Quarterly Journal of Political Science, 2023, 18: 513–542

Ban the Box? Information, Incentives, and Statistical Discrimination

John W. Patty[1] and Elizabeth Maggie Penn[2]*

[1] *Professor of Political Science and Quantitative Theory & Methods, Emory University, Atlanta, GA, USA; jwpatty@gmail.com*
[2] *Professor of Political Science and Quantitative Theory & Methods, Emory University, Atlanta, GA, USA; elizabeth.m.penn@gmail.com*

ABSTRACT

"Banning the Box" refers to a policy campaign aimed at prohibiting employers from soliciting applicant information that could be used to statistically discriminate against categories of applicants (in particular, those with criminal records). In this article, we examine how the concealing or revealing of informative features about an applicant's identity affects hiring both directly and, in equilibrium, by possibly changing applicants' incentives to invest in human capital. We show that there exist situations in which an employer and an applicant are in agreement about whether to ban the box. Specifically, depending on the structure of the labor market, banning the box can be (1) Pareto dominant, (2) Pareto dominated, (3) benefit the applicant while harming the employer, or (4) benefit the employer while harming the applicant. Our results have policy implications spanning beyond employment decisions, including the use of credit checks by landlords and standardized tests in college admissions.

Keywords: Discrimination; racial profiling; equity; moral hazard

*We thank anonymous reviewers, Scott Ashworth, Jon Bendor, Steve Callander, Tom Clark, Josh Clinton, Daniel Diermeier, Marc Fleurbaey, Dana Foarta, Michael Hanley, Navin Kartik, Jenny Kim, Cesar Martinelli, Andrea Moro, Marcus Pivato, Mattias Polborn, Aaron Roth, Miguel Rueda, Mehdi Shadmehr, Jessica Sun, Scott Tyson, seminar audiences at Columbia, Emory, NYU, Rochester, Stanford, University of Montreal, UNC Chapel Hill, Vanderbilt, and the 2021 Online Social Choice Seminar for incredibly helpful comments.

Online Appendix available from:
http://dx.doi.org/10.1561/100.00022021_app
MS submitted on 4 February 2022; final version received 18 November 2022
ISSN 1554-0626; DOI 10.1561/100.00022021

Introduction

Discrimination is pervasive across political, economic, and social settings, including the markets for housing, credit, and employment. Eliminating it is a longstanding and vexing policy challenge. Several widely discussed approaches to achieving more equitable outcomes involve eliminating the possibility of discrimination by withholding sensitive information about individuals. This approach is sometimes referred to as "fairness as blindness," and has been utilized in a variety of important decision-making contexts. These contexts include college admissions (need-blindness), symphony orchestras (blind auditions), and, our focus in this article, "Ban the Box" (or BTB) employment policies that withhold information concerning felony status from potential employers.

We present a simple model of the possible effects of this "informational" approach to reducing discrimination. Our primary motivation is to understand the conditions under which such methods can improve outcomes in a moral hazard model of employment. Such models have played a central role in understanding statistical discrimination (see the section "Related Literature"), and our model is directly inspired by/borrowed from this literature. A strength of this approach for understanding discrimination is that it incorporates the possible *endogeneity* of a worker's qualifications to the information that they expect the employer to use when deciding whether to hire him or her.

A second, closely related, motivation is to understand when such "fairness as blindness" methods are robust in the sense that they benefit *both* workers and employers. This is where we obtain the most important conclusion from our model: there is a large set of situations in which "blinding the employer" to an applicant's sensitive information (e.g., race, gender, ethnicity, credit history, or criminal record) can make both employers and applicants better off *in equilibrium*. This conclusion's implications are two-fold. In addition to demonstrating that blindness can yield a Pareto improvement, this conclusion *requires* the possibility of endogeneity between the applicants' qualifications for the job and the employer's hiring rule.

"Ban the Box" Policies

Our theory is motivated by a policy proposal popularly known as "Ban the Box". Such policies have been both adopted voluntarily by some employers, such as Starbucks, Target, and Walmart, and imposed by law in various states and localities. In practice, BTB policies generally preclude employers from considering a job applicant's criminal history at least initially in the

hiring process.[1] Empirically, the overall effect of these policies on employment outcomes is at best mixed: there is evidence of BTB policies decreasing discrimination against applicants with a felony conviction, but increasing discrimination against Black (and sometimes Latino) applicants in the United States.[2] This prompts the question of what we can learn about the potential effects from a theoretical perspective.[3] The answer to this question — namely whether the negative effect of discrimination on all workers under BTB outweighs the positive effect of eliminating disparate treatment of members of the disadvantaged group — depends on many factors, including "the extent to which those with criminal histories benefit from suppressing information, the extent to which those without criminal histories are harmed, and the relative size of these two classes of applicants within the group itself" (Raphael, 2020, p. 7).

For simplicity, our theory considers only two groups of workers, and consequently we describe these two groups as "convicted felons" and "all other people." However, the debate about BTB policies often involves discussions of more than two groups (e.g., the intersection of felon status and race, leading to at least four "groups" of workers). Specifically, much of the debate in the United States is about how omitting information about felon status will differentially affect workers of color relative to white workers. Interpreting our theory's conclusions in the context of this much richer debate requires one to step back from the model and, for example, consider the analysis in parallel — one set of parameters for (say) white workers and another set of parameters for workers of color. Placing these parallel analyses side-by-side allows us to consider the impact of BTB on the outcomes experienced by workers from different racial groups as a function of whether information about the second, "felon status" group membership is included in the hiring process.[4] Consequently,

[1]The point at which such consideration is permissible varies across jurisdictions. It is common for the prohibition to extend until a conditional offer of employment is made. A full consideration of the effect of this timing is very interesting. However, space precludes us from treatment of this issue in this article.

[2]See, among others, Doleac and Hansen (2020).

[3]For a comprehensive review of the theoretical and empirical literatures about BTB, see Raphael (2020).

[4]This article's focus is closely related to the emerging literature on "algorithmic fairness" (e.g., Patty and Penn, 2022), and a subtle point with respect to this connection is that many notions of algorithmic fairness are *dyadic* in nature in the sense that they judge the fairness of an algorithm (e.g., the employer's hiring process) on the basis of comparisons between pairs of groups. Accordingly, while interpreting the model with more than two groups is more challenging, this is not the result of our assumption that there are only two groups. Rather, it is due to the conceptual ambiguity of simultaneously comparing more than two groups. For similar reasons, we do not explore other important distinctions, such as having been convicted of a felony versus having been charged, but not convicted. We thank an anonymous reviewer for suggesting that we clarify this point.

we believe that our focus on two groups does not undermine application of the model's insights to situations with more than two groups. Moreover, although BTB is frequently discussed regarding consideration of *criminal history* in the hiring process, the logic identified by our model regarding the ambiguity of BTB's effects on both employment and welfare extends to the effects of including other types of information, including information about credit history or standardized testing, in evaluative processes such as housing applications and college admissions (e.g., Bartik and Nelson, 2019; Maturana *et al.*, 2020; Niu *et al.*, 2022).

The Many Dimensions of Ban the Box: Our analysis focuses on one human capital dimension of the possible impacts of BTB. The potential social, economic, and policy impacts of BTB are much broader than this, of course. For example, criminal history might be per se relevant for certain jobs and careers (such as law enforcement and/or positions requiring access to classified information). Furthermore, employment in the real world often involves multiple individuals working together and/or with third parties (such as clients). Accordingly, it is important to note that we abstract from these and other ethical/efficiency concerns one might consider when choosing whether to adopt a BTB-style policy in the first place. While this obviously limits the scope of our analysis, we believe it clarifies one approach to considering the various feedback loops that might govern the actual impacts of informational policies such as BTB. In addition, it arguably strengthens the robustness of our main finding — that sometimes BTB can be Pareto optimal for employees and employers — by setting aside other potential salutary effects of BTB to focus on one dimension of the policy's impact. We now turn to describe the model.

