Journal of Economic Behavior and Organization 210 (2023) 379–390

Contents lists available at ScienceDirect

Journal of Economic Behavior and Organization

journal homepage: www.elsevier.com/locate/jebo

Discrimination with inaccurate beliefs and confirmation bias[☆]

Christian A. Ruzzier[a,*], Marcelo D. Woo[b]

[a] Department of Economics, Universidad de San Andres, Vito Dumas 284, B1644BID Victoria, Buenos Aires, Argentina
[b] School of Economics, University of Nottingham, University Park, NG7 2RD, Nottingham, United Kingdom

ARTICLE INFO

Article history:
Received 24 February 2022
Revised 9 February 2023
Accepted 17 April 2023
Available online 1 May 2023

JEL classification:
D90
D91
J71
Learning
Confirmation bias
Stereotypes
Discrimination
Segregation
Labor market

ABSTRACT

We examine patterns of discrimination when employers hold incorrect beliefs about the relationship between group membership and productivity, and suffer from confirmation bias when updating their beliefs. As a result, employers do not correct them fully, leading to persistent wage discrimination. Negative stereotypes generate discrimination against minority workers upon entry to the labor market, but are not enough to have discrimination in the long run, and reversals in discrimination are possible. We also discuss whether interventions aimed at reducing discrimination would succeed if confirmation bias is an important source of discrimination, and consider segregation in an extension with heterogeneous employers.

1. Introduction

Discrimination – less favorable treatment of members of a minority group with respect to otherwise identical members of a majority group – continues to be widespread. Systematic evidence of persistent discrimination is provided in Rodgers (2006) and, more recently, in Bertrand and Duflo (2017). Discriminatory treatment in the labor market, in particular, has been credited as the major cause behind the existing disparities among social groups (Darity et al., 1998) – for reviews of theory and empirical findings, see Cain (1986), Altonji and Blank (1999), and Lang and Lehman (2012).

It is customary to classify theoretical models of employment discrimination according to the source of the discrimination. In taste- (or preference-) based theories, originating in the work of Becker (1957), employers suffer a disutility when interacting with members of a particular group. Models of statistical discrimination (Phelps, 1972; Arrow, 1973), on the other hand, assume potential employers have no animus against any particular group, but cannot observe the productivity of individual group members, and thus use group identity to form beliefs about an employee's productivity. For the most part, employers' beliefs are taken as accurate, but there is a recent literature pointing to the importance of statistical discrimination based on inaccurate or incorrect beliefs (see Bohren et al., 2020, for a discussion of the sources for inaccurate beliefs, and a review of the extant evidence).

☆ This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.
* Corresponding author.
E-mail addresses: cruzzier@udesa.edu.ar (C.A. Ruzzier), marcelo.woo@nottingham.ac.uk (M.D. Woo).

https://doi.org/10.1016/j.jebo.2023.04.018
0167-2681/

This paper builds on the inaccurate statistical discrimination literature to examine patterns of discrimination when employers hold incorrect beliefs, and suffer from confirmation bias when evaluating employees from different social groups. Confirmation bias – the tendency 'to misinterpret ambiguous evidence as confirming [one's] current hypothesis about the world' (Rabin and Schrag, 1999) – has been systematically studied by Social Psychology since the 1960s and by Economics more recently.[1] Learning, in turn, is important in labor markets and within organizations, as it has implications for the wages, job assignments, promotions, and sector affiliations of individuals, and for changes in these variables over workers' careers (Gibbons et al., 2005).

There are several good reasons to expect confirmation bias to play an important role in our setting. Whether rooted in cognitive processes (specific cognitive limitations, lack of understanding of logic), or caused by motivational forces (maintain self-esteem and positive regard by others, protect one's ego, avoid certain type of errors, accomplish specific practical goals), and whatever the form confirmation bias takes, the evidence shows that, on the net, when people err, it tends to be in the direction of confirmation (see the extensive and widely cited reviews of Klayman, 1995, and Nickerson, 1998). In the labor market and within organizations, in particular, the specialized press and practitioners in general seem to regard confirmation bias as an important issue, especially in hiring. Wright and Schepker (2015), for instance, blame confirmation bias for mistakes in choosing C-suite candidates. Kahneman et al. (2019) argue further that the same bias is present in unstructured decision-making in general (whether in job interviews or in other, more strategic decisions).[2]

On the other hand, performance evaluations are in many instances plagued by ambiguity and subjectivity, and evidence is open for interpretation. The presence of ambiguous evidence or behavior that must be interpreted is fertile ground for confirmation bias, as is widely acknowledged in the literature (Keren, 1987, Griffin and Tversky, 1992, Klayman, 1995, Rabin and Schrag, 1999). Moreover, research shows that in these contexts, when people perceive themselves as objective evaluators, their judgments tend to be relatively more influenced by stereotypic beliefs –and that, however, many organizational contexts seem to encourage a sense of personal objectivity (Uhlmann and Cohen, 2007). When faced with the overload of information associated with a high number of applicants, employers appear to rely heavily on stereotypes to reduce uncertainty and simplify their decisions, and confirmation bias plays a role in those decisions (Uhlmann and Silberzahn, 2014).

We develop a simple model of wage formation in which potential employers (firms) start with incorrect beliefs about the relationship between group membership (e.g., race or gender) and productivity – that is, firms hold inaccurate stereotypes (cf. Altonji and Pierret, 2001) about workers. Information about workers' productivity is publicly learned over time. Learning is thus symmetric, but when firms update their beliefs about a worker's productivity to set wages, they suffer from confirmation bias, so that firms tend to interpret signals about productivity as if they were closer to their pre-existing beliefs.

In this setup, we show that firms that start with incorrect beliefs do not correct them fully, leading to persistent wage discrimination in the labor market, i.e., different wages for otherwise identical workers belonging to different social groups. Put differently, with confirmation-biased learning, stereotypes have long-lasting effects on wages, whereas with Bayesian learning (as usually assumed) stereotypes would eventually be corrected – discrimination cannot persist in the long run in such a setup.

When workers only differ in group membership and potential employers hold a negative stereotype about the minority group, the minority worker is persistently discriminated against because her expected long-run wage is lower than the long-run wage of an otherwise identical majority worker. The magnitude of the discrimination will depend on the entrenchment of initial beliefs (how strongly held are the initial stereotypes and how willing are potential employers to change them). In the general case, when different groups have different perceived variabilities and the market interprets signals differently depending on the group identity of the worker, negative stereotypes generate discrimination against minority workers upon entry to the labor market, but are not enough to have discrimination in the long run, and reversals in discrimination are possible.

Our setup is amenable to thinking about whether interventions aimed at reducing discrimination would succeed if confirmation bias is an important source of discrimination. For instance, reducing stereotyping is predicted to reduce discrimination, consistent with evidence, but would only eliminate it in the long run in very special circumstances. Attracting employers' attention to within-group differences among minority workers or reducing noise, ambiguity or subjectivity in the evaluation of minority workers all have the potential to reduce discrimination (as some evidence suggests), but will achieve so only if minority workers are perceived as less productive than they truly are at the outset.

We then extend the model to a situation where employers in the same labor market are heterogeneous in terms of their initial beliefs (for example, employers who have the same group identity as a potential employee may have different initial beliefs about worker productivity from employers who do not), and show that the same set of factors that produce discrimination can also produce segregation. The model predicts that, in a cross section of occupations, after controlling for relevant differences across workers, discrimination against minority workers should be more pronounced in partially segre-

[1] Some have gone as far as to state that confirmation bias is "perhaps the best known and most widely accepted notion of inferential error to come out of the literature on human reasoning" (Evans, 1989, p. 41), and that "if one were to attempt to identify a single problematic aspect of human reasoning that deserves attention above all others, the confirmation bias would have to be among the candidates for consideration" (Nickerson, 1998, p. 175).

[2] More systematic evidence is still in short supply (see Whysall, 2018 for a review of cognitive biases and discrimination in recruitment, selection, and promotion), and mainly concerns hiring decisions (see, e.g., Uhlmann and Cohen 2007, and Uhlmann and Silberzahn 2014, on gender discrimination in hiring).

gated occupations than in fully segregated ones. It also predicts, consistent with evidence, a negative correlation between the average pay (of all workers combined) and the proportion of all employees in that job that belong to the minority group.

We model confirmation bias as a source of persistent discrimination among individuals belonging to different social groups, over which society (and hence the market) might have different priors (or stereotypes). Wason (1960), Lord et al. (1979), and Anderson et al. (1980) are important early contributions on confirmation bias in Social Psychology. Within Economics, one of the first papers to address confirmation bias formally was Rabin and Schrag, 1999. In their model, agents receive noisy signals about the state of the world. To model confirmation bias, they assume an agent misinterprets –with exogenous probability– a signal as supporting her current hypothesis. Confirmation bias is shown to lead to overconfidence in the favored hypothesis, and even an infinite amount of information may be insufficient to overcome its effects: over time an agent may with positive probability come to believe with near certainty in the wrong hypothesis.

Where others have used Rabin and Schrag, 1999 to introduce confirmation bias in their models (like Müller, 2010, to study job rotation within a firm as a means to mitigating confirmation bias; Pouget et al., 2017, to analyze the impact of confirmation bias on financial markets; or Rabin and Schrag, 1999 themselves, to explore the consequences of confirmation bias for incentive provision in a principal-agent relationship), we adopt the model of confirmation bias and belief formation proposed by Fryer et al. (2019) –which formalizes the type of confirmation bias described by Lord et al. (1979) and Darley and Gross (1983). Fryer et al. (2019) introduce a human memory storage limitation as a foundation for the bias assumed in Rabin and Schrag (1999), and develop a model of confirmation bias for the case of continuous signals, which provides some new results showing that bias always occurs (with probability one) and depends on early signals and not just the prior.[3]

We use this model as a building block, and add it to a labor market setting with the aim of exploring the consequences of employers being subject to this cognitive bias on the wage gap. In addition, we enrich the market by introducing heterogeneity in prior beliefs (at least some of which will necessarily be inaccurate) about the productivity of different social groups, which allows us to characterize not only differences in wages but also patterns of labor market segregation.

The concept of confirmation bias has seen productive applications in several fields in economics in recent years, like industrial organization (e.g., Mullainathan and Shleifer, 2005), finance (e.g., Pouget et al., 2017), political economy (e.g., Mullainathan and Washington, 2009), marketing (e.g., Narasinham et al., 2005), and organizational design (e.g., Müller, 2010). We contribute to this list by studying the consequences of confirmation bias for (wage) discrimination in the labor market.

The paper also contributes to a growing literature that has been recently expanding the classical models of statistical discrimination.[4] Our work is closest to extensions of classical models that introduce behavioral elements, especially those considering inaccurate beliefs as a source of discrimination.[5] Barron and Ditlmann (2022) differentiate between two different forms of belief-based discrimination, explicit (against women who are equally qualified than men) and implicit (against women who are differently qualified), and produce experimental evidence on these forms. Bohren et al. (2019) develop a model with evaluators who can have a misspecified model of inference about the distribution of ability conditional on gender, or about the preferences or beliefs of other evaluators. They show that such inaccurate beliefs may lead to discrimination based on gender, and enrich the traditional taxonomy of sources of discrimination by distinguishing between preferences (driving taste-based discrimination), rational beliefs (driving accurate statistical discrimination), and inaccurate beliefs (driving inaccurate statistical discrimination).

In Bohren et al. (2019), employers are Bayesian, except for a potential misspecification of their initial beliefs. Our paper departs from their setup by considering employers who, beyond any misspecification in their model of inference, exhibit non-Bayesian learning. This relates our paper to the broader literature on non-Bayesian updating (see the extensive review in Benjamin, 2019). On this front, Campos-Mercade and Mengel (2021) ask experimental subjects to hire one of two potential candidates with potentially different and unobservable productivity, and an observable education signal –and find evidence of substantial conservatism, and of excess discrimination (relative to a Bayesian benchmark), stemming from the neglect of workers' education certificates. Eyting (2022) connects non-Bayesian learning and discrimination based on inaccurate beliefs. In particular, the author identifies the role of motives in generating inaccurate beliefs that lead to discriminatory outcomes. In an experimental setting, Eyting (2022) finds that subjects who play the role of employers exhibit asymmetric learning, weighing more the signals that are aligned with their motives.[6]

2. A model of learning with confirmation bias

We start from a simple dynamic model of learning and wage determination (see, e.g., Freeman, 1977; Harris and Holmstrom, 1982; and Farber and Gibbons, 1996). Consider a competitive labor market in which a large number of identical, risk-neutral firms compete for the labor services of risk-neutral workers of initially unknown productivity $\eta \in \mathbb{R}$. Information is

[3] Schwartzstein (2014) provides a different foundation for confirmation bias, and also provides a discussion of how it can lead to persistent discrimination.

[4] These developments are aptly reviewed in Onuchic (2022).

[5] See, e.g., Bordalo et al. (2016); Bohren et al. (2019, 2020); Coffman et al. (2021a, 2021b, 2021c); Campos-Mercade and Mengel (2021); Erkal et al. (2021); Lepage (2022).

[6] We note that since, in the experiment, the employers' motives are identified by their prior beliefs about the productivity distribution of each group (before an information treatment that corrects such beliefs), the results may also be consistent with confirmation-biased learning.

incomplete but symmetric: both workers and potential employers in the market share the initial belief that η is normally distributed with mean μ_1 and variance $1/h_1$ (where h_1 is the precision). Firms learn about each worker's productivity by observing the worker's output through time ($t = 1, 2, 3, \ldots$), which is given by the following technology:

$$y_t = \eta + \varepsilon_t.$$

The stochastic noise term ε_t is assumed to be independently and identically distributed as $\varepsilon_t \sim \mathcal{N}(0, 1/h_\varepsilon)$ – hence, output is a noisy signal of a worker's true productivity, and learning is gradual.

Workers have the following (publicly-known) utility function:

$$U(w) = \sum_{t=1}^{\infty} \beta^{t-1} w_t,$$

where $\beta \in (0, 1)$ is a discount factor, and w is a wage stream. Only short-term, non-contingent wages can be offered in this market, and we assume that wages are paid in advance. The wage offered at the beginning of any period t may, however, depend on the history of output realizations up to time t, $y^t \equiv (y_1, y_2, \ldots, y_t)$, through the updated beliefs about worker productivity – we denote such a wage by $w_t(y^{t-1})$.

The only decision workers make in this model is which firm they work for. Workers can leave firms in any period at no cost, and will work for the firm offering the highest current wage. Since the labor market is competitive, in each period wages are bid up to expected output (conditional on y^t), that is, until firms earn no profits:

$$w_t(y^{t-1}) = E[y_t \mid y^{t-1}] = E[\eta \mid y^{t-1}]. \tag{1}$$

In each period, firms make wage offers determined by their current beliefs about worker productivity; then each worker chooses for which firm to work, and finally individual outputs for that period are realized and observed by everyone.

To introduce confirmation bias into the learning process, we depart from traditional Bayesian models and consider each potential employer to be an interpretive evaluator as proposed by Fryer et al. (2019). An interpretive evaluator first interprets the signal given her prior, and then updates her beliefs following Bayes' rule, but using her interpretation rather than the raw information. This double updating amounts to weighing the prior belief twice and leads to confirmation bias.

Let $\widehat{y}$ denote the interpretation of the signal y, and $\widehat{\mu}$ the posterior mean based on the interpretation $\widehat{y}$. The two steps involved in the learning process are the following:

a. Interpretation of the signal

$$\widehat{y}_{t+1} = \widehat{\mu}_t \cdot \left(\frac{h_t^\mu}{h_{t+1}^y}\right) + y_t \cdot \left(\frac{h_\varepsilon}{h_{t+1}^y}\right),$$

where h_t^y and h_t^μ denote the precisions of the interpretation $\widehat{y}_t$ and of the interpretive belief $\widehat{\mu}_t$ at time t. The interpretive evaluator interprets (possibly ambiguous) information y_t based on her pre-existing belief $\widehat{\mu}_t$, following Bayes' rule. This has the effect of 'pulling' the signal towards her pre-existing belief, and can be thought of as a model of the 'information assimilation bias' pointed out by Lord et al. (1979).

b. Belief updating based on the interpretation

$$\widehat{\mu}_{t+1} = \widehat{\mu}_t \cdot \left(\frac{h_t^\mu}{h_{t+1}^\mu}\right) + \widehat{y}_{t+1} \cdot \left(\frac{h_{t+1}^y}{h_{t+1}^\mu}\right).$$

Note that firms update their prior beliefs based on their *interpretation*$\widehat{y}$.

Besides the double updating, the learning process is well known (see Degroot, 1970). The posterior means $\widehat{\mu}_{t+1}$ and precisions h_{t+1}^μ are given by:

$$\widehat{\mu}_{t+1} = \mu_1 \cdot \left(\frac{2^t h_1}{2^t(h_1 + h_\varepsilon) - h_\varepsilon}\right) + \left(\frac{2^t h_\varepsilon}{2^t(h_1 + h_\varepsilon) - h_\varepsilon}\right) \cdot \left[\sum_{j=1}^{t} y_j \left(\frac{1}{2}\right)^j\right],$$

$$h_{t+1}^\mu = 2^t(h_1 + h_\varepsilon) - h_\varepsilon.$$

See the online appendix for the derivation of these expressions and an analysis of the evolution of beliefs over time. From (1), the wage function is given by $w_t(y^{t-1}) = \widehat{\mu}_t$.

The expected wage in an equilibrium with confirmation bias is:

$$E[w_t(y^{t-1})] = \mu_1 \cdot \left(\frac{2^{t-1} h_1}{2^{t-1}(h_1 + h_\varepsilon) - h_\varepsilon}\right) + \eta \cdot \left(\frac{2^{t-1} h_\varepsilon}{2^{t-1}(h_1 + h_\varepsilon) - h_\varepsilon}\right) \left[\sum_{j=1}^{t-1} \left(\frac{1}{2}\right)^j\right].$$

As we can see, the confirmation-biased equilibrium wage is determined by the initial belief μ_1 to a larger extent than in the standard Bayesian case, where the weight attached to the prior mean would be $h_1/(h_1 + (t-1)h_\varepsilon)$.

In the long run, as the number of output realizations grows large ($t \to \infty$), the market's posterior mean, and thus wages, converge in expectation to something that is wrong with probability one, as in Fryer et al. (2019), as long as $h_1 > 0$ – a fairly general condition:

$$E[w_t(y^{t-1})] \xrightarrow[t\to\infty]{} \left(\frac{h_1}{h_1 + h_\varepsilon}\right)\mu_1 + \left(\frac{h_\varepsilon}{h_1 + h_\varepsilon}\right)\eta, \tag{2}$$

whereas in the pure Bayesian case (no confirmation bias), $E[w_t(y^{t-1})] \xrightarrow[t\to\infty]{} \eta$. Therefore, there will be a *persistent* effect of the initial belief on wages.

Let w^{LR} denote the (expected) long-run wage (the right-hand side of [2]). The weight of the initial belief μ_1 decreases in the signal-to-prior precision ratio h_ε/h_1. This means that, ceteris paribus, a relatively smaller initial prior precision h_1 (larger prior variance) would move w^{LR} closer to a worker's true productivity. However, the very idea of a significantly large prior variance $1/h_1$ seems to be at odds with the notion of confirmation bias. Increasing the signal precision h_ε has the same effect on w^{LR} as reducing h_1.

We define the retribution bias (*RB*) as the difference between the expected wage with and without confirmation bias.[7] It is straightforward to show that

$$RB \xrightarrow{t\to\infty} (w^{LR} - \eta) = \frac{h_1}{h_1 + h_\varepsilon}(\mu_1 - \eta) \tag{3}$$

As shown by equation [3], in the long run, the retribution bias is determined by the prior-mean error (PME) and the entrenchment of beliefs. The PME is measured by the distance between the initial (subjective) assessment of productivity and the true productivity of the worker, $(\mu_1 - \eta)$. If potential employers are confirmation-biased and hold negative initial beliefs about the productivity of the worker ($\mu_1 < \eta$), there will be persistent under-retributions of his or her productivity, no matter how large the number of signals is. The entrenchment of beliefs is captured by the ratio $h_1/(h_1 + h_\varepsilon)$. The higher the entrenchment (the degree of 'conviction' about one's initial assessments), the greater the retribution bias that will persist given a PME.

3. Discrimination in the labor market

We now extend our learning model with confirmation bias to study patterns of discrimination. Assume, for simplicity, that workers belong to two different social groups or categories (such as gender, ethnicity, or any other), indexed by $g \in \{M, m\}$. Group M is the majority group, while m makes reference to any minority or disadvantaged group facing a negative stereotype (like women, Blacks, Muslims, immigrants, etc.). Employers cannot observe individual worker productivity, but since group membership is observable they can condition their initial beliefs on g, that is:

$$\eta^g \sim \mathcal{N}(\mu_1^g, 1/h_1^g),$$

where μ_1^g is the prior mean and h_1^g is the prior precision for group $g \in \{M, m\}$. We also allow the precision of the signal to differ across groups, and write h_ε^g. To emphasize the role of group identity in discrimination, we will assume that both groups have the same average true productivity, but that employers have negative stereotypes (generalizations about groups that are applied to individual group members simply because they belong to that group; Heilman, 2012) about the minority group: minority workers are seen as less productive, so that $\mu_1^m < \mu_1^M$.[8,9]

Consider two individuals with the same true productivity $\overline{\eta}$, but belonging to different groups. Upon entry to the labor market, there will be (statistical) discrimination against the minority worker, because with no information other than group identity, $w_1^m = \mu_1^m < \mu_1^M = w_1^M$. But will discrimination persist? Most economic models that regard pre-market decisions as exogenous have shown that statistical discrimination can indeed persist, provided the number of signals is limited (Arrow, 1973; Aigner and Cain, 1977; Altonji and Pierret, 2001; Bohren et al., 2019). However, with Bayesian learning (which those models assume), wages would eventually converge to true productivity as the number of signals grows large, and hence both workers would expect to receive the same wage – discrimination cannot persist in the long run in such a setup. As hinted above, this need not be the case with confirmation-biased learning.[10]

Under confirmation-biased learning, the long-run wages for two individuals belonging to different groups but who are otherwise identical (i.e., have equal productivities and output histories) will be:

$$w_M^{LR} = \mu_1^M\left(\frac{h_1^M}{h_1^M + h_\varepsilon^M}\right) + \bar{\eta}\left(\frac{h_\varepsilon^M}{h_1^M + h_\varepsilon^M}\right), \text{ and}$$

[7] In the limit, *RB* can be interpreted as a measure of economic discrimination due to confirmation-biased learning. In the limit, $RB < 0$ implies that the worker is not receiving 'pay or remuneration commensurate with their productivity' (Aigner and Cain, 1977), as measured by η.

[8] Our assumption corresponds to what (Bohren et al., 2019) call biased belief-based partiality (towards members of group M).

[9] Because groups have the same true average productivity, beliefs are necessarily inaccurate in the sense of Bohren et al. (2020).

[10] Elmslie and Sedo (1996) and Goldsmith et al. (2004) propose an alternative mechanism that produces long-run effects of initial discrimination based upon Festinger's theory of cognitive dissonance. In response to discrimination, workers may adjust their beliefs about the quality of the job that they can expect to attain, thereby reducing their labor supply. Early discrimination may then negatively affect future employability and wages. The model focuses on the effects of initial discrimination on the supply side of labor, whereas ours focuses on biased beliefs by employers.

$$w_m^{LR} = \mu_1^m \left(\frac{h_1^m}{h_1^m + h_\varepsilon^m} \right) + \bar{\eta} \left(\frac{h_\varepsilon^m}{h_1^m + h_\varepsilon^m} \right).$$

We define long-run discrimination as $LRD \equiv w_m^{LR} - w_M^{LR}$. Therefore,

$$LRD = \left[\mu_1^m \left(\frac{h_1^m}{h_1^m + h_\varepsilon^m} \right) - \mu_1^M \left(\frac{h_1^M}{h_1^M + h_\varepsilon^M} \right) \right] + \bar{\eta} \left[\left(\frac{h_\varepsilon^m}{h_1^m + h_\varepsilon^m} \right) - \left(\frac{h_\varepsilon^M}{h_1^M + h_\varepsilon^M} \right) \right].$$

Letting $x \equiv h_1/h_\varepsilon$ we can write equivalently:

$$LRD = \left[\mu_1^m \left(\frac{x^m}{1 + x^m} \right) - \mu_1^M \left(\frac{x^M}{1 + x^M} \right) \right] + \bar{\eta} \left[\left(\frac{1}{1 + x^m} \right) - \left(\frac{1}{1 + x^M} \right) \right]. \tag{4}$$

A minority worker is discriminated against when her expected long-run wage is lower than the long-run wage of an otherwise identical majority worker; that is, when $LRD < 0$.

Before deriving general conditions for $LRD < 0$, let us analyze particular cases. In the spirit of Phelps (1972), and similar to Coate and Loury (1993) and Bohren et al. (2019), assume that groups only differ in their prior means – i.e., $h_1^M = h_1^m = h_1$ and $h_\varepsilon^M = h_\varepsilon^m = h_\varepsilon$, or $x^m = x^M = x$. Then:

$$LRD = \left(\mu_1^m - \mu_1^M \right) \left(\frac{x}{1 + x} \right).$$

It is clear that initial group-level discrimination, $\mu_1^m < \mu_1^M$, leads to persistent between-group discrimination under confirmation bias – i.e., $LRD < 0$ for all x.[11] The magnitude of the discrimination will depend on the entrenchment of beliefs, as measured by $x/(1 + x) = h_1/(h_1 + h_\varepsilon)$. A relatively high h_1 (stronger *a priori* convictions about initial beliefs) leads to more discrimination for a given difference in initial beliefs, whereas a higher h_ε (better chances to convey information through less noisy signals) moves LRD towards zero.

An alternative set of parameters explored in the literature assumes, contrary to what we have done here, that prior means are identical across groups $\mu_1^m = \mu_1^M = \mu_1$ (so that individuals are perceived to have the same average productivity), but that precisions differ. In that case,

$$LRD = (\mu_1 - \bar{\eta}) \left(\frac{x^m}{1 + x^m} - \frac{x^M}{1 + x^M} \right).$$

For instance, signals might be noisier for minority workers (Phelps, 1972; Aigner and Cain, 1977; Lundberg and Startz, 1983; Lundberg, 1991; Cornell and Welch, 1996; Farmer and Terrell, 1996), and hence $h_\varepsilon^m < h_\varepsilon^M$. Lower signal precision could also be interpreted as greater subjectivity in judgment, as in Bohren et al. (2019). Alternatively, potential employers might think that minority group members tend to resemble each other quite a bit, and thus $h_1^m > h_1^M$: the minority group has a lower perceived variability (Park and Judd, 1990).[12] Either assumption (and the existing evidence) suggests that $x^m > x^M$ is the most relevant case. Then there will be long-run discrimination as long as the PME is negative for these individuals – i.e., whenever $\mu_1 < \bar{\eta}$ and individuals' productivities are under-appreciated by the market at the outset.

We can now turn to the exploration of the general case (with $\mu_1^m < \mu_1^M$). To that end, it is useful to write (4) as follows:

$$LRD = (\mu_1^m - \bar{\eta}) \frac{x^m}{1 + x^m} - \left(\mu_1^M - \bar{\eta} \right) \frac{x^M}{1 + x^M}. \tag{5}$$

Discrimination, a feature of behavior, depends on the primitives of the model, and can go either way. Notice in particular that $\mu_1^m < \mu_1^M$ is not sufficient to have $LRD < 0$, and that reversals in discrimination are possible.[13] The following proposition summarizes the results of a sensitivity analysis:

Proposition 1. *Long-run discrimination against the minority worker ($LRD < 0$):*

- *obtains whenever $\mu_1^m < \bar{\eta} < \mu_1^M$, for any x^m,x^M, and is increasing in both x^m and x^M in this case;*
- *is more likely the lower μ_1^m and the higher μ_1^M;*
- *is more (less) likely the higher is x^m whenever $\mu_1^m < \bar{\eta}$ ($\mu_1^m > \bar{\eta}$); and*
- *is less (more) likely the higher is x^M whenever $\mu_1^M < \bar{\eta}$ ($\mu_1^M > \bar{\eta}$).*

The proof follows from simple inspection of (5) and is omitted.

Proposition 1 helps us think about the expected effects of policies aimed at reducing discrimination under confirmation bias. Procedures that involve hiding the group identity of the worker at the time of evaluation (like the blind auditions

[11] Discrimination can also arise as the result of favoritism or partiality towards the in-group rather than hostility against out-groups. Ahmen (2007) shows experimental evidence of this.

[12] This may be related to the outgroup homogeneity effect (Judd and Park, 1988; Linville, 1998) – the tendency for people to see outgroup members as more alike than ingroup members – if potential employers belong to the majority group.

[13] For example, set $\mu_1^M = 3$, $\mu_1^m = 2.8$, $\eta = 1$,$x^M = 1$, $x^m = 2$. Then $\mu_1^m - \mu_1^M = -0.2$ but $LRD = +0.2$.

analyzed by Goldin and Rouse (2000); or the double-blind refereeing in academic journals discussed by Blank (1991)) would be akin to making $\mu_1^m = \mu_1^M$. Such interventions would eliminate discrimination upon entry to the labor market and would reduce discrimination in the long run – but would not eliminate discrimination unless $x^M = x^m$. Empathy training programs have been shown to reduce stereotyping (Aboud and Levy, 2000), which would amount to bringing μ_1^m closer to μ_1^M in our setup, and we predict that this intervention would reduce discrimination, consistent with evidence (Mcgregor, 1993; Batson et al., 1997; Stephan and Finlay, 1999; and Galinsky and Moskowitz, 2000).

Attracting individuals' attention to within-group differences in the minority group would imply a reduction in h_1^m and hence in x^m. This approach has the potential to reduce discrimination if the minority worker was under-appreciated to begin with (i.e., if $\mu_1^m < \bar{\eta}$). Evidence from laboratory and field experiments suggests that increasing the perceived variability of the minority group indeed reduces discrimination – it might even reduce stereotyping, which would reinforce the effect (see., e.g., Brauer and Er-Rafiy, 2011). Reducing noise in the measurement of minority workers' output – that is, increasing h_ε^m – would also imply a reduction in x^m, and have similar effects on discrimination. Evidence consistent with reduced discrimination when minority workers' signals are interpreted with greater precision is provided in Sarsons et al. (2021) and Bohren et al. (2019). Increased h_ε^m can also be interpreted as reduced ambiguity, which makes discrimination less likely to arise according to evidence in Nieva and Gutek (1980) and Helman and Haynes (2008).

The *LRD* in Eq. (4) is an absolute-level wage gap, and it is a function of the workers' underlying productivity, $\bar{\eta}$. As one moves up the income distribution (i.e., as the wages of BOTH workers, w_m^{LR} and w_M^{LR}, increase), the relative size of this gap might change. Since a simultaneous increase in both wages can only occur through an increase in the workers productivity, the question boils down to whether the absolute wage gap, *LRD*, increases or decreases with underlying productivity, $\bar{\eta}$.[14] Taking the derivative of Eq. (4) with respect to quickly shows that:

$$\frac{\partial LRD}{\partial \bar{\eta}} < 0 \iff x^m > x^M \tag{6}$$

As we have discussed previously, $x^m > x^M$ is the most likely case. Hence, as one moves up the income distribution, the wage gap would be reduced. If we associate higher wages and higher productivity with higher skill jobs, then this prediction fits nicely with the existing evidence on the wage gap between blacks and whites in the US - one of the main stylized facts found by Lang and Lehman (2012) in their extensive review of racial discrimination in the labor market. To the extent that more education is associated to higher productivity, the prediction also squares well with the evidence in Ñopo et al. (2012) on international gender wage gaps.

4. Discrimination and segregation

In the previous section we have characterized patterns of wage discrimination in labor markets in which all employers statistically discriminate based on the same (incorrect) beliefs and suffer from confirmation bias. A natural and important next step to consider would be a situation where employers in the same labor market are heterogeneous in terms of their initial beliefs (for example, employers who have the same group identity as a potential employee may have different initial beliefs about worker productivity from employers who do not). Such an extension would also allow us to discuss whether the same set of factors that produce a pay differential would also produce segregation (Blau and Jusenius, 1976). Pay gaps and occupational segregation by group membership have long been associated with employment discrimination (Bergman, 1974) and tend to coexist in the labor market. Furthermore, occupational segregation by gender, for instance, explains most of the gender wage gap (Bielby and Baron, 1986; Groshen, 1991; Blau and Khan, 2017).

We continue to assume that workers belong to group $g \in \{M, m\}$, and that employers cannot observe individual worker productivity, but only their group belonging. Each firm has only one position to fill, and hence, can hire at most one worker in each period. A firm that does not hire a worker makes zero profits in that period. We model the heterogeneity of employers by assuming that they differ in their initial beliefs about each type of worker, μ_{fg}. For simplicity, we assume precisions are equal across employer types $f \in \{M, m\}$.[15]

As before, labor supply is perfectly inelastic, so wages are determined solely by the demand side, and we assume workers are interchangeable in production. Employers, on the other hand, hold different initial beliefs on worker productivity, i.e., $\mu_{Mg} \neq \mu_{mg}$ for some g. Let $F_f > 0$ denote the number of type f firms, $F \equiv F_M + F_m$, and $\pi^f \equiv F_f / F$. Equivalently for workers, we define $L \equiv L_M + L_m > 0$, and $p^g \equiv L_g / L$. We assume $F > L$ (there are more firms than workers overall), and $p^m > \pi^m$ (e.g., because minority workers are underrepresented in managerial positions).

We consider a competitive labor market with no search costs, so that in each period all firms can make bids to all workers at no cost. Given our assumptions, workers need only look at current wage offers and accept the best offer in each period. If an outside offer ties the offer of the current employer, we assume the worker stays.[16]

[14] We thank an anonymous reviewer for suggesting this comparative statics exercise.

[15] In choosing the same labels for worker and employer groups we are focusing on employers who may or may not have the same group identity as potential employees. Alternatively, different employers could have the same beliefs within groups, but with different groups interpreted as different occupations (like 'masculine' and 'feminine').

[16] Alternatively, we can assume there exists an infinitesimal cost of changing jobs.

Firms compete à la Bertrand in wage offers each period. To determine wage offers by firms, first note that in any period $t > 0$ the willingness to pay (WTP) of a type-f firm for the services of a worker of group g with a signal history y_i^{t-1} is determined by its current confirmation-biased belief about that worker's productivity, which is given by:

$$\hat{\mu}_t\left(\mu_{fg}, y^{t-1}\right) = \mu_{fg}\left(\frac{2^{t-1}h_1}{2^{t-1}(h_1 + h_\varepsilon) - h_\varepsilon}\right) + \hat{\upsilon}_t\left(y^{t-1}\right) \tag{7}$$

where

$$\hat{\upsilon}_t\left(y^{t-1}\right) \equiv \sum_{j=1}^{t-1} y_j\left(\frac{2^{t-1-j}h_\varepsilon}{2^{t-1}(h_1 + h_\varepsilon) - h_\varepsilon}\right)$$

summarizes the *value of a history of signals at t* (i.e., its contribution to the posterior belief). For any t, conditional on a history y^{t-1}, the WTP is strictly increasing in the firm's prior μ_{fg}. Moreover, two workers with the same output history would receive different wage offers to the extent that the firm's prior beliefs on them differ.[17]

Bertrand competition implies that wages offered by all firms to any given worker (of group g with history y^{t-1}) will be determined by the WTP of the *marginal* employer of that social group. The marginal employer of group g will be the firm type with the highest WTP for such a worker if there is no ingroup congestion for group g ($L_g < F_g$). If there is congestion in one group, the marginal employer is the firm type with the lowest WTP, and the other firm type can lower its equilibrium wage offer to that of the marginal employer and enjoy a positive expected profit.

Many cases can arise depending on the ordering of the different initial beliefs of employers. For the sake of brevity, we focus here on the cases we regard as most relevant or plausible.[18] To this end, we assume the following:

- (A1) $\mu_{MM} > \mu_{Mm}$ (i.e., employers in the majority group regard majority workers as more productive on average than minority workers)
- (A2) $\mu_{mm} > \mu_{Mm}$ (i.e., minority workers are regarded more favorably by minority employers than by majority employers)
- (A3) $\mu_{MM} > \mu_{mM}$ (i.e., majority workers are regarded more favorably by majority employers than by minority employers)

This ordering corresponds to a case in which prior beliefs favor ingroup matching (homophily): for each worker group g, the ingroup employer has a higher prior than the outgroup employer (and, conditional on history, a higher WTP). Other assumptions might well generate different predictions, but ingroup favoritism seems like a natural starting point for exploring the consequences of confirmation bias in a setting with heterogeneous employers –and there is plenty of evidence (see, e.g., Lewis and Sherman, 2003; Nunley et al., 2011; Doleac and Stein, 2013; Wright and Schepker, 2015; Whysall, 2018; Kline et al., 2022) that such ingroup biases are pervasive in this context, making it a plausible assumption.

Suppose first that there is no ingroup congestion (NIC), i.e., $F_g > L_g\ \forall g$. Then, under (A1)-(A3) there exists a unique full segregation equilibrium in which all firms offer wages that are equal to their current beliefs $\hat{\mu}_t\left(\mu_{fg_i}, y_i^{t-1}\right)$ about the productivity of each worker, M-workers accept offers from M-firms (which are higher than those offered to them by m-firms), and m-workers accept offers from m-firms (which are higher than those offered to them by M-firms).[19] In this equilibrium all firms make zero expected profits and there will be discrimination except in the case in which $\mu_{MM} = \mu_{mm}$.

In contrast, if NIC is violated (i.e., there is a worker group g that saturates job slots, $L_g > F_g$), then under (A1)-(A3) there is a partial segregation equilibrium.[20] For concreteness, assume it is the minority group that cannot find jobs only in minority firms. In such an equilibrium all M-workers are employed by M-firms, which pay their WTP and hence make no expected profits. Some m-workers cannot be employed by m-firms due to slot constraints, and hence M-firms become their marginal employer. This drives m-firms' wage offers down to that of M-firms, and all firms pay in equilibrium a wage equal to the lowest WTP, that of the M-type.

m-workers employed by M-firms receive a lower wage than M-workers with a similar history of signals employed by the same type of firms (because $\mu_{MM} > \mu_{Mm}$ by (A1)). m-workers employed by m-firms are also discriminated against, as they receive wages lower than comparable M-workers in M-firms. Therefore, minority workers are discriminated against by both ingroup and outgroup firms.

When firms have heterogeneous prior beliefs, the equilibrium wage gap between M and m workers with the same history y^{t-1} is given by:

$$w_t^m\left(m, y_i^{t-1}\right) - w_t^M\left(M, y_i^{t-1}\right) = \left(\frac{2^{t-1}h_1}{2^{t-1}(h_1 + h_\varepsilon) - h_\varepsilon}\right)(\mu_{Mm} - \mu_{MM}) \tag{8}$$

under partial segregation, and by

$$w_t^m\left(m, y_i^{t-1}\right) - w_t^M\left(M, y_i^{t-1}\right) = \left(\frac{2^{t-1}h_1}{2^{t-1}(h_1 + h_\varepsilon) - h_\varepsilon}\right)(\mu_{mm} - \mu_{MM}) \tag{9}$$

[17] Once again, due to confirmation bias, differences will persist in the long run.

[18] The analysis of every case not covered here is available from the authors upon request.

[19] While segregation depends on beliefs in the model, beliefs could in turn be affected by segregation. See Levy and Razin (2017) for a model of the coevolution of segregation, beliefs, and discrimination.

[20] Duncan and Duncan's (1955) index of dissimilarity (or segregation) would be equal to $(F_g/L_g) * 100$, which is below 100.

under full segregation. The full segregation wage gap is negative for any t (discrimination against minority workers) if and only if $\mu_{mm} < \mu_{MM}$, whereas it is always negative under partial segregation by (A1). Whatever the equilibrium, discrimination is a function of a) the entrenchment of prior beliefs, and b) the difference in prior means of the *marginal employers* of m- and M-workers.

Long-run discrimination is given by

$$\lim_{t\to\infty}\left(\frac{2^{t-1}h_1}{2^{t-1}(h_1+h_\varepsilon)-h_\varepsilon}\right)(\mu_{mm}-\mu_{MM}) = \left(\frac{h_1}{h_1+h_\varepsilon}\right)(\mu_{mm}-\mu_{MM}) \tag{10}$$

under full segregation, and by

$$\lim_{t\to\infty}\left(\frac{2^{t-1}h_1}{2^{t-1}(h_1+h_\varepsilon)-h_\varepsilon}\right)(\mu_{Mm}-\mu_{MM}) = \left(\frac{h_1}{h_1+h_\varepsilon}\right)(\mu_{Mm}-\mu_{MM}) \tag{11}$$

under partial segregation.

In a cross section of occupations, after controlling for relevant differences across workers, discrimination against minority workers should be more pronounced in partially segregated occupations than fully segregated ones for any t. From (8) and (9), the difference between the short-run partial-segregation and the short-run full-segregation wage gaps is given by

$$\left(\frac{2^{t-1}h_1}{2^{t-1}(h_1+h_\varepsilon)-h_\varepsilon}\right)(\mu_{Mm}-\mu_{mm}),$$

which is negative by (A2).

The comparison between the full- and partial-segregation equilibria shows that M-firms pay less to their workers on average under partial segregation, which squares well with the substantial evidence showing a negative correlation between the average pay (of men and women combined) and the proportion of all employees in that job that is female, if we take women as the minority group; see, e.g., Killingsworth (1987); Macpherson and Hirsch (1995), and Boraas and Rodgers (2003).

In a time series, both in full- and partial-segregation equilibria, discrimination against m-workers is worse in the short run, consistent with evidence showing that labor-market experience mitigates discrimination against, e.g., immigrants (Baert et al., 2017; and Fays et al., 2020).

5. Discrimination and competition

Wherever stereotypes come from (e.g., from employers using the representative heuristic of Bordalo et al., 2016), our model shows how confirmation-biased employers tend to perpetuate their consequences for wage gaps. Hence, discriminatory behavior survives the passing of time. But will discriminators survive in the long run? This is a related, but different, question. As Schwartzstein, 2014, (p. 1447) emphasizes, "the logic of confirmation bias does not by itself pin down which incorrect beliefs we can expect to persist". Thus, something else is needed to pin down beliefs. Competition among employers might achieve precisely that.

The issue has been discussed at least since Becker, 1957 in the context of taste-based discrimination, where relative wages of minority workers are determined by the most prejudiced employer with whom they come into contact – the marginal discriminator (Charles and Guryan, 2008). Therefore, it is the marginal, rather than the average, prejudice of employers that determines the wage gap.

This logic of the marginal competitor applies equally well to models of statistical discrimination like ours (see, e.g., Nunley et al., 2011). To be concrete, assume n potential employers have only one position to fill, and hence, can hire at most one worker in each period. A firm that does not hire a worker makes zero profits in that period. Employers bid each period for the services of a worker from group $g \in \{M, m\}$. All employers have the same belief about the productivity of M-workers, $\mu_i^M = \mu^M$ for all i, and they only differ in the belief they have on the productivity of m-workers, μ_i^m $(i = 1, \dots, n)$. In particular, employers can be ranked according to how biased they are against m-workers:

$$\mu_i^m = \mu^M - b_i$$

with n being the most biased employer, and $b_i \in (0, \mu^M)$ for all i.

Firms compete à la Bertrand in wage offers each period. For any t, conditional on a history y^{t-1}, the willingness to pay (WTP) is strictly increasing in the firm's prior μ_i^g. Moreover, two workers with the same output history would receive different wage offers to the extent that the firm's prior beliefs on them differ. Irrespective of actual employer, the wage of a majority worker will be determined by μ^M, but the wage of a minority worker will be determined by the WTP of the second highest bidder –the marginal employer –i.e., $\mu_2^m = \mu^M - b_2$. In general, with n firms bidding for the services of $k < n$ workers of group m, whether offers for workers are all simultaneous or sequential, the worst-off m-worker's wage will be determined by $\mu_{k+1}^m = \mu^M - b_{k+1}$.

The marginal employer of group m will then be the $(k+1)$-th firm. Increased competition in the labor market, construed as an increase in the number of employers, will reduce discrimination if it reduces the bias of the marginal employer – i.e., if entrants come from the left tail of the employer bias distribution (the most likely case, since they are the ones who can profit from entry). Because less biased employers have a higher willingness to pay, increasing the number of employers in the labor market raises the probability that the k firms with the highest willingness to pay are less biased, and thus raises the wage offer (Nunley et al., 2011).

Employers with more extreme biases are "driven out of the market" as competition increases (they cease to be the marginal employer). We thus expect labor markets with higher competitive pressure on the demand side to show smaller wage gaps (consistent with evidence in, e.g., Meng, 2004; Nunley et al., 2011; Caminade et al., 2012; Doleac and Stein, 2013).

Our setup is more amenable to thinking about the consequences of competition in the labor market than in the product market in which our employers operate. In the context of taste-based discrimination, prejudiced employers sacrifice profits by discriminating. As the usual argument goes, such employers would be ultimately driven from the market in the long run in a competitive setting (Arrow, 1972; Comanor, 1973). More recent work has shown that wage gaps can persist in the long run if there is some form of imperfect information, imperfect competition, or adjustment costs (Charles and Guryan, 2008). All these arguments refer to competition in the product market. It is not clear, however, what the relationship between product-market competition and statistical discrimination is (Berkovec et al., 1998; Heyman et al., 2013; Cooke et al., 2019; Fays et al., 2020). For instance, if increased competitive pressure forces firms to improve management practices, it could improve the screening of job candidates, reducing the reliance on group stereotypes (Heyman et al., 2013; Cooke et al., 2019). Improved screening could be interpreted as a reduced x^m in our model. Whether this would reduce (or increase) discrimination, however, depends also on the other parameters of the model, as can be seen in equation (5).

6. Concluding remarks

We have developed a model of inaccurate statistical discrimination that highlights confirmation bias as a source of discrimination. We have shown that discrimination in the long run depends not only on the stereotypes that apply to different social groups, but also on how potential employers evaluate a worker's output and how willing they are to modify their initial beliefs. We have used the model to predict how different interventions to reduce discrimination would work in the face of confirmation-biased learning, and to show how the same factors that generate discrimination can also produce segregation.

The limitations of our modeling assumptions are worth mentioning at this point. Our focus has been on labor-market discrimination, which means that we have omitted a worker's pre-labor market human capital investments (see, e.g., Coate and Loury, 1993; or Farmer and Terrell, 1996). Because we had a single job in the analysis, we were not able to address job assignment and promotions, which are important ingredients of a fully-fledged theory of wage dynamics and careers in organizations (Gibbons and Waldman, 1999a) – extending our model in this direction, for instance along the lines of Gibbons and Waldman (1999b), would be straightforward and is an interesting avenue for future research.

We have explored the consequences of incorrect beliefs and confirmation bias for wage discrimination, and generated several predictions. We view this as a theoretical article and have not attempted a rigorous empirical evaluation. However, to flesh out the analysis, we have reported some evidence from other empirical studies that is consistent with those predictions, and may be useful in thinking about the determinants of wage discrimination. It is beyond the scope of our article to provide more systematic evidence on the predictions of the model, but designing an experiment with appropriate parameterizations to directly test the theoretical predictions should be high on the research agenda. Since our model refers to the persistence of wage discrimination, rather than to binary hiring decisions (as in much of the empirical work on non-Bayesian discrimination), we must recognize some of the challenges involved in designing such an experiment, in particular, the elicitation of long-run beliefs.

Declaration of Competing Interest

None.

Data availability

No data was used for the research described in the article.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.jebo.2023.04.018

References

Aboud, F.E., S.R. Levy, 2000. Interventions to reduce prejudice and discrimination in children and adolescents. In: Oskamp S. (Ed.), Reducing Prejudice and Discrimination. Lawrence Erlbaum Associates, Inc., Mahwah, NJ.
Ahmen, A.M., 2007. Group identity, social distance, and intergroup bias. J Econ Psychol 28 (3), 324–337. doi:10.1016/j.joep.2007.01.007.
Aigner, D.J., Cain, G.G., 1977. Statistical theories of discrimination in labor markets. Ind. Labor Relat. Rev. 30 (2), 175–187. doi:10.1177/001979397703000204.
Altonji, J.G., Blank, R.M., 1999. Race and gender in the labor market. In: Ashenfelter, O., Card, D. (Eds.), Handbook of Labor Economics, Vol. 3C. North-Holland, Amsterdam doi:10.1016/S1573-4463(99)30039-0.
Altonji, J.G., Pierret, C.R., 2001. Employer learning and statistical discrimination. Q J Econ 116 (1), 313–350. doi:10.1162/003355301556329.
Anderson, C.A., Lepper, M.R., Ross, L., 1980. Perseverance of social theories: the role of explanation in the persistence of discredited information. J Pers Soc Psychol 39 (6), 1037–1049. doi:10.1037/h0077720.

Arrow, K.J., 1972. Some mathematical models of race in the labor market. In: Pascal, A.H. (Ed.), Racial Discrimination in Economic Life. Lexington Books.
Arrow, K.J., 1973. The Theory of Discrimination. In: Ashenfelter, O., Rees, A. (Eds.), Discrimination in Labor Markets. Princeton University Press, Princeton.
Baert, S., Albanese, A., Du Gardein, S., Ovaere, J., Stappersa, J., 2017. Does work experience mitigate discrimination? Econ. Lett. 155, 35–38. doi:10.1016/j.econlet.2017.03.011.
Barron, K., Ditlmann, R., Gehrig, S., Schweighofer-Kodritsch, S., 2022. Explicit and implicit belief-based gender discrimination: A hiring experiment. CESifo Working Paper Series 9731, CESifo. https://doi.org/10.2139/ssrn.4097858.
Batson, C.D., Sager, K., Garst, E., Kang, M.S., Rubchinsky, K., Dawson, K., 1997. Is empathy-induced helping due to self–other merging? J Pers Soc Psychol 73 (3), 495–509. doi:10.1037/0022-3514.73.3.495.
Becker, G.S., 1957. The Economics of Discrimination. University of Chicago Press, Chicago.
Benjamin, D.J., 2019. Errors in probabilistic reasoning and judgment biases. In: Bernheim, B.D., DellaVigna, S., Laibson, D. (Eds.), Handbook of Behavioral Economics – Foundations and Applications, Vol. 2. North Holland doi:10.1016/bs.hesbe.2018.11.002.
Bergman, B.R., 1974. Occupational segregation, wages and profits when employers discriminate by race or sex. East Econ J 1 (2), 103–110.
Berkovec, J.A., Canner, G.B., Gabriel, S.A., Hannan, T.H., 1998. Discrimination, competition, and loan performance in FHA mortgage lending. Rev Econ Stat 80 (2), 241–250. doi:10.1162/003465398557483.
Bertrand, M., Duflo, E., 2017. Field experiments on discrimination. In: Banerjee, A.V., Duflo, E. (Eds.), Handbook of Economic Field Experiments, Vol. 1. North-Holland, Amsterdam doi:10.1016/bs.hefe.2016.08.004.
Bielby, W.T., Baron, J.N., 1986. Men and women at work: sex segregation and statistical discrimination. American Journal of Sociology 91 (4), 759–799. doi:10.1086/228350.
Blank, R., 1991. The effects of double-blind versus single-blind refereeing: experimental evidence from the american economic review. Am Econ Rev 81 (5), 1041–1067. https://www.jstor.org/stable/2006906
Blau, F.D., Jusenius, C.L., 1976. Economists approaches to sex segregation in the labor market: an appraisal. Signs (Chic) 1 (3), 181–199. doi:10.1086/493286.
Blau, F.D., Khan, L.M., 2017. The gender wage gap: extent, trends, and explanations. J Econ Lit 55 (3), 789–865. doi:10.1257/jel.20160995.
Bohren, A., Haggag, K., Imas, A., Pope, D., 2020. Inaccurate statistical discrimination: an identification problem. PIER Working Paper 19-010, July. Available at https://ssrn.com/abstract=3406060.
Bohren, A., Imas, A., Rosemberg, M., 2019. The dynamics of discrimination: theory and evidence. Am Econ Rev 109 (10), 3395–3436. doi:10.1257/aer.20171829.
Boraas, S., Rodgers, IIIW.A., 2003. How does gender play a role in the earnings gap? an update. Monthly Labor Review 126 (3), 9–15.
Bordalo, P., Coffman, K., Gennaioli, N., Shleifer, A., 2016. Stereotypes. Q J Econ 131 (4), 1753–1794. doi:10.1093/qje/qjw029.
Brauer, M., Er-Rafiy, A., 2011. Increasing perceived variability reduces prejudice and discrimination. J Exp Soc Psychol 47 (5), 871–881. doi:10.1016/j.jesp.2011.03.003.
Cain, G.G., 1986. The economic analysis of labor market discrimination: a survey. In: Ashenfelter, O., Layard, R. (Eds.), Handbook of Labor Economics, Vol. 1. North-Holland, Amsterdam doi:10.1016/S1573-4463(86)01016-7.
Caminade, J., List, J. A., Livingston, J. A., Picel, J., 2012. When the weak become weaker: The effect of market power on third degree price discrimination. Unpublished manuscript, Bentley University, May.
Campos-Mercade, P., Mengel, F., 2021. Non-bayesian statistical discrimination. Available at SSRN. https://doi.org/10.2139/ssrn.3843579.
Charles, K.K., Guryan, J., 2008. Prejudice and wages: an empirical assessment of Becker's the economics of discrimination. Journal of Political Economy 116 (5), 773–809. doi:10.1086/593073.
Coate, S., Loury, G.C., 1993. Will affirmative-action policies eliminate negative stereotypes? Am Econ Rev 83 (5), 1220–1240. https://www.jstor.org/stable/2117558
Coffman, K. B., Collis, M., Kulkarni, L., 2021a. Stereotypes and belief updating. HBS Working Paper 19–068, Harvard Business School, June.
Coffman, K.B., Flikkema, C.B., Shurchkov, O., 2021b. Gender stereotypes in deliberation and team decisions. Games Econ Behav 129, 329–349. doi:10.1016/j.geb.2021.06.004.
Coffman, K. B., Ugalde-Araya, P., Zafar, B., 2021c. A (dynamic) investigation of stereotypes, belief-updating, and behavior. NBER Working Paper 29382, National Bureau of Economic Research, October. https://www.nber.org/papers/w29382.
W.S. Comanor, 1973, Racial discrimination in American industry, Economica, 40(160), 363-378, doi.org/10.2307/2553319.
Cooke, D., Fernandes, A.P., Ferreira, P., 2019. Product market competition and gender discrimination. Journal of Economic Behavior and Organization 157, 496–522. doi:10.1016/j.jebo.2018.10.005.
Cornell, B., Welch, I., 1996. Culture, information, and screening discrimination. J Polit Econ 104 (3), 542–571. doi:10.1086/262033.
Darity, J.R., William, A., Mason, P.L., 1998. Evidence on discrimination in employment: codes of color, codes of gender. Journal of Economic Perspectives 12 (2), 63–90. doi:10.1257/jep.12.2.63.
Darley, J.M., Gross, P.H., 1983. A hypothesis-confirming bias in labelling effects. J Pers Soc Psychol 44 (1), 20–33.
Degroot, M., 1970. Optimal Statistical Decisions. McGraw Hill, New York.
Doleac, J.L., Stein, L.C.D., 2013. The visible hand: race and online market outcomes. The Economic Journal 123 (572), F469–F492. doi:10.1111/ecoj.12082.
Duncan, O., Duncan, B., 1955. A methodological analysis of segregation indexes. Am Sociol Rev 20 (2), 210–217. doi:10.2307/2088328.
Elmslie, B., Sedo, S., 1996. Discrimination, social psychology, and hysteresis in labor markets. J Econ Psychol 17 (4), 465–478. doi:10.1016/0167-4870(96)00021-9.
Erkal, N., Gangadharan, L., Koh, B.H., 2021. Gender biases in performance evaluation: the role of beliefs versus outcomes. University of East Anglia School of Economics Working Paper Series 2021-09, School of Economics, University of East Anglia, Norwich, UK. https://ueaeco.github.io/working-papers/papers/ueaeco/UEA-ECO-21-09.pdf
Evans, J.S.B.T., 1989. Bias in Human Reasoning: Causes and Consequences. Erlbaum, Hillsdale, NJ.
Eyting, M., 2022. Why do we discriminate? the role of motivated reasoning. SAFE Working Paper, No. 356, Leibniz Institute for Financial Research SAFE, Frankfurt a. M., September. https://doi.org/10.2139/ssrn.4210315.
Farber, H.S., Gibbons, R., 1996. Learning and wage dynamics. Q J Econ 111 (4), 1007–1047. doi:10.2307/2946706.
Farmer, A., Terrell, D., 1996. Discrimination, Bayesian updating of employer beliefs, and human capital accumulation. Econ Inq 34 (2), 204–219. doi:10.1111/j.1465-7295.1996.tb01373.x.
Fays, V., Mahy, B., Rycx, F., Volral, M., 2020. Wage discrimination based on the country of birth: do tenure and product market competition matter? Appl Econ 53 (13), 1–21. doi:10.1080/00036846.2020.1838431.
Freeman, S., 1977. Wage trends as performance displays productive potential: a model and application to academic early retirement. Bell Journal of Economics 8 (2), 419–443. doi:10.2307/3003295.
Fryer, R.G., Harms, P., Jackson, M.O., 2019. Updating beliefs when evidence is open to interpretation: implications for bias and polarization. J Eur Econ Assoc 17 (5), 1470–1501. doi:10.1093/jeea/jvy025.
Galinsky, A.D., Moskowitz, G.B., 2000. Perspective-taking: decreasing stereotype expression, stereotype accessibility, and in-group favoritism. J Pers Soc Psychol 78 (4), 708–724. doi:10.1037/0022-3514.78.4.708.
Gibbons, R., Katz, L.F., Lemieux, T., Parent, D., 2005. Comparative advantage, learning, and sectoral wage determination. J Labor Econ 23 (4), 681–723. doi:10.1086/491606.
Gibbons, R., Waldman, M., 1999a. Careers in organizations: theory and evidence. In: Ashenfelter, O., Card, D. (Eds.), Handbook of Labor Economics, Vol 3B. North-Holland, Amsterdam doi:10.1016/S1573-4463(99)30022-5.
Gibbons, R., Waldman, M., 1999b. A theory of wage and promotion dynamics inside firms. Q J Econ 114 (4), 1321–1358. doi:10.1162/003355399556287.

Goldin, C., Rouse, C., 2000. Orchestrating impartiality: the impact of “blind” auditions on female musicians. Am Econ Rev 90 (4), 715–741. doi:10.1257/aer.90.4.715.

Groshen, E.L., 1991. The structure of the female/male wage differential: is it who you are, what you do, or where you work? Journal of Human Resources 26 (3), 457–472. doi:10.2307/146021.

Harris, M., Holmstrom, B., 1982. A theory of wage dynamics. Review of Economic Studies 49 (3), 315–333. doi:10.2307/2297359.

Heilman, M.E., 2012. Gender stereotypes and workplace bias. Res Organ Behav 32, 113–135. doi:10.1016/j.riob.2012.11.003.

Helman, M.E., Haynes, M.C., 2008. Subjectivity in the appraisal process: a facilitator of gender bias in work settings. In: Borgida, E., Fiske, S.T. (Eds.), Beyond Common Sense: Psychological Science in the Courtroom. Blackwell Publishing, Malden, MA doi:10.1002/9780470696422.ch7.

Heyman, F., Svaleryd, H., Vlachos, J., 2013. Competition, takeovers, and gender discrimination. ILR Review 66 (2), 409–432. doi:10.1177/001979391306600205.

Judd, C.M., Park, B., 1988. Out-group homogeneity: judgments of variability at the individual and group levels. J Pers Soc Psychol 54 (5), 778–788. doi:10.1037/0022-3514.54.5.778.

Killingsworth, M.R., 1987. Heterogeneous preferences, compensating wage differentials, and comparable worth. Q J Econ 102 (4), 727–742. doi:10.2307/1884278.

Kline, P.M., Rose, E.K., Walters, C.R., 2022. Systemic discrimination among large u.s. employers. Q J Econ 137 (4), 1963–2036. doi:10.1093/qje/qjac024.

Lang, K., Lehman, J.Y.K., 2012. Racial discrimination in the labor market: theory and empirics. J Econ Lit 50 (4), 959–1006. doi:10.1257/jel.50.4.959.

Lepage, L. P., 2022. Bias Formation and Hiring Discrimination. Unpublished manuscript, Stockholm University, March.

Levy, G., Razin, R., 2017. The coevolution of segregation, polarized beliefs, and discrimination: the case of private versus state education. American Economic Journal: Microeconomics 9 (4), 141–170. doi:10.1257/mic.20160201.

Lewis, A.C., Sherman, S.J., 2003. Hiring you makes me look bad: social-identity based reversals of the ingroup favoritism effect. Organ Behav Hum Decis Process 90 (2), 262–276. doi:10.1016/S0749-5978(02)00538-1.

Linville, P.W., 1998. The heterogeneity of homogeneity. In: Darley, J.M., Cooper, J. (Eds.), Attribution and social interaction: The legacy of Edward E. Jones. American Psychological Association, Washington, DC doi:10.1037/10286-008.

Lord, C.G., Ross, L., Lepper, M.R., 1979. Biased assimilation and attitude polarization: the effects of prior theories on subsequently considered evidence. J Pers Soc Psychol 37 (11), 2098–2109. doi:10.1037/0022-3514.37.11.2098.

Lundberg, S.J., 1991. The enforcement of equal opportunity laws under imperfect information: affirmative action and alternatives. Q J Econ 106 (1), 309–326. doi:10.2307/2937919.

Lundberg, S.J., Startz, R., 1983. Private discrimination and social intervention in competitive labor markets. Am Econ Rev 73 (3), 340–347. https://www.jstor.org/stable/1808117

Macpherson, D.A., Hirsch, B.T., 1995. Wages and gender composition: why do women’s jobs pay less? J Labor Econ 13 (3), 426–771. doi:10.1086/298381.

Mcgregor, J., 1993. Effectiveness of role playing and antiracist teaching in reducing student prejudice. J Educ Res 86 (4), 215–226. doi:10.1080/00220671.1993.9941833.

Meng, X., 2004. Gender earnings gap: the role of firm specific effects. Labour Econ 11 (5), 555–573. doi:10.1016/j.labeco.2003.09.006.

Mullainathan, S., Shleifer, A., 2005. The market for news. Am Econ Rev 95 (4), 1031–1053. doi:10.1257/0002828054825619.

Mullainathan, S., Washington, E., 2009. Sticking with your vote: cognitive dissonance and political attitudes. American Economic Journal: Applied Economics 1 (1), 86–111. doi:10.1257/app.1.1.86.

Müller, D., 2010. On horns and halos: Confirmation bias and job rotation. bonn econ discussion paper 05/2010. Bonn Graduate School of Economics.

Narasinham, C., He, C., Anderson, E.T., Brenner, L., Desai, P., Kuksov, D., Messinger, P., Moorthy, S., Nunes, J., Rottenstreich, Y., Staelin, R., Wu, G., Zhang, Z.J., 2005. Incorporating behavioral anomalies in strategic models. Mark Lett 16 (3/4), 361–373. doi:10.1007/s11002-005-5898-9.

Nickerson, R.S., 1998. Confirmation bias: a ubiquitous phenomenon in many guises. Review of General Psychology 2 (2), 175–220. doi:10.1037/1089-2680.2.2.175.

Nieva, V.F., Gutek, B.A., 1980. Sex effects on evaluation. Academy of Management Review 5 (2), 267–276. doi:10.5465/amr.1980.4288749.

Ñopo, H., Daza, N., Ramos, J., 2012. Gender earnings gaps around the world: a study of 64 countries. Int J Manpow 33 (5), 464–513. doi:10.1108/01437721211253164.

Nunley, J.M., Owens, M.F., Howard, R.S., 2011. The effects of information and competition on racial discrimination: evidence from a field experiment. Journal of Economic Behavior & Organization 80 (3), 670–679. doi:10.1016/j.jebo.2011.06.028.

Onuchic, P., 2022. Recent contributions to theories of discrimination. papers 2205.05994. ArXiv.org, revised Dec 2022. https://doi.org/10.48550/arXiv.2205.05994.

Park, B., Judd, C.M., 1990. Measures and models of perceived group variability. J Pers Soc Psychol 59 (2), 173–191. doi:10.1037/0022-3514.59.2.173.

Phelps, E.S., 1972. The statistical theory of racism and sexism. Am Econ Rev 62 (4), 659–661. https://www.jstor.org/stable/1806107

Pouget, S., Sauvagnat, J., Villeneuve, S., 2017. A mind is a terrible thing to change: confirmatory bias in financial markets. Rev Financ Stud 30 (6), 2066–2109. doi:10.1093/rfs/hhw100.

Rabin, M., Schrag, J.L., 1999. First impressions matter: a model of confirmatory bias. Q J Econ 114 (1), 37–82. doi:10.1162/003355399555945.

Rodgers, IIIW.M., 2006. Handbook on the Economics of Discrimination. Edward Elgar Publishing, Inc., Northampton, MA.

Sarsons, H., Gerxhani, K., Reuben, E., Schram, A., 2021. Gender differences in recognition for group work. J Polit Econ 129 (1), 101–147. doi:10.1086/711401.

Schwartzstein, J., 2014. Selective attention and learning. J Eur Econ Assoc 12 (6), 1423–1452. doi:10.1111/jeea.12104.

Stephan, W.G., Finlay, K., 1999. The role of empathy in improving intergroup relations. Journal of Social Issues 55 (4). doi:10.1111/00224537.00144.

Wason, P.C., 1960. On the failure to eliminate hypotheses in a conceptual task. Q J Exp Psychol 12 (3), 129–140. 10.1080%2F17470216008416717

Whysall, Z., 2018. Cognitive biases in recruitment, selection, and promotion: the risk of subconscious discrimination. In: Caven, V., Nachmias, S. (Eds.), Hidden Inequalities in the Workplace. Palgrave Macmillan doi:10.1007/978-3-319-59686-0_9.

Wright, P.M., Schepker, D.J., 2015. Why boards go awry in their hiring decisions. The Wall Street Journal. https://www.wsj.com/articles/why-boards-go-awry-in-their-hiring-decisions-1445824934