian employers hire red educated workers as well. When $\frac{\alpha_1}{\alpha_1+\alpha_2} > \frac{1}{2}$, while this proposition is no longer true, we see that naive employers still hire green workers to a much larger extent than bayesians. Hence, even when $\alpha_1 > \alpha_2$ is not satisfied, the results are in line with what we expect: naive employers discriminate against the disadvantaged group much more often than bayesians.

The case for hiring of uneducated workers is equivalent to the one described above, with the difference that the equivalent of Proposition A.1 for uneducated workers requires that $\alpha_1 \geq \alpha_2$ rather than $\alpha_1 < \alpha_2$.

Proposition A.2. *When $\alpha_1 \geq \alpha_2$, if bayesians discriminate against uneducated red workers, Naive employers will discriminate against uneducated red workers too.*

Proof. Very similar to Proposition A.1's proof. □

Once again, the simulations show an equivalent case to that of Figure A.1.

A.3 Continuous ability

We now assume that ability come from a continuous distribution, rather than the discrete distribution containing low, medium and high ability workers discussed in the main text. Denote by $N_i(s)$ the distribution from which ability for individuals of identity $i \in \{r, g\}$ are drawn. We assume that $N_i(s)$ follows a normal distribution $N_i(\mu_i, \sigma_i)$ for $i \in \{r, g\}$, where $\mu_r < \mu_g$ (meaning that red workers are worse on average). In line with the main text, we assume that the payoff of the employer is $X(s) = a + bs$ where $b > 0$. We denote $P_e(s)$ the probability that an individual of ability s obtains a university degree. For simplicity, we assume that $P_e = 0$ if $s < l$ and $P_e = 1$ if $s \geq l$, although we expect results to hold for most increasing functions of $P_e(s)$ with respect to ability.

Bayesians will then prefer educated green workers whenever

$$\int_{-\infty}^{\infty} \frac{N_g(s)P_e(s)X(s)}{\int_{-\infty}^{\infty} N_g(s)P_e(s)ds} ds > \int_{-\infty}^{\infty} \frac{N_r(s)P_e(s)X(s)}{\int_{-\infty}^{\infty} N_r(s)P_e(s)ds} ds,$$

and green uneducated green workers whenever

$$\int_{-\infty}^{\infty} \frac{N_g(s)(1 - P_e(s))X(s)}{\int_{-\infty}^{\infty} N_g(s)(1 - P_e(s))ds} ds > \int_{-\infty}^{\infty} \frac{N_r(s)(1 - P_e(s))X(s)}{\int_{-\infty}^{\infty} N_r(s)(1 - P_e(s))ds} ds.$$

In contrast, since naives do not update based on the information signal, they prefer green workers (whether educated or not educated) whenever

$$\int_{-\infty}^{\infty} N_g(s)X(s)ds > \int_{-\infty}^{\infty} N_r(s)X(s)ds.$$

Then, the following result follows.

Proposition A.3. *When ability is normally distributed, naive employers discriminate against red educated workers strictly more than Bayesian employers.*

Proof. Note that for naives

$$\int_{-\infty}^{\infty} N_i(s)X(s)ds = \int_{-\infty}^{\infty} N_i(\mu_i, \sigma_i)(a + bs)ds = \int_{-\infty}^{\infty} \frac{\exp \frac{(s-\mu_i)^2}{2\sigma_i^2}}{\sqrt{2\pi}\sigma_i}(a + bs)ds = a + b\mu_i.$$

Therefore, naives will prefer green workers whenever

$$a + b\mu_g > a + b\mu_r \implies \mu_g > \mu_r,$$

which is always true.

For bayesians, the expected payoff of hiring an educated worker of identity i is

$$\int_{-\infty}^{\infty} \frac{N_g(s)P_e(s)X(s)}{\int_{-\infty}^{\infty} N_g(s)P_e(s)ds} ds = \frac{\int_{-\infty}^{l} N_i(\mu_i, \sigma_i)(a + bs)0ds + \int_{l}^{\infty} N_i(\mu_i, \sigma_i)(a + bs)ds}{\int_{-\infty}^{l} N_i(\mu_i, \sigma_i)0ds + \int_{l}^{\infty} N_i(\mu_i, \sigma_i)ds} =$$

$$= \frac{(a + b\mu_i)\text{erfc}\left(\frac{l-\mu_i}{\sqrt{2}\sigma_i}\right) + \frac{2b\sigma_i \exp\left(\frac{-(l-\mu_i)^2}{2\sigma_i^2}\right)}{\sqrt{2\pi}}}{\text{erfc}\left(\frac{l-\mu_i}{\sqrt{2}\sigma_i}\right)} =$$

$$= a + \frac{\sqrt{\frac{2}{\pi}} b\sigma_i \exp(-\frac{(l-\mu_i)^2}{2\sigma_i^2})}{\text{erfc}\left(\frac{l-\mu_i}{\sqrt{2}\sigma_i}\right)} + b\mu_i.$$

Therefore, bayesians will prefer green workers whenever

$$a + \frac{\sqrt{\frac{2}{\pi}} b\sigma_g \exp(-\frac{(l-\mu_g)^2}{2\sigma_g^2})}{\text{erfc}\left(\frac{l-\mu_g}{\sqrt{2}\sigma_g}\right)} + b\mu_g > a + \frac{\sqrt{\frac{2}{\pi}} b\sigma_r \exp(-\frac{(l-\mu_r)^2}{2\sigma_r^2})}{\text{erfc}\left(\frac{l-\mu_r}{\sqrt{2}\sigma_r}\right)} + b\mu_r.$$

$$\mu_g - \mu_r > \frac{\sqrt{\frac{2}{\pi}} \sigma_r \exp(-\frac{(l-\mu_r)^2}{2\sigma_r^2})}{\text{erfc}\left(\frac{l-\mu_r}{\sqrt{2}\sigma_r}\right)} - \frac{\sqrt{\frac{2}{\pi}} \sigma_g \exp(-\frac{(l-\mu_g)^2}{2\sigma_g^2})}{\text{erfc}\left(\frac{l-\mu_g}{\sqrt{2}\sigma_g}\right)}.$$

Naives always prefer green workers. Therefore, we only need to find an example in which, given these assumptions, bayesians prefer red workers. Let $\mu_g = 0$, $\sigma_g = 1$, $\mu_r = -1$, $\sigma_r = 1.5$, and $l = 1$. Then, the expected payoff of hiring a green worker is $a + 1.52b$ and the expected payoff of hiring a red worker is $a + 1.69b$. Since $b > 0$, Bayesian employers will prefer red workers while naive employers will prefer green workers. □

Proposition A.4. *When ability is normally distributed, naive employers discriminate against uneducated red workers strictly more than Bayesian employers.*

B Theoretical Results: Endogenous Case

This section derives the results for the case in which workers choose whether to pursue education. First we consider the case in which all employers are either Bayesian or naive, which corresponds to Proposition 2. Next we consider the case in which a share of the employers are Bayesians and a share are naive employers, which corresponds to Proposition 3.

B.1 Proof of Proposition 2

Full Set of Equilibria if Employers are Naive. Recall that employers observe a red and a green worker, their education level, and have to decide whom to hire. Hence, the worker's game when they face a Naive employer can be captured by Table B.1, where the row player is the red worker and the column player is the green worker. There, p^i_a is the (expected) probability that the worker of identity i and type a attains education.

	E	¬ E
E	$(1-p^g_{a'})p^r_a w - c$, $p^g_{a'}w - c$	$p^r_a w - c$ 0
¬ E	$0, p^g_{a'}w - c$	$0, 0$

Table B.1. Worker payoffs if employer is naive. Row player is red worker of ability i and column player is green worker of ability s'.

From this payoff matrix we can study under what conditions each of the equilibria would hold. This is, each worker knows his ability, the distribution of ability across both colors, and the decisions that the employer would make given these parameters and the education signals. With these parameters, both workers weigh the expected payoff of pursuing education against the expected payoff of not pursuing it. The table illustrates the parameter conditions needed to sustain each equilibrium. We summarize them here.

1. $(E, E; E, E)$ is part of an equilibrium iff $(1-(g_m p_m + g_h))p_m w - c > 0$.
2. $(\neg E, E; E, E)$ is part of an equilibrium iff:
   - $(1-(g_m p_m + g_h))p_m w - c < 0$
   - $(1-(g_m p_m + g_h))w - c > 0$
   - $p_m w - c > 0$
3. $(\neg E, \neg E; E, E)$ is part of an equilibrium iff:
   - $(1-(g_m p_m + g_h))w - c < 0$
   - $p_m w - c > 0$
4. $(\neg E, E; \neg E, E)$ is part of an equilibrium iff:
   - $(1-g_h)\pi_w - c > 0$
   - $p_m w - c < 0$

5. $(\neg E, \neg E; \neg E, E)$ is part of an equilibrium iff

    - $(1-g_h)w - c < 0$
    - $w - c > 0$

Full Set of Equilibria if Employers are Bayesian. The payoffs for workers facing a Bayesian employer are depicted in Table B.2. The left payoff matrix corresponds to the case where condition **B** is satisfied ($\frac{g_h}{g_m} > \frac{r_h}{r_m}$), meaning that the employer prefers the green worker, while the right payoff matrix corresponds to the case where **B** is not satisfied. We assume that the employer randomizes between hiring one or the other worker if both are equally preferred.

	E	¬ E
E	$(1-p^g_{s'})p^r_s w - c$, $p^g_{s'} w - c$,	$p^r_s w - c$ 0
¬ E	$0, p^g_{s'} w - c$	$0, 0$

	E	¬ E
E	$p^r_s w - c$, $(1-p^r_s)p^g_{s'} w - c$,	$p^r_s w - c$ 0
¬ E	$0, p^g_{s'} w - c$	$0, 0$

Table B.2. Worker payoffs if the employer is Bayesian depending on whether condition **(B)** is satisfied (left panel) or not (right panel). Row player is red worker of type a and column player green player of type a'.

By considering this scenario, the following set of equilibria exist:

1. $(E, E; E, E)$ is part of an equilibrium iff:

    - $(1-(g_m p_m + g_h))p_m w - c > 0$ and condition **(B)** holds or
    - $(1-(r_m p_m + r_h))p_m w - c > 0$ and condition **(B)** does not hold.

2. $(E, E; \neg E, E)$ is:

    - Not an equilibrium if condition **(B)** holds. (In this case, either g_m wants to pursue education or r_m does not want to pursue education).
    - Part of an equilibrium if condition **(B)** does not hold, $(1-g_h)p_m w - c > 0$, and $(1-(r_m p_m + r_h))p_m w - c < 0$.

3. $(\neg E, E; E, E)$ is:

    - Not an equilibrium if condition**(B)** does not hold. (In this case, either r_m wants to pursue education or g_m does not want to pursue education).
    - Part of an equilibrium if condition**(B)** holds, if $(1-r_h)p_m w - c > 0$, and $(1-(g_m p_m + g_h))p_m w - c < 0$.

4. $(\neg E, E; \neg E, E)$ is part of an equilibrium. It requires:

    - $(1-r_h)w + r_h\frac{w}{2} - c > 0$
    - $(1-g_h)w + g_h\frac{w}{2} - c > 0$
    - $(1-g_h)p_m w - c < 0$
    - $(1-r_h)p_m w - c < 0$

These conditions imply that $p_m < \frac{2-i_h}{2-2i_h} \geq 1$

5. $(\neg E, E; \neg E, \neg E)$ is part of an equilibrium iff
   - $(1 - r_h)w + r_h\frac{w}{2} - c < 0$.
6. $(\neg E, \neg E; \neg E, E)$ is part of an equilibrium iff
   - $(1 - g_h)w + g_h\frac{w}{2} - c < 0$.

To sum up, Table B.3 shows the full set of equilibria when education is endogenous for Bayesian and Naive employers. Note that this analysis proves Proposition 2, which states that red workers study weakly less than green workers of the same ability when the employer is Naive. More concretely, there are no equilibria in which red workers are more likely to educate than green workers when the employers are Naive, although such equilibria exist if employers are Bayesian.

			Symmetric Equilibria	
		$(E, E; E, E)$	$(\neg E, E; \neg E, E)$	
Bayesian	**B**	$c < (1 - (g_m p_m + g_h))p_m w$	$(1 - i_h)p_m w < c < (1 - \frac{i_h}{2})w, \forall i = r, g$	
	not B	$c < (1 - (r_m p_m + r_h))p_m w$	$(1 - i_h)p_m w < c < (1 - \frac{i_h}{2})w, \forall i = r, g$	
Naive		$c < (1 - (g_m p_m + g_h))p_m w$	$p_m w < c < (1 - g_h)w$	
			Asymmetric Equilibria - More Green	
		$(\neg E, \neg E; \neg E, E)$	$(\neg E, \neg E; E, E)$	$(\neg E, E; E, E)$
Bayesian	**B**	$c > (1 - \frac{g_h}{2})w$	-	$(1 - r_h)p_m w > c > (1 - (g_m p_m + g_h))p_m w$
	not B	$c > (1 - \frac{g_h}{2})w$	-	-
Naive		$c > (1 - g_h)w$	$(1 - (g_m p_m + g_h))w < c < p_m w$	$(1 - (g_m p_m + g_h))p_m w < c < (1 - (g_m p_m + g_h))w$
			Asymmetric Equilibria - More Red	
		$(\neg E, E; \neg E, \neg E)$	$(E, E; \neg E, \neg E)$	$(E, E; \neg E, E)$
Bayesian	**B**	$c > (1 - \frac{r_h}{2})w$	-	-
	not B	$c > (1 - \frac{r_h}{2})w$	-	$(1 - g_h)p_m w > c > (1 - (r_m p_m + r_h))p_m w$
Naive		-	-	-

Table B.3. Equilibrium education decisions and parameter conditions under which they can be supported in equilibrium. Condition **B** is $\frac{g_h}{g_m} > \frac{r_h}{r_m}$.

B.2 Endogenous case with more general payoffs

In the theory section of the main body of the paper we consider the simplest case in which employer's payoff increases linearly in worker's ability. In that case, condition **(N)** is always satisfied and the result for Proposition 2 follows. The theoretical result, however, may be lost when one considers more general payoff schemes, for example situations in which employers are especially interested in hiring a high-ability worker.

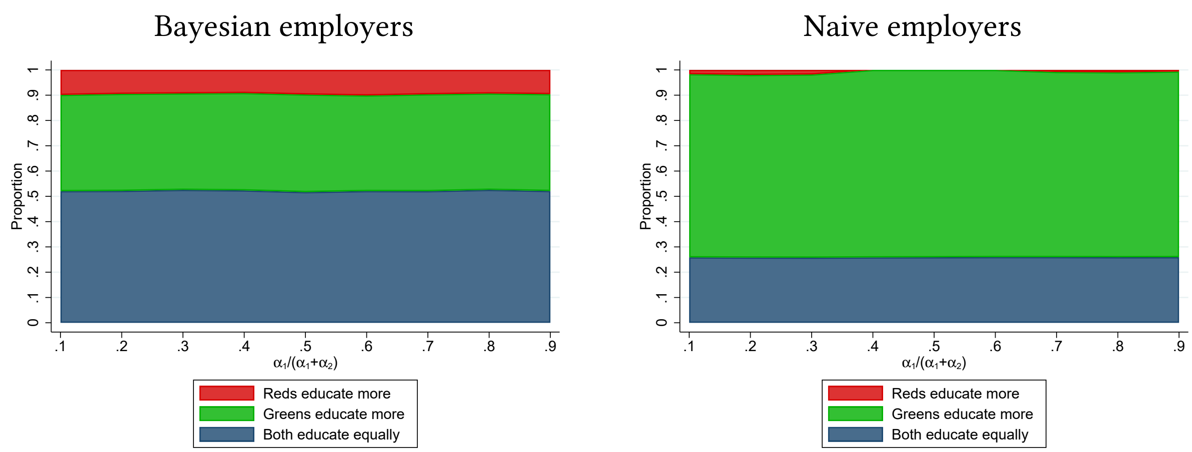


Figure B.1. Proportion of equilibria when hiring costs are high

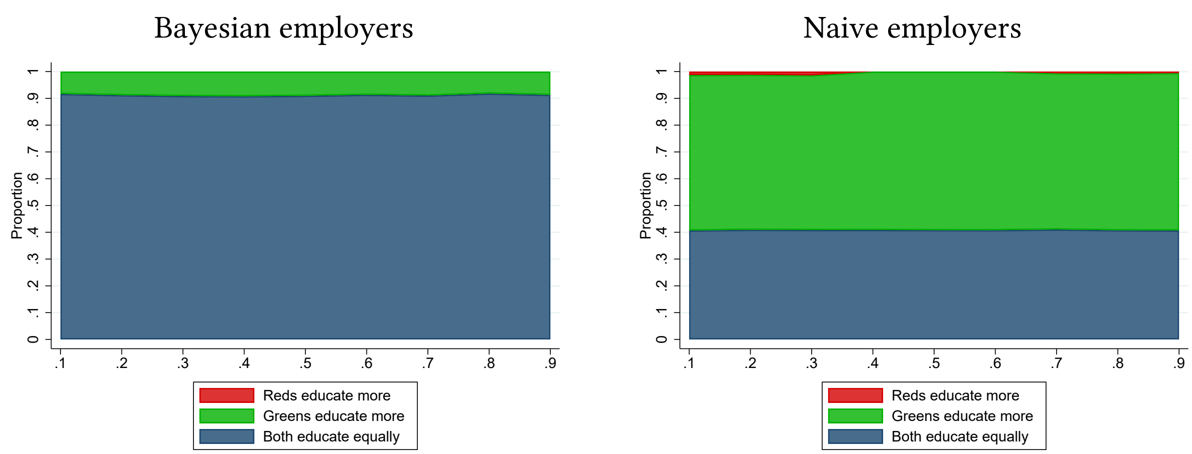


Figure B.2. Proportion of equilibria when hiring costs are low

Figures B.1 and B.2 study these more general scenarios using numerical simulations. More concretely, we generate parameter combinations and study what proportion of each of the equilibria survives under each setting. To do so, in each simulation we generate two groups of 4 workers where each worker is randomly assigned an ability l, m, or h (we generate groups based on 4 workers to mimic the experiment in the paper, but results are very similar assuming larger groups). The group that on average has the higher average ability is defined as the green group, and the other one is defined as the red group. We assume that $p_m = 0.5$ (the results are similar assuming different values). Figure B.1 further assumes relatively high costs, where $w = 10$ and $c = 8$, while Figure B.2 assume relatively low costs, where $w = 10$ and $c = 2$. As in Figure A.1, the X-axis captures $\frac{\alpha_1}{\alpha_1+\alpha_2}$, where a low value implies increasing marginal returns to ability, and a high value implies decreasing marginal returns to ability. For each value $\frac{\alpha_1}{\alpha_1+\alpha_2} \in \{0.1, 0.2, ..., 0.8, 0.9\}$ we generate 10,000 numerical simulations and for each of those we count the number of equilibria in which reds educate more, green educate more, and both educate equally. The graphs display the proportion of each kind of equilibrium across all the simulations.

Both figures show that the intuitions of the results in the main text largely carry over to this more general setting. While Naive employers can sometimes support equilibria in which red workers educate more, these cases are extremely uncommon. More generally, across all parameter combinations naive employers are much more likely to support equilibria in which greens educate more than reds, while most equilibria with Bayesian employers imply that both workers educate equally.

B.3 Proof of Proposition 3

In this section, we derive the full set of equilibria in a situation in which γ firms are naive and $1-\gamma$ firms are Bayesian. We assume workers to know the proportion of firms that are naive and Bayesian, but to not know whether the firm that they are applying to is Bayesian or naive.

Note that there can only exist 8 possible equilibria. To see this, note that since $p_l = 0$, a low ability worker would never want to pursue education. Since $p_h = 1$ and $w - c > 0$, there cannot exist any equilibrium in which no one pursues education (since a high ability worker would then decide to study). Note also that there cannot exist any equilibrium in which, for a given color, high ability workers do not study while medium ability workers study. To see this, note first that, since firms cannot see workers' ability, if it is not worth it for a high ability to study (because the cost of studying is higher than the expected benefit), neither is it worth it for a medium ability worker to study (since the cost to study is the same as high ability workers, but the potential benefit is reduced by p_m). Therefore, it can only be that, within a given color, medium ability workers study but high ability workers do not if, when only medium ability workers study, firms are more likely to choose this color. However, this will never be the case: if firms are naive, they only rely on the base-rate and they therefore do not react to knowing that high ability workers are not studying. If firms are Bayesian, these firms will learn that any worker that studies from the color in which mediums study is medium ability. They will therefore always prefer to choose the other color if in the other color high abilities are studying. If in the other color only mediums are studying, then it would become dominant for high abilities to start studying because then firms would always pick them.

Here, we consider these eight equilibria, how they depend on the condition (B) discussed in the main text, and how γ interacts with them:

1. $(E, E; E, E)$ is part of an equilibrium iff:
   - If **(B)**:
     $(1 - (g_m p_m + g_h))p_m w - c > 0$
   - If ¬**(B)**
     $\gamma((1 - g_m p_m - g_h)p_m w) + (1-\gamma)(p_m w) - c > 0$
     $\gamma(p_m w) + (1-\gamma)((1 - r_m p_m - r_h)p_m w)) - c > 0$
2. $(E, E; \neg E, E)$ is part of an equilibrium if:
   - If **(B)**: Not an equilibrium
   - If ¬**(B)**:
     $(1 - g_h)p_m w - c > 0$
     $\gamma(p_m w) + (1-\gamma)((1 - r_m p_m - r_h)p_m w)) - c < 0$
3. $(\neg E, E; E, E)$ is part of an equilibrium iff:
   - If **(B)**:
     $((1 - g_m p_m - g_h)p_m w) - c < 0$
   - If ¬**(B)**:
     $\gamma((1 - g_m p_m - g_h)p_m w) + (1-\gamma)p_m w - c < 0$
     $\gamma((1 - g_m p_m - g_h)w) + (1-\gamma)w - c > 0$
     $\gamma(p_m w) + (1-\gamma)((1 - r_h)p_m w) - c > 0$
4. $(E, E; \neg E, \neg E)$ is not part of an equilibrium.

5. $(\neg E, \neg E; E, E)$ is part of an equilibrium iff:

   - $p_m w - c > 0$
     $\gamma((1 - p_m g_m - g_h)w) + (1-\gamma)w - c < 0$

6. $(\neg E, E; \neg E, E)$ is part of an equilibrium iff:

   - $((1 - g_h)p_m w) - c < 0$
     $\gamma((1 - g_h)w) + (1-\gamma)((1 - g_h)w + g_h \frac{w}{2}) - c > 0$
     $\gamma(wp_m) + (1-\gamma)((1 - r_h)p_m w) - c < 0$
     $\gamma w + (1-\gamma)((1 - r_h)w + r_h \frac{w}{2}) - c > 0$

7. $(\neg E, E; \neg E, \neg E)$ is part of an equilibrium iff

   - $wp_m - c < 0$
     $\gamma w + (1-\gamma)((1 - r_h)w + r_h \frac{w}{2}) - c < 0$

8. $(\neg E, \neg E; \neg E, E)$ is part of an equilibrium iff

   - $\gamma((1 - g_h)w) + (1-\gamma)((1 - g_h)w + g_h \frac{w}{2}) - c < 0$
     $wp_m - c < 0$

Note that by Proposition 2 we know that when $\gamma = 1$ (all employers are Naive) there exist no equilibria in which red workers educate more, while when $\gamma = 0$ (all employers are Bayesian) these equilibria do exist. Note further that γ enters all the expressions above linearly, which implies that there exist a γ^* such that for $\gamma > \gamma^*$ there will only exist equilibria where greens educate more than reds. This proves Proposition 3.

C Additional tables

Table C.1. Sample characteristics from Experiment I

Demographics		*Ethnicity*		*Social class*		*Decisions*	
Female	0.633	White	0.461	Working class	0.461	Quiz correct	0.209
Age between 20-24 years old	0.506	Black	0.104	Middle class	0.474	Cognitive ability (max 6)	2.19
Age between 25-30 years old	0.159	Asian	0.156	Upper class	0.065	Risk aversion (max 10)	5.93
Age between >30 years old	0.182	Other	0.279				
British	0.312						
Europe	0.354						
Asia	0.205						

Table C.2. Name perceptions from Experiment II

	Origin UK	White	Christian	Working class	Intelligence	Likability
Julia	66	94	89	14	6.07	6.34
Becky	80	94	82	50	5.19	5.83
Anna	70	90	91	27	6.06	6.26
Megan	68	89	87	36	5.53	5.91
Liam	91	93	85	62	5.15	5.93
Joseph	74	85	87	25	5.97	6.08
Alan	96	97	87	60	5.56	5.95
Matthew	88	97	91	24	6.08	6.08
Average females	71	91.75	87.25	31.75	5.71	6.09
Average males	87.25	93	87.5	42.5	5.69	6.01

D Additional Figures

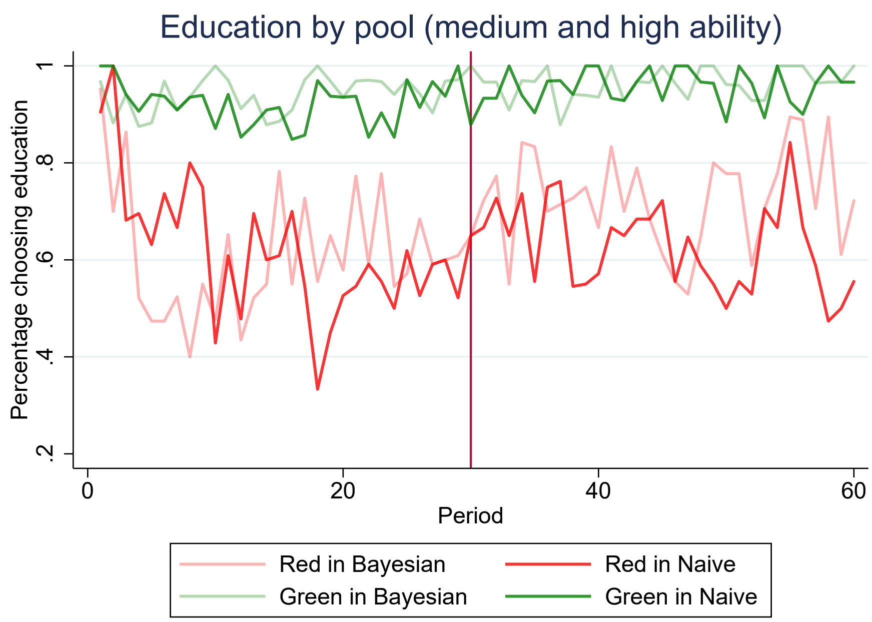
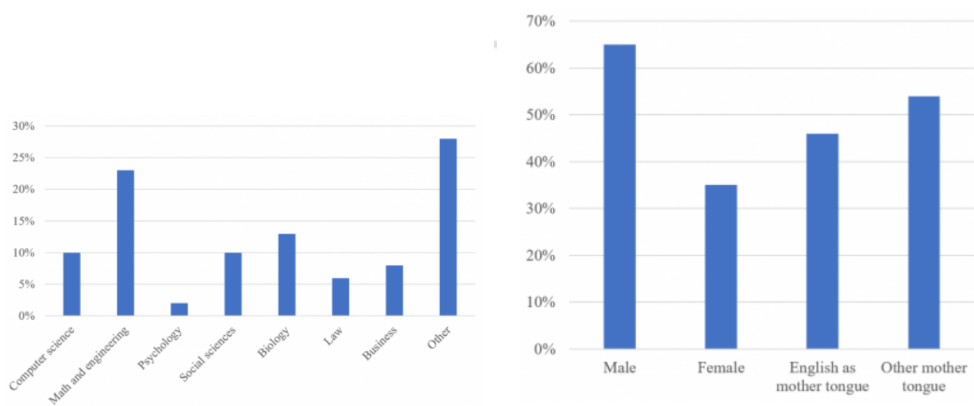


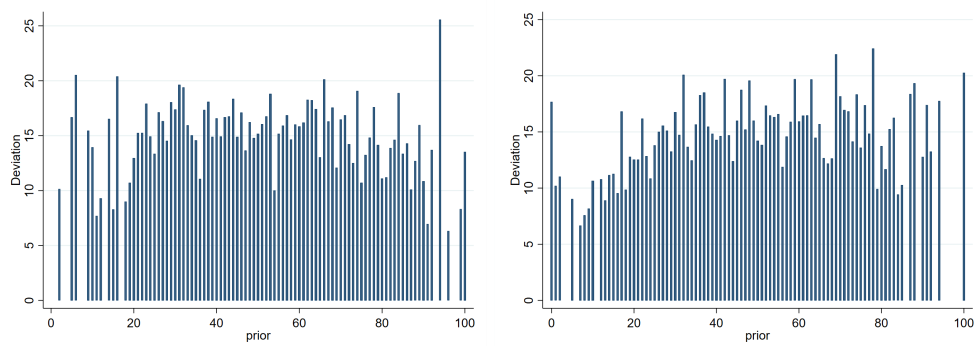
Figure D.1. Education Decisions in Experiment 1.



(a) Fields of Study (b) Other Characteristics

Figure D.2. Bar charts showing how information on top performers was presented to participants.

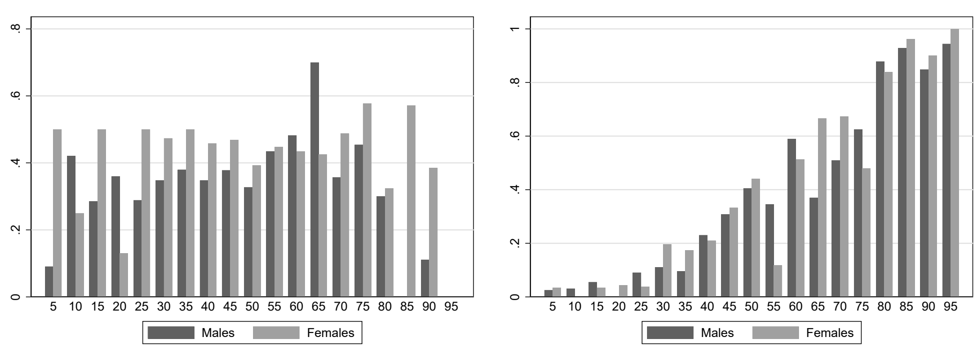
Figure D.3. Distribution of prior beliefs in Experiment 2.



(a) Male Candidates

(b) Female Candidates

Figure D.4. Average Deviation from Bayesian posterior depending on prior for male, Panel(a), and female, Panel (b), candidates.



(a) Priors

(b) Posteriors

Figure D.5. Fraction of hired workers by (a) Priors and (b) Posteriors for male and female candidates. Note that the most common priors for men are between 50-65 and the most common priors for women between 35-50.

E Instructions Experiment I

Welcome and thanks for participating in this experiment. Please read these instructions carefully. They are identical for all the participants with whom you will interact during this experiment. If you have any questions please raise your hand. One of the experimenters will come to you and answer your questions. From now on communication with other participants is not allowed. Please do also switch off your mobile phone at this moment. If you do not conform to these rules we are sorry to have to exclude you from the experiment.

You will receive 4 GBP just for showing up. During the experiment you can earn additional monetary rewards. How much you earn depends on your choices and those of others and is explained below. All your decisions will be treated confidentially.

This session consists of two parts. The first part consists of the main body of the experiment, and will take approximately 1 hour and 15 minutes. The second part is a questionnaire that will take approximately 30 minutes. Hence, we expect this experiment to take a bit less than 2 hours.

THE EXPERIMENT In the experiment you will be randomly assigned either the role of worker or the role of employer. If you are assigned the role of worker, you will also be randomly assigned an ability level (high, medium or low) and a colour (yellow or orange). Your role, ability level and colour will remain fixed during the entire experiment.

The experiment consists of 60 rounds. In each round, the computer will create one pool of 4 yellow workers and one pool of 4 orange workers. The computer will then randomly pick one yellow worker and one orange worker and assign them to an employer. Both the employer and the workers will see how many workers of each ability there are in each pool, but the employer will not see the ability of the particular workers that he/she is assigned to. There are then two decisions:

1. Each of the workers decide whether to pursue education or not. Pursuing education is costly and might or might not lead to graduation. The higher a worker's ability, the higher the chance that they graduate.
2. The employer sees whether each of the workers graduated. Then the employer decides whether to hire the yellow worker, the orange worker, or not hire any worker.

Workers' decisions As a worker you decide whether to pursue education or not. Pursuing education costs 1 GBP. If you decide to pursue education you will successfully graduate

- with a 100% chance (for sure) if you have high ability
- with an 80% chance if you have medium ability and
- with a 0% chance (for sure NOT) if you have low ability.

If you decide not to pursue education you will for sure NOT graduate.

In addition you receive

- a payment of 8 GBP if you are hired.
- a payment of 0 GBP if you are NOT hired.

In the second part of the experiment, you will be offered additional money if you correctly answer a quiz about other participants' behavior during the experiment. We will tell you more when you reach that part.

Employers' decisions As an employer you decide whether to hire the yellow worker, the orange worker, or neither worker. You cannot hire both workers.

When you make your decision you will see each worker's colour and whether they graduated. You will NOT be able to see their ability level. However you will see for each colour, how many workers of low, medium and high ability there are in the pool of yellow and in the pool of orange workers.

You receive

- 20 GBP if you hire a graduated and high-skilled worker
- 15 GBP if you hire a graduated and medium-skilled worker
- 10 GBP if you hire a graduated and low-skilled worker
- 0 GBP if you hire a worker who did NOT graduate
- 8 GBP if you decide NOT to hire

At the end of each round you will be informed about the ability level of the worker you hired and your round payment. You will then move to the next round.

Payment: At the end of the experiment we will pay you:

- Your earnings in one of the 60 randomly drawn rounds of the experiment
- + the amount that you receive for answering the quiz (only as a worker)
- + the amount that you receive for answering the questionnaire
- + 4 GBP show up fee

Enjoy the Experiment!

E.1 Screenshots of Experiment I

After reading the instructions and answering a set of 12 questions that made sure that participants understood the experiment, the experiment started. Below, we attach screenshots from the experiment from the perspective of the worker.

Figure E.1 represents an example of the first screen that workers saw in each round, in which they were asked whether they would like to pursue education. After answering it, and while they waited for the employers to make their decision, they were asked whether they thought they would be hired or not. Figure E.2 represents an example of the last screen that workers saw, summarizing their earnings for the round.

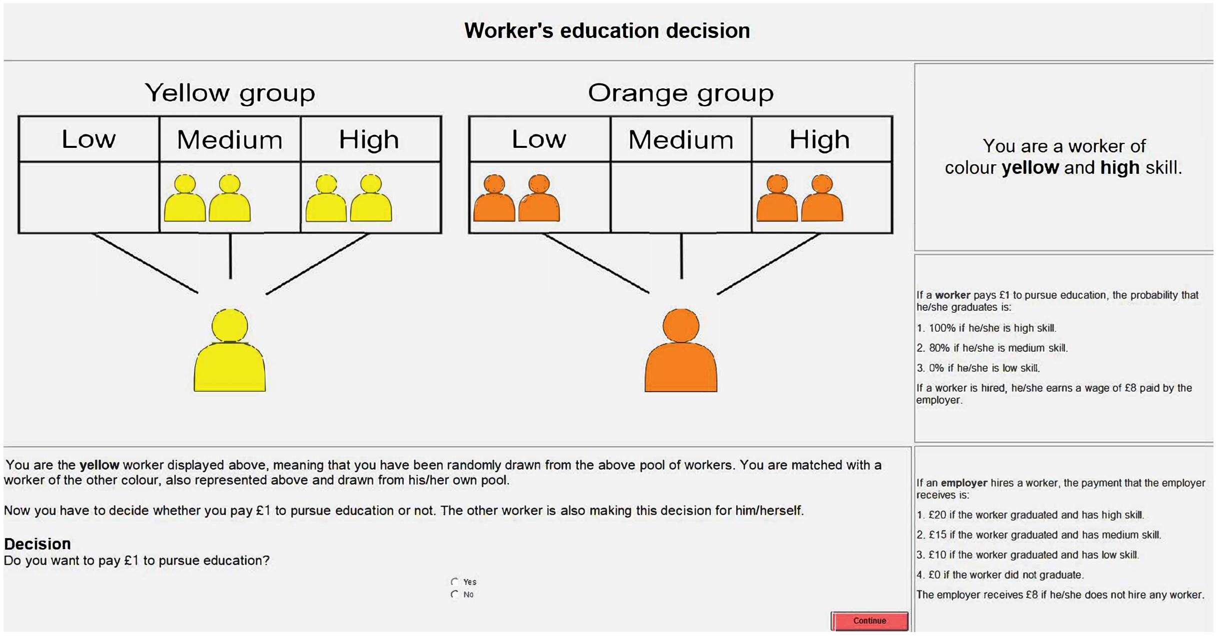


Figure E.1. Education decision from a yellow worker with high skill

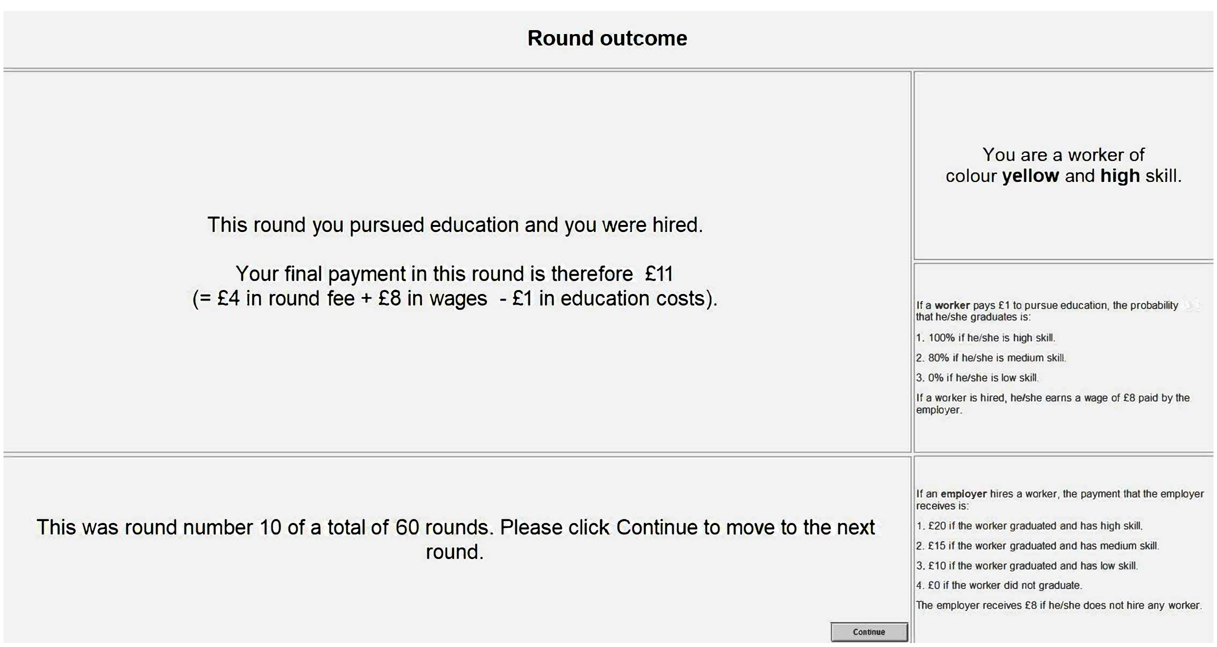


Figure E.2. Round outcome from a yellow worker with high skill

Below, we attach screenshots from the experiment from the perspective of the employer. Employers first saw the same picture shown to the workers, represented in E.1, and had to wait while workers made their decision. After the workers made their decisions, Figure E.3 represents an example of the screen employers saw when they had to make their decision. Finally, Figure E.4 represents an example of the last screen that employers saw, summarizing their earnings for the round.

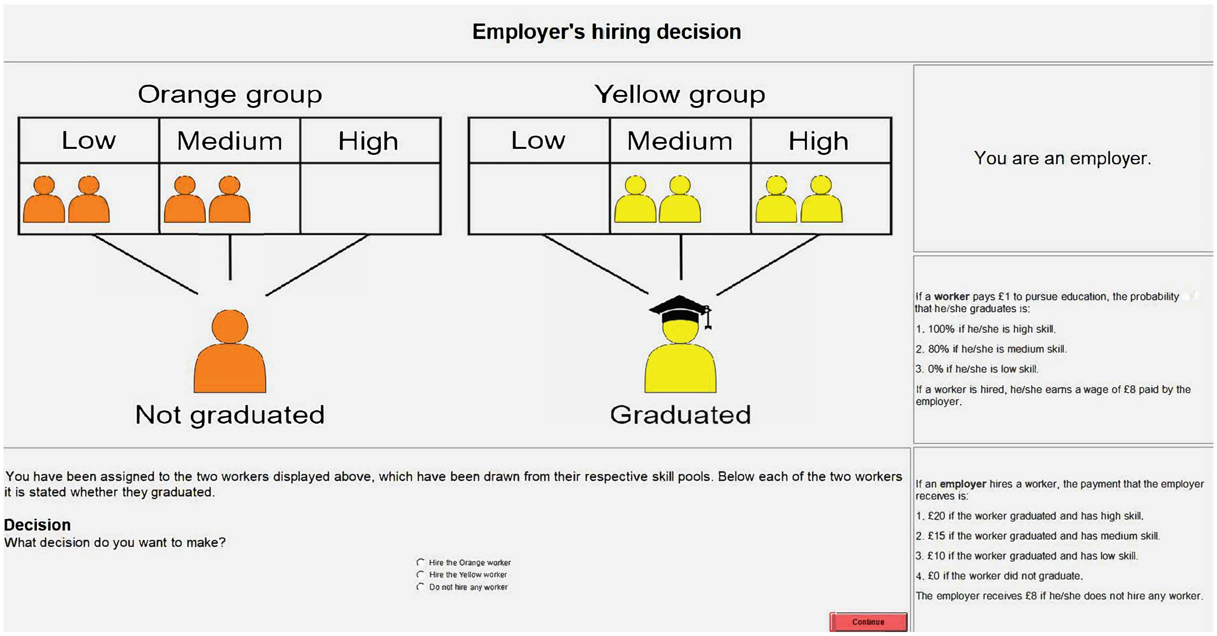


Figure E.3. Hiring decision from an employer

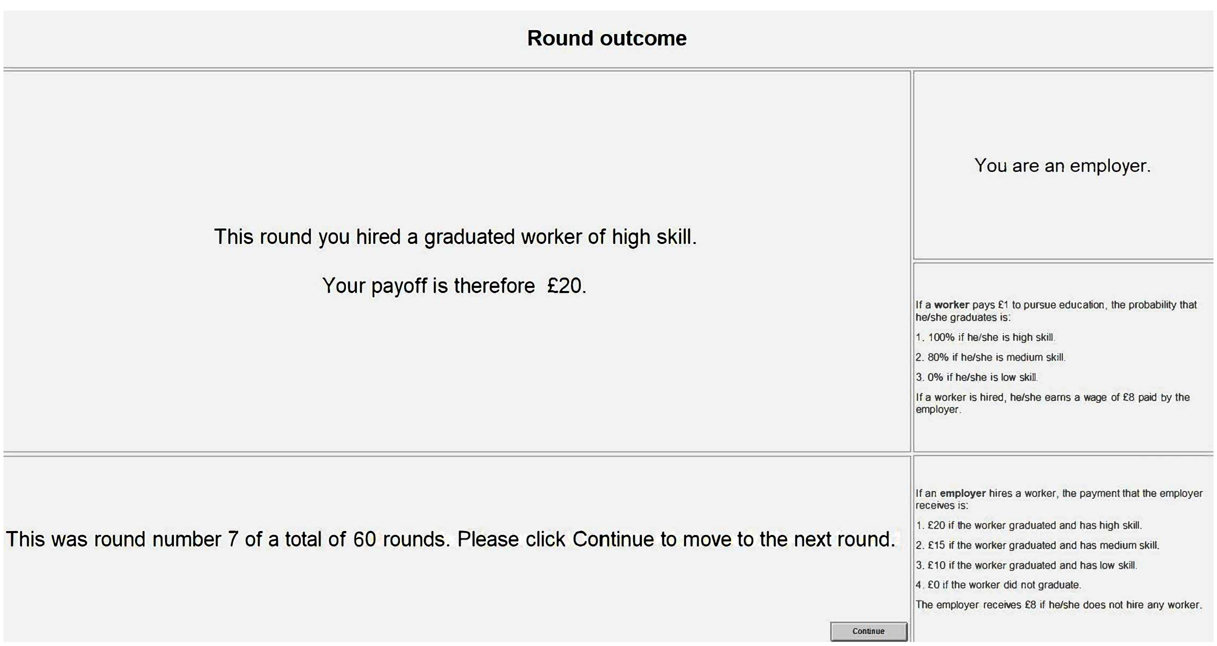


Figure E.4. Round outcome from an employer

After finishing the experiment, subjects were shown the following final questionnaire.

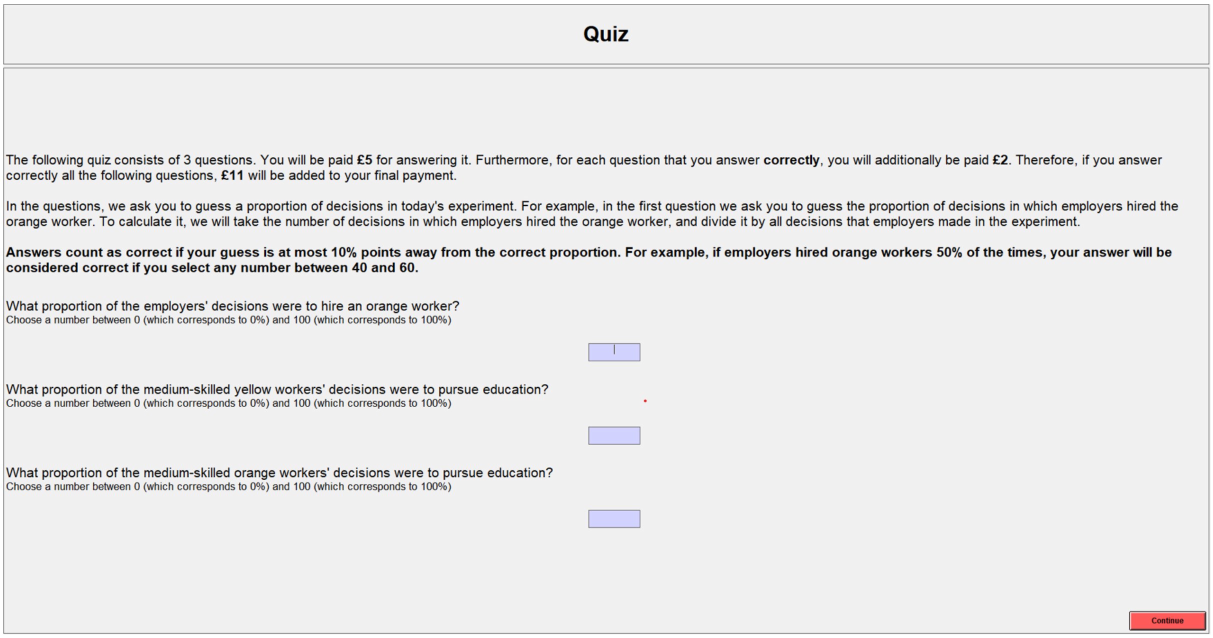


Figure E.5. Final questionnaire

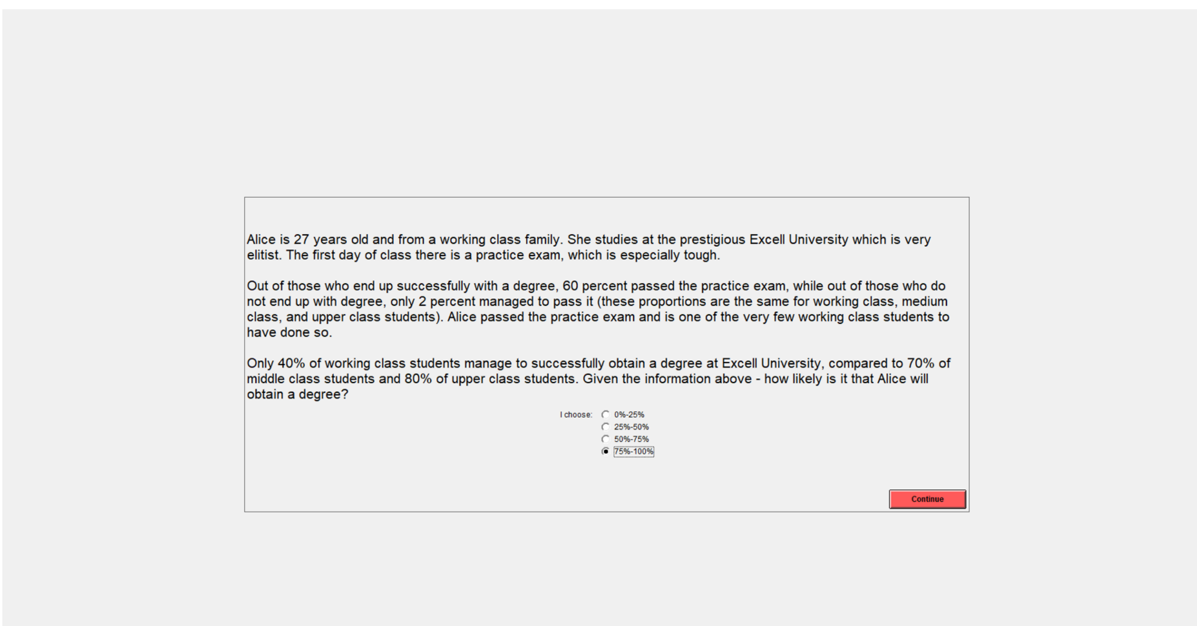


Figure E.6. Final questionnaire

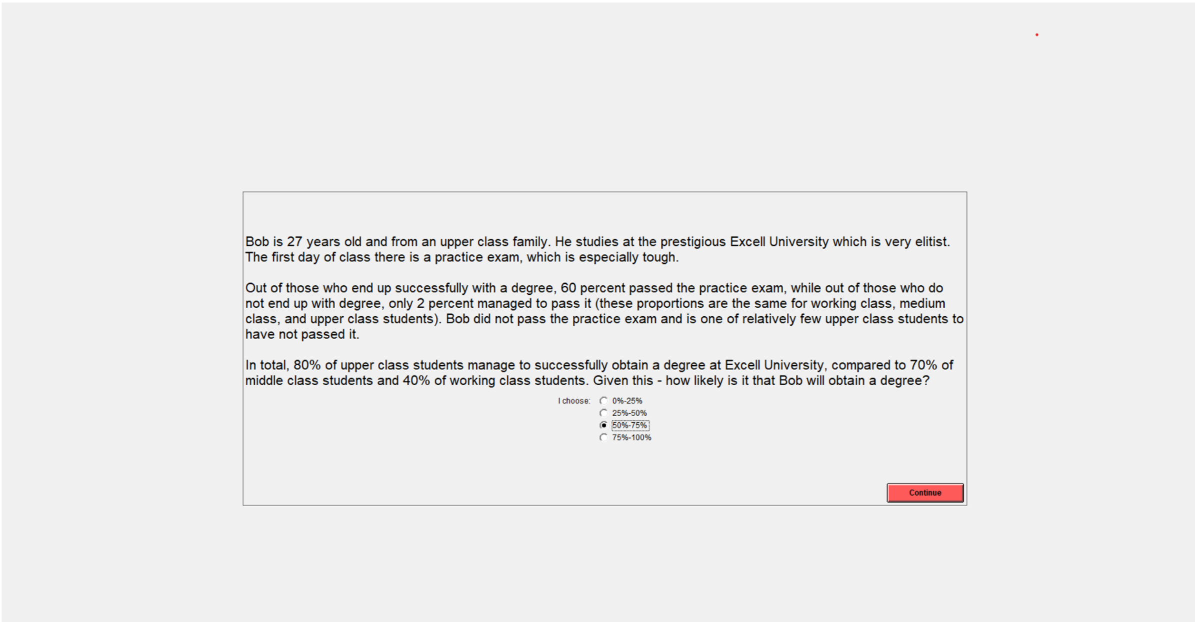


Figure E.7. Final questionnaire

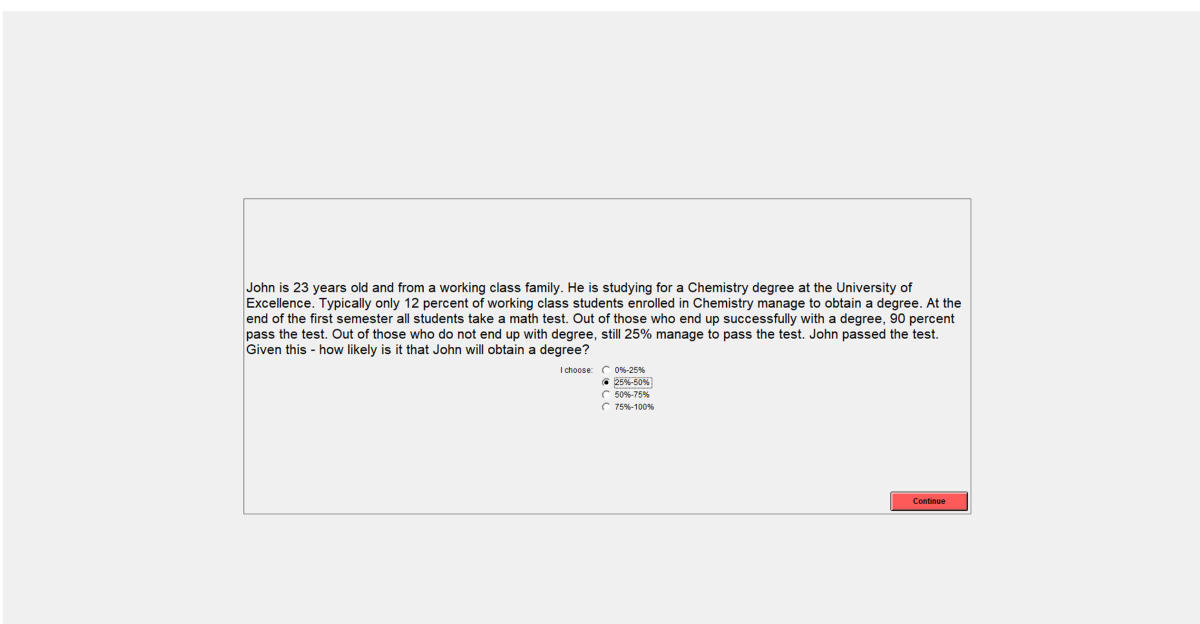


Figure E.8. Final questionnaire

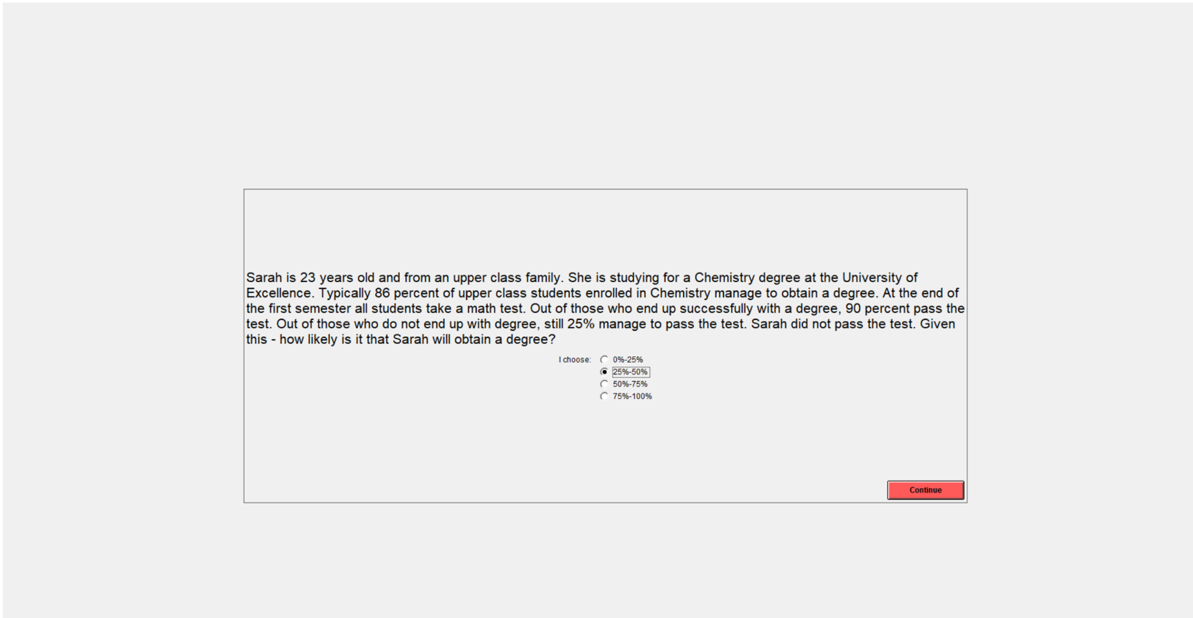


Figure E.9. Final questionnaire

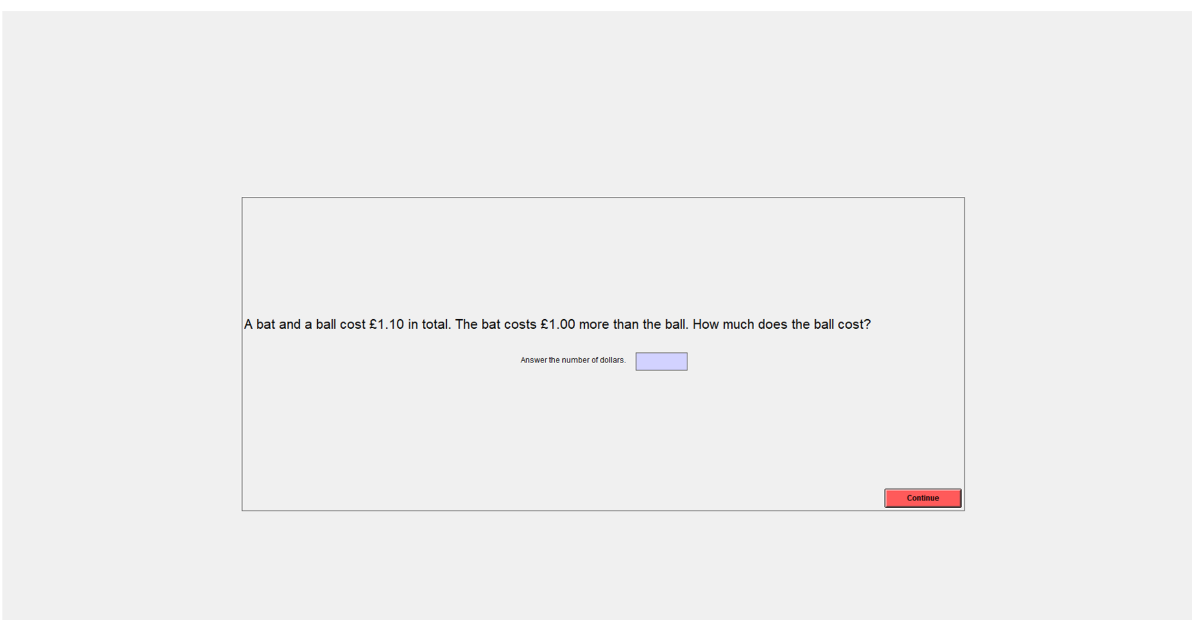


Figure E.10. Final questionnaire

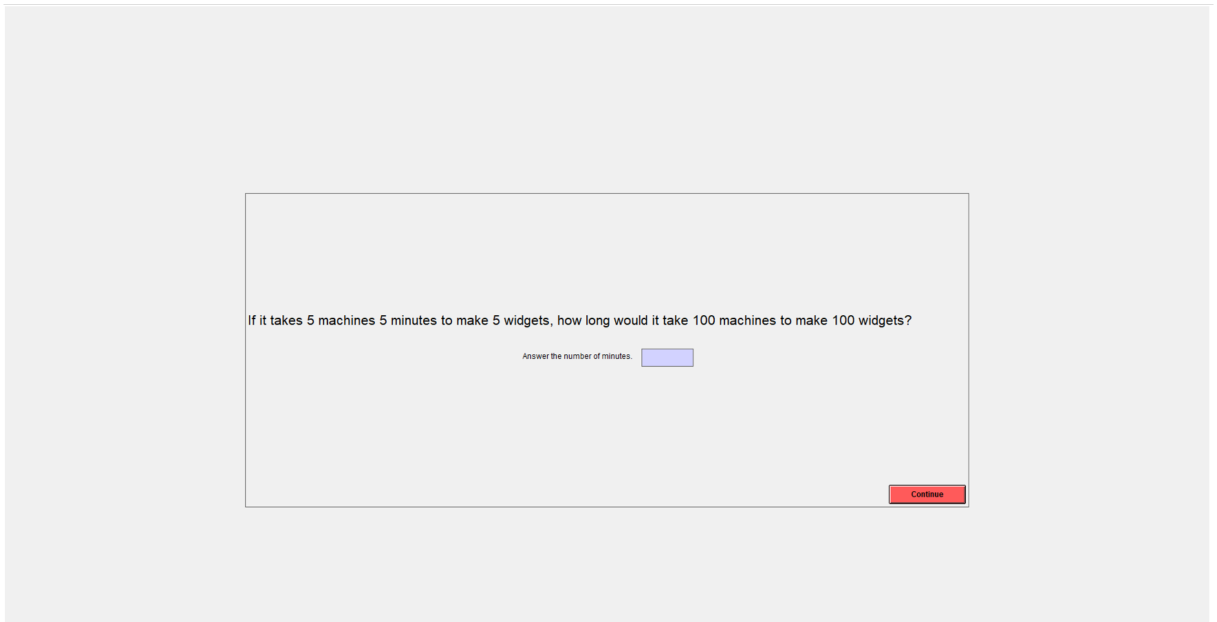


Figure E.11. Final questionnaire

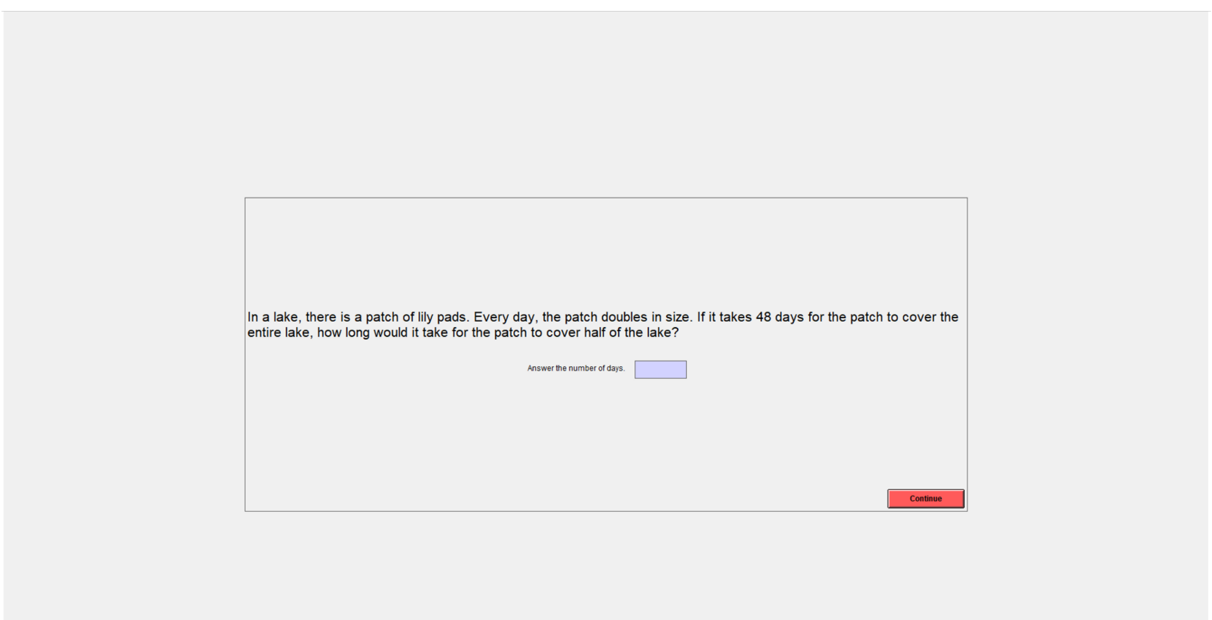


Figure E.12. Final questionnaire

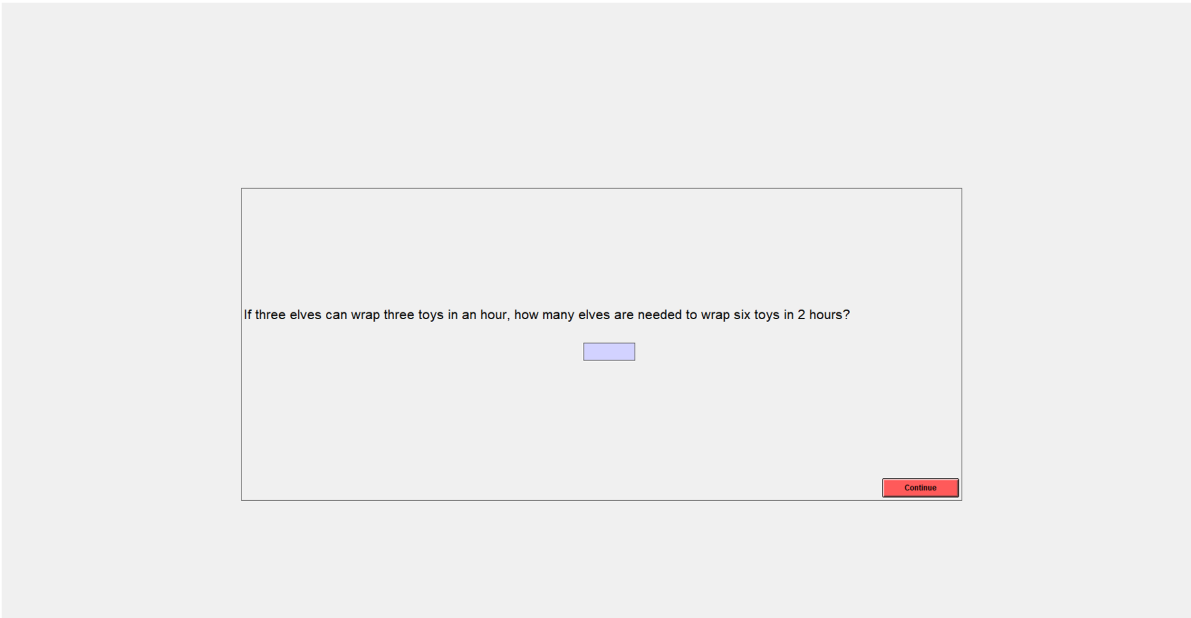


Figure E.13. Final questionnaire

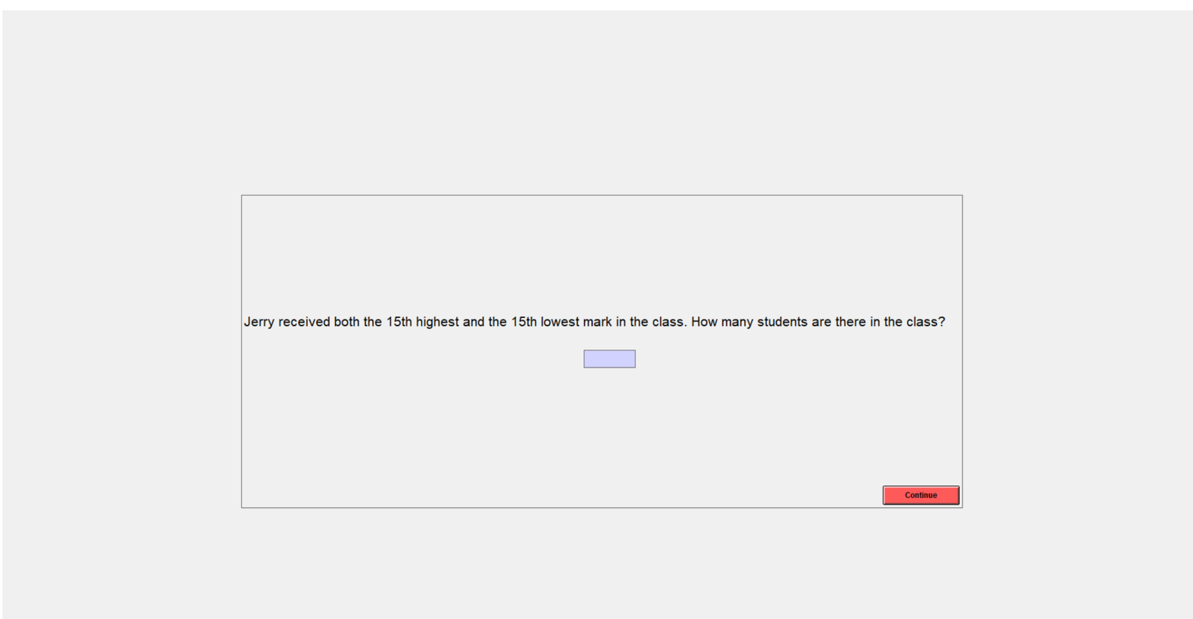


Figure E.14. Final questionnaire

Figure E.15. Final questionnaire

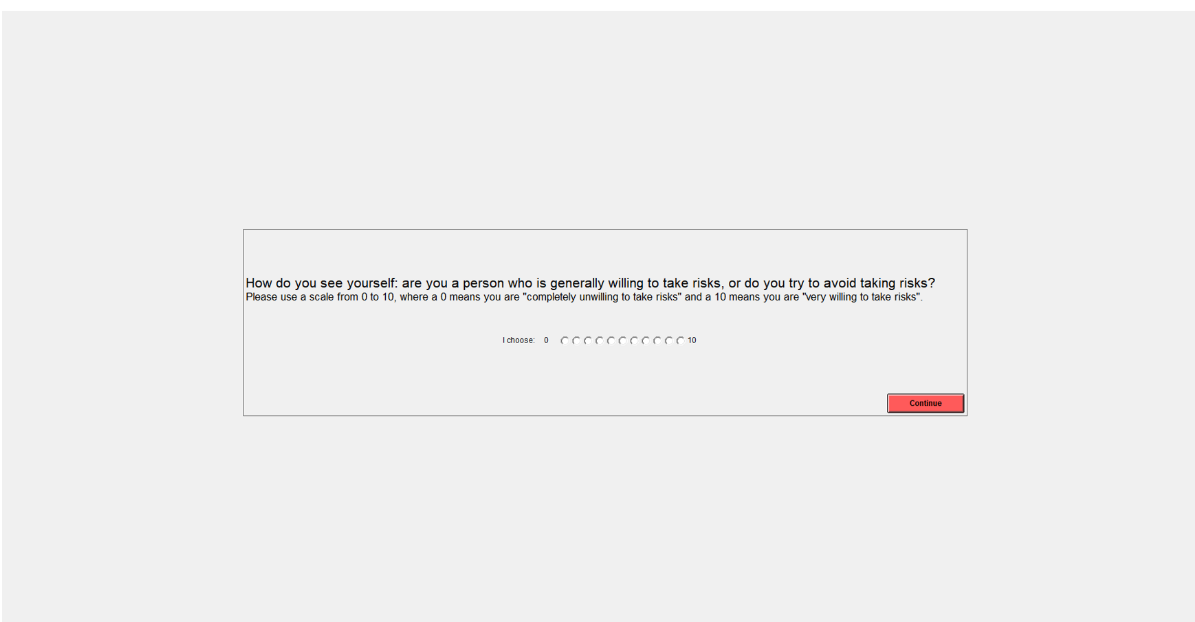


Figure E.16. Final questionnaire

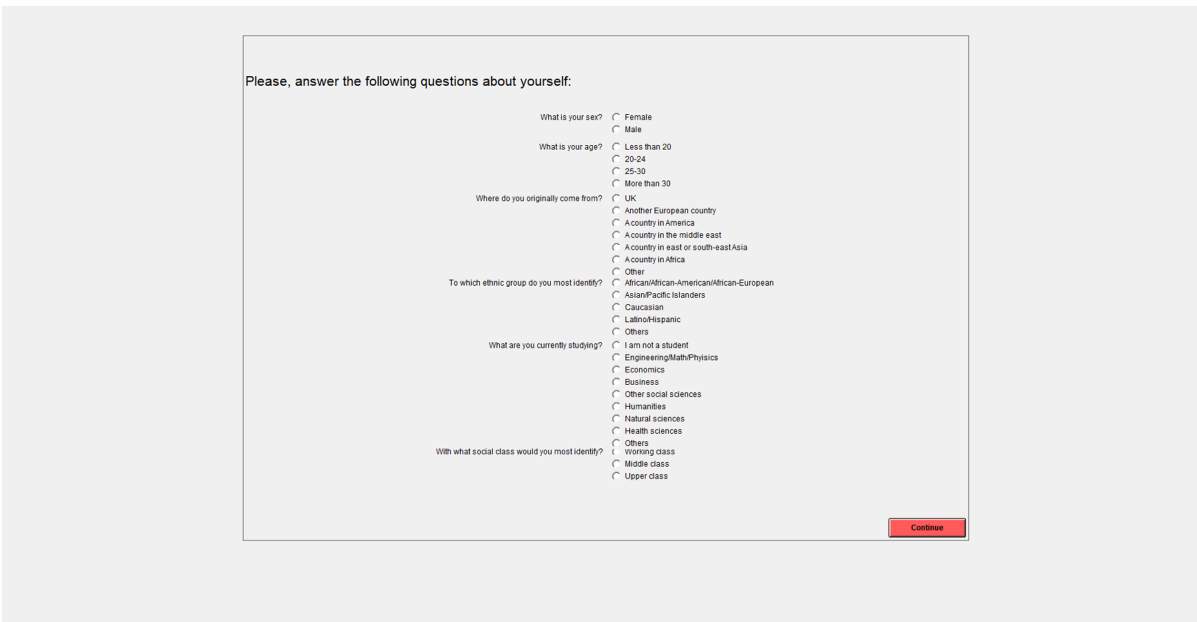


Figure E.17. Final questionnaire

F Instructions Experiment II

These are the Instructions given to participants at the very beginning of the online experiment. Please read the following lines very carefully, you will only be able to proceed with the study if you answer correctly a set of understanding questions below. This survey consists of two parts: PART 1 takes about 10 minutes to complete and PART 2 takes about 2 minutes to complete. We will first describe PART 1 and, after you have filled it out, describe PART 2.

BACKGROUND. We performed a Math and Logic test with 80 young university students from a university in the East of England, who we will call the candidates. The candidates were placed in 4 groups of 20 candidates each.

ASSESSMENT. In the following screens, we will show you a CV of one of the candidates, which includes the candidate's first name, age, gender, region of residence, marital status, and field of study at the university. You do not know the performance of this candidate: your goal is to assess how likely it is that the performance of this candidate is in the top half of his/her group. In this case we will say the candidate is a "Top Performer". In other words, you will have to guess how likely it is that the performance of this person is within the top 10 of his/her group of 20 candidates.

INFORMATION SIGNALS. After you have made your first assessment, we will give you a signal of the candidate. The signal will be either Positive or Negative.
If the candidate is a Top Performer, the signal is Positive with a 70% chance and Negative with a 30% chance.
If the candidate is not a Top Perfomer, the signal is Positive with a 30% chance and Negative with a 70% chance.

YOUR CHOICES. You will make six choices. For each of the choices, you will have to indicate how likely it is that the candidate is a Top Performer.
Choice 1. You will see the CV.
Choice 2. You will see the CV and one signal.
Choice 3. You will additionally see a second signal.
Choice 4. You will additionally see a third signal.
Choice 5. You will additionally see a fourth signal.
Choice 6. You will additionally see a fifth signal.

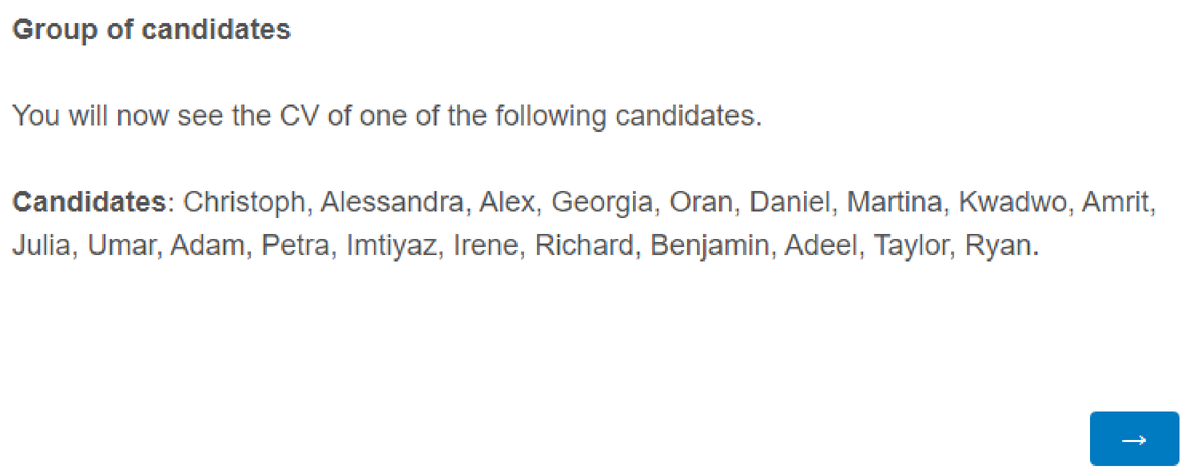
You will have to assess in total 4 candidates. So you will make Choice 1-6 four times, one for each different candidate. Hence you will make 24 choices in total in PART 1.

PAYMENT. At the end of the study, a computer will randomly pick one of the 28 choices that you made (which consist of 24 choices in PART 1 and 4 choices in PART 2) and you will be paid according to this choice. In PART 1, you will be paid either £0 or £5 based on the accuracy of your reported probability. The higher the accuracy, the higher the chances that you receive £5. Click here to see the exact method used to determine the probabilities. This method ensures that the best for you is to report your best guess. We will explain the payment structure from PART 2 later on.

In addition to this payment, you will be paid £1 fixed fee.

After answering the control questions to make sure that they understood the instructions, participants were briefly shown an example of the tasks that the candidates had to solve. Then, they were shown the summary statistics represented in Figure D.2 on candidates' average performance in the task.

Participants then started evaluating candidates. In what follows, we attach examples of the screens that participants saw during the experiment. The signals were then added to Figure F.5, one by one, as described in the main text of the paper.



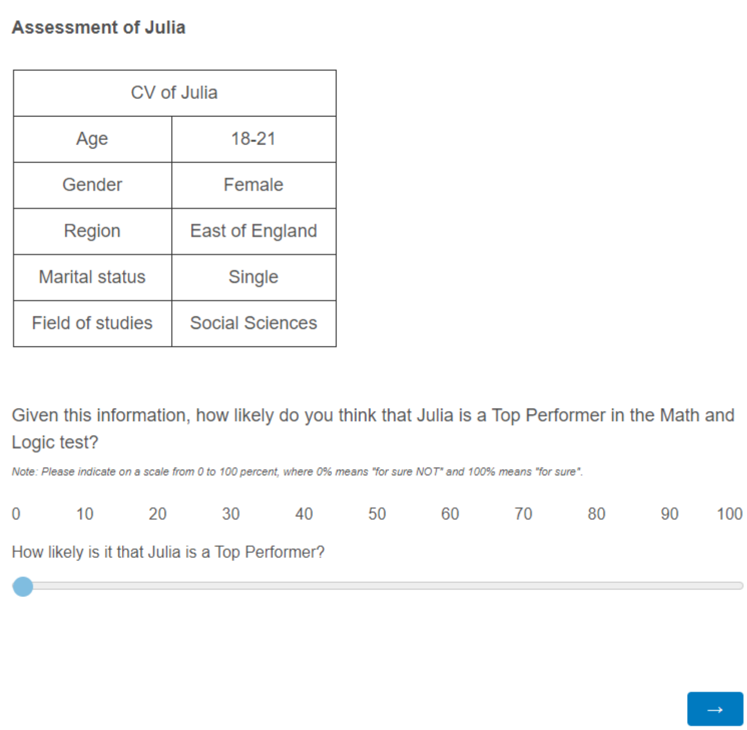
Group of candidates

You will now see the CV of one of the following candidates.

Candidates: Christoph, Alessandra, Alex, Georgia, Oran, Daniel, Martina, Kwadwo, Amrit, Julia, Umar, Adam, Petra, Imtiyaz, Irene, Richard, Benjamin, Adeel, Taylor, Ryan.

→

Figure F.1. Decision screens for Experiment 2



Assessment of Julia

CV of Julia	
Age	18-21
Gender	Female
Region	East of England
Marital status	Single
Field of studies	Social Sciences

Given this information, how likely do you think that Julia is a Top Performer in the Math and Logic test?

Note: Please indicate on a scale from 0 to 100 percent, where 0% means "for sure NOT" and 100% means "for sure".

0 10 20 30 40 50 60 70 80 90 100

How likely is it that Julia is a Top Performer?

→

Figure F.2. Decision screens for Experiment 2

Participants then faced the hiring decisions, where they had to decide whether they hired each of subjects they evaluated (note that participants were again reminded of the signals that they had seen)

PART 2

The second part of this study takes 2 minutes to complete. You will be asked whether you want to **hire each of the 4 candidates to solve a math and logic task for you** (you will hence make a total of 4 choices, one for each candidate, in random order).

If the Choice selected for payment is one of the following 4 choices, you will be paid as follows:

YOUR PAYMENT

- If you **do not hire the candidate,** you get £2.5.
- If you **hire the candidate**, you get £0 if the worker is **not a Top performer**, and £5 if the worker **is a Top performer**. Furthermore, we will send £2.5 to this worker (we will actually send him/her the money).

→

Figure F.3. Decision screens for Experiment 2

Hiring decision - Julia

CV of Julia	
Age	18-21
Gender	Female
Region	East of England
Marital status	Single
Field of studies	Social Sciences
Signal 1	
Signal 2	
Signal 3	
Signal 4	
Signal 5	

Note: Remember that if Julia is a Top Performer you see a Positive signal with 70% chance and a Negative signal with 30% chance. If Julia is not a Top Performer, you see a Positive signal with 30% chance and a Negative signal with 70% chance.

Now you have to choose whether to hire Julia:

- If you do not hire Julia, you get £2.5.
- If you hire Julia, we will pay her £2.5. You will then receive £5 if Julia was a Top Performer, and £0 if Julia was not a Top Performer.

What do you do?

Not hire Julia

Hire Julia

→

Figure F.4. Decision screens for Experiment 2

Finally, participants answered the final questionnaire.

What is your sex?

- Male
- Female

What year were you born?

Where do you originally come from?

- UK
- Another European country
- A country in Africa
- A country in America
- A country in East Asia
- A country in the Middle East
- A country in South Asia
- Other

Which ethnic group do you most identify as?

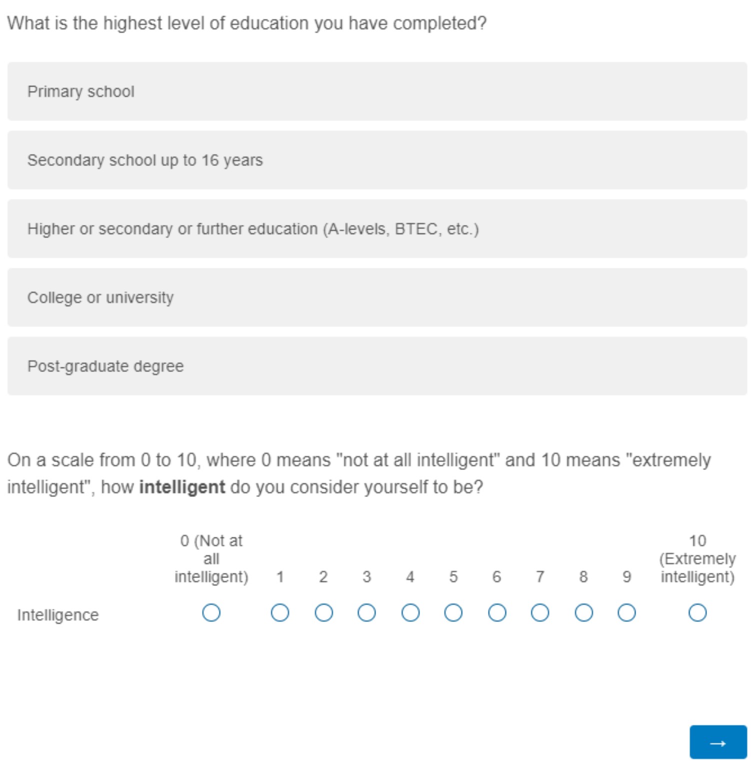
- Black African
- Black Carribean
- Other Black
- East Asian
- South Asian
- White British
- Other White
- Others

Is English your mother tongue?

- Yes
- No (please insert the name of your mother tongue)

Which social class do you most identify with?

- Working class
- Middle class
- Upper class



What is the highest level of education you have completed?

- Primary school
- Secondary school up to 16 years
- Higher or secondary or further education (A-levels, BTEC, etc.)
- College or university
- Post-graduate degree

On a scale from 0 to 10, where 0 means "not at all intelligent" and 10 means "extremely intelligent", how **intelligent** do you consider yourself to be?

	0 (Not at all intelligent)	1	2	3	4	5	6	7	8	9	10 (Extremely intelligent)
Intelligence	○	○	○	○	○	○	○	○	○	○	○

→

Figure F.5. Final questionnaire for Experiment 2