Overview of the Model and Results

Our model of the labor market is stylized and focuses only on the moral hazard dimension of the employment market, in that an employer is unable to directly observe an applicant's fitness for a job. We construct a model of incomplete and imperfect information with two players, a **worker**, W, and an **employer**, E. The worker can acquire **qualification** for the job ($q = 1$) or not ($q = 0$), with an eye toward gaining employment by the employer. The worker's cost of acquiring qualification, c, is private information to the worker, and his or her decision to acquire qualification is also private information. The employer then observes a noisy, but positively correlated signal of the worker's qualification, $\theta(q)$. We refer to $\theta(q)$ as the worker's **test result**. Conditional on this result, the employer decides whether to hire the worker ($h = 1$) or not ($h = 0$). The game then concludes, and the two players receive payoffs based on the worker's choice of qualification, q, and the employer's hiring decision, h.

We add a wrinkle to this standard model of moral hazard by considering the effect of the employer's information about the distribution of the worker's cost of qualification, c. We assume that there are two groups of workers, distinguished only by their distributions of costs: one group of workers has a distribution of costs that first order stochastically dominates the other group's distribution of costs. The group with higher costs to qualification is referred to as the *disadvantaged group* and the other group is referred to as the *advantaged group*. When the box is present, the employer can observe both the worker's group identity and the worker's test result. When the box is banned, the employer can only observe the worker's test result.

Our theory is aimed at considering how removing the information about group identity required for "direct" discrimination can affect investment, employment, and welfare, in equilibrium. To keep our analysis as compact as possible without distorting the underlying analysis, we rule out wage discrimination and focus solely on employment discrimination in terms of differential standards for employment at a prevailing, common market wage.[5]

We show that, in equilibrium, blindness can (1) hurt employers while helping workers, (2) hurt workers while helping employers, (3) hurt both employers and workers, (4) help both employers and workers, or (5) have no effect. Thus, within our setting, the welfare effects of attempting to achieve fairness through blindness are truly ambiguous without more information. The specific equilibrium arrived at depends on the distribution of costs for workers; the market wage relative to this cost; the relative sizes of the two groups; and the informativeness of the test result that the employer observes.

The intuition for our results hinges on the fact that, when the box is banned, the employer must pool the workers from the advantaged and disadvantaged groups, and set a common hiring standard for both. When the advantaged group is comparatively large relative to the disadvantaged group, this pooling serves as a commitment device for the employer to hire more aggressively from the disadvantaged group. More aggressive hiring, in turn, can make investment in qualification more attractive to the disadvantaged group.

The above dynamic can ultimately be Pareto improving for both the employer and the worker, as it grows the pool of talent from which the employer hires, while simultaneously raising the prospects of employment for the worker. At the same time, if the advantaged group is comparatively small relative to the disadvantaged group then banning the box can be Pareto dominated. In this case, when the groups are pooled the employer may no longer be able to commit to hiring anyone. No individual obtains qualification in equilibrium, and the labor market shuts down. Interestingly, there are also regions of the parameter space in which banning the box induces the

[5]We discuss relaxing the assumption that the employer is a wage taker in the Extensions and Conclusion section.

advantaged group to obtain qualification at a higher rate, by committing the employer to hire from this group *less* aggressively.

Related Literature

A long literature on discrimination[6] has noted that it can arise from various sources, including a "taste" for one group over another (e.g., Becker, 1971), belief-based "statistical" discrimination (e.g., Arrow, 1973; Phelps, 1972), and implicit biases in evaluating individuals (e.g., Bertrand *et al.*, 2005).[7] A feature common to all three of these sources of discrimination is that the employer must be able to observe (or infer) applicants' group memberships.[8] Consequently, as we have discussed, eliminating or withholding this information may help forestall discrimination at its root source.[9]

Models of Discrimination: Becker (1971) presented the seminal analysis of the economics of *taste-based* discrimination. Phelps (1972) and Arrow (1973) presented early models of statistical discrimination, including how such discrimination can be self-enforcing. Coate and Loury (1993) extended this line of inquiry to consider how discrimination affects the incentive to invest in human capital and, relatedly, whether affirmative action policies might break this self-enforcing nature. Moro and Norman (2004) combine the Arrow (1973) and Coate and Loury (1993) models within a task-assignment context. Fryer Jr (2007) considers the interaction of discrimination at the hiring stage and in subsequent promotion decisions. Bjerk (2008) extends the study of dynamic, statistical discrimination by considering how differences in an employer's informational precision early in an applicant's career might affect the promotion path.

Of these, the model developed by Coate and Loury (1993) is the most closely related to ours,[10] so it is useful to consider the distinction between our model and theirs. In Coate and Loury's model, the two groups of applicants

[6]While the term "discrimination" itself has a wide array of closely related definitions, in this article we say that discrimination occurs whenever a decision-maker treats one group of applicants differently than another group, simply as a function of their group memberships (i.e., holding other factors equal).

[7]For a review of theories of statistical discrimination, see Fang and Moro (2011).

[8]We use hiring as our running example in this article, but the implications are more general in scope.

[9]A related issue (that for reasons of space we do not confront as squarely as we could in this article) is the degree to which employers can, or should, infer sensitive information about applicants from seemingly innocuous covariates. Our theory does indicate the importance of this question to the degree that it clearly, if partially, illustrates the situations in which such an incentive would emerge *in equilibrium*.

[10]We say this because Coate and Loury's model was part of the inspiration for this research. We thank a reviewer for pointing out an arguably tighter conceptual linkage between our analysis and the model developed by Lundberg and Startz (1998).

are identical from an ex ante perspective. In our model, the two groups are different in the sense that one group faces lower barriers to qualification than the other. It is important to note that this assumption is conservative relative to theirs in the sense that it offers an initial justification for the employer to include the box to distinguish between the two groups. This is conservative because our *main conclusion is that there will still exist situations in which the employer strictly benefits from banning the box.*

In Coate and Loury's model (as in Arrow's), discrimination can emerge in equilibrium in spite of the fact that the groups are identical in ex ante terms. Discrimination in such settings results from equilibrium multiplicity in the labor market and occurs when each applicant's group membership essentially serves as an equilibrium selection device. Thus, within their setting, the impact of banning the box would depend upon which equilibrium would be played in the absence of the box. Similarly, our model — where the groups of workers are not identical from an ex ante perspective — may also support multiple equilibria. However, our arguments do not leverage this multiplicity: we focus throughout only on the (generically unique) Pareto efficient equilibrium. Accordingly, our comparison of labor markets with and without the box presumes that the box plays no role in equilibrium selection.[11]

Models of Information and Discrimination: We are not the first to consider the impact of an employer's information about applicants on hiring and employment outcomes (e.g., Arrow, 1973; Autor and Scarborough, 2008; Bartik and Nelson, 2019; Phelps, 1972) but, to our knowledge, we are among the first to consider the impact of this information on *hiring* when the worker's qualification for the job is endogenously determined. Most closely related along these lines is the model developed by Kim and Loury (2019), who allow for endogenous qualification *and* endogenous group identification. The two main differences between their analysis and ours are that (1) we consider the impact of the availability of the information (i.e., "blindness") on equilibrium outcomes and (2) Kim and Loury's model allows employers to offer different wages to workers from different groups. The first difference — the effects of different information structures on discrimination and outcomes — is our central interest in this analysis. On the second difference, our model follows a common tradition in this literature (e.g., Coate and Loury, 1993).[12]

[11] This equilibrium selection issue and how it connects a few seemingly disparate models of statistical discrimination, is also addressed briefly in Patty and Penn (2022).

[12] While we have examined the model with endogenous wages (see the section "Extensions and Conclusion"), we have not explored the possibility of "limited wage flexibility." Specifically, we conjecture that our analysis is robust to the employer having *some* ability to offer different wages to different groups of workers, as long as the ability to set different wages does not allow the employer to ensure that he or she would use the same hiring rule (in terms of the signal, $\theta(q)$) for both groups.

The Model

Our theory is based on a two-player game involving a **worker**, W, and an **employer**, E. In order to better illustrate the incentives of this baseline model, we consider first a setting with only one group of workers.

The Worker's Information and Potential: The worker has a (privately observed) binary real-valued type, $c \in C \equiv \{c_L, c_H\}$, with $0 < c_L < c_H$. The type determines the cost of becoming qualified ($q = 1$) relative to remaining unqualified ($q = 0$) and is distributed as follows:

$$\Pr[c = c_L] = p,$$
$$\Pr[c = c_H] = 1 - p.$$

As mentioned in the introduction, we refer to the parameter p as the **potential** of the worker's group. This is because, in the cases of interest in our analysis (Assumption 1), p is the maximum ex ante probability that a worker in that group might actually be qualified in equilibrium.

The Worker's Choices: The worker first observes his or her **cost of qualification**, $c \in \{c_L, c_H\}$ (with $0 < c_L < c_H$), and then chooses whether to become qualified (denoted by $q = 1$) or not (denoted by $q = 0$). If the worker chooses to become qualified, he or she incurs a net cost of c.

The Employer's Information: The worker's qualification (i.e., W's choice of q) is not directly observed by the employer. Rather, an informative — but noisy — signal of his or her choice, denoted by $\theta \in \Theta \equiv \{1, 2, 3\}$, is generated as follows:

	$q = 0$	$q = 1$
$\Pr[\theta = 1 \mid q]$	ϕ_0	0
$\Pr[\theta = 2 \mid q]$	$1 - \phi_0$	$1 - \phi_1$
$\Pr[\theta = 3 \mid q]$	0	ϕ_1

(1)

Note that if E observes either $\theta = 1$ or $\theta = 3$, then the test result reveals the qualification of the worker, q, with certainty. On the other hand, a test result of $\theta = 2$ is a "**garbled test result**" that can potentially be sent by both qualified and unqualified types. Accordingly, for each qualification choice, $q \in \{0, 1\}$, $\phi_q \in (0, 1)$ is the conditional probability that θ is "correct" in the sense of revealing q. We refer to the conditional distribution of θ described in (1) as a *test* of qualification, so that θ represents the *outcome* of the worker's test.

The Employer's Choices: After (1) the worker's cost of qualification, c, is realized by the worker, (2) the worker chooses his or her qualification, q, and (3) conditional on W's choice of q, the test result θ is realized and observed by

the employer, the employer then finally chooses whether to hire W (denoted by $h = 1$) or not (denoted by $h = 0$).

Sequence of Play: Summarizing the description above, our model's decision sequence is as follows:

1. the worker observes c,
2. the worker chooses $q \in \{0, 1\}$,
3. the employer observes θ,
4. the employer chooses $h \in \{0, 1\}$,
5. the process concludes and players receive their payoffs.

Payoffs: The players' payoffs, given c, q, and h, are as follows:

$$\begin{aligned} u_W(q, h \mid c) &= wh - cq, \\ u_E(h \mid q, \theta) &= (Bq - w)h, \end{aligned} \tag{2}$$

where $w > 0$ and $B > w$ are exogenous parameters that are assumed to be common knowledge throughout. The parameter w represents the **wage** paid by E to W if E hires W, and $B > w$ represents E's **benefit** from hiring ($h = 1$) a qualified worker ($q = 1$). Finally, as noted earlier, $c \in \{c_L, c_H\}$ captures W's cost of obtaining qualification.

Strategies: A (possibly mixed) qualification strategy for W is a mapping $\chi : C \to [0, 1]$, where $\chi(c) \equiv \Pr[q = 1 \mid c]$ denotes the probability that the worker chooses $q = 1$, given his or her cost, c. Similarly, a (possibly mixed) hiring strategy for the employer is a mapping $\eta : \Theta \to [0, 1]$, where $\eta(\theta) \equiv \Pr[h = 1 \mid \theta]$ denotes the probability E hires W, for each $\theta \in \Theta$. We refer to the employer's hiring strategy, η, as **aggressive** when $\eta(2) = 1$, **conservative** when $\eta(2) = 0$, and **mixed** when $\eta(2) \in (0, 1)$.

Beliefs: The employer's beliefs about q, given θ, are denoted by $\mu(\theta) \equiv \Pr[q = 1 \mid \theta]$. Our equilibrium concept of choice, sequential equilibrium, will require that these beliefs be correct. We now turn to the analysis of the model.

Equilibrium Analysis

Our equilibrium concept is sequential equilibrium (Kreps and Wilson, 1982), a refinement of perfect Bayesian equilibrium. Sequential equilibria are typically more complicated to verify than perfect Bayesian equilibria, but have the benefit of ruling out some perfect Bayesian equilibria in which E holds "unreasonable off the path beliefs." In our setting, this refinement is particularly useful because

it rules out an otherwise ubiquitous perfect Bayesian "pooling" equilibrium in which the employer never hires workers (even after observing $\theta = 3$) and workers never become qualified. This is not a sequential equilibrium: in any sequential equilibrium, E's beliefs about q must satisfy the following:[13]

$$\mu(3) = 1.$$

With this in hand, we can simplify notation and write E's beliefs simply as $\mu \equiv \mu(2) \in [0,1]$.[14] E's beliefs, μ, are consistent with W's strategy, χ, if μ satisfies the following:

$$\mu = \frac{(1-\phi_1)(p\chi(c_L) + (1-p)\chi(c_H))}{\begin{array}{l}((1-\phi_1)(p\chi(c_L) + (1-p)\chi(c_H)) + (1-\phi_0)(p(1-\chi(c_L)) \\ \quad + (1-p)(1-\chi(c_H)))\end{array}}. \tag{3}$$

Equilibrium Hiring: The sequentially rational hiring strategy for E, given μ, is defined by the following:

$$\eta(\theta \mid \mu) = \begin{cases} 0 & \text{if } \theta = 1, \\ 0 & \text{if } \theta = 2 \text{ and } \mu < \dfrac{w}{B}, \\ 1 & \text{if } \theta = 2 \text{ and } \mu > \dfrac{w}{B}, \\ 1 & \text{if } \theta = 3, \end{cases} \tag{4}$$

and any hiring probability is sequentially rational conditional upon $\theta = 2$ and $\mu = \frac{w}{B}$. With this in hand, we will write E's strategy simply as $\eta \equiv \Pr[h = 1 \mid \theta = 2]$.

Equilibrium Qualification: Turning to the worker, first note that if $w < c_L$, then $q = 1$ is strictly dominated for W, so that $\chi(c_L) = 0$ and $\eta = 0$ in any equilibrium. On the other hand, when $c_H < w$, there may exist equilibria in which all workers obtain qualification with probability 1 (i.e., $\chi(c_L) = \chi(c_H) = 1$). While these equilibria are interesting in their own right, they do not accurately reflect the role we intend for the parameter p to play in the model — an upper bound on the probability that $q = 1$ (i.e., the maximum "equilibrium potential" of the worker's group). Accordingly, we

[13] To see this, consider any sequence of "fully mixed" strategies by the worker, $\{\chi_\tau\}_{\tau=1}^{\infty}$ with $\chi_\tau(c) \in (0,1)$ for both $c \in \{c_L, c_H\}$, and consider the sequence of beliefs, $\{\mu_\tau^*\}_{\tau=1}^{\infty}$ such that μ_τ^* is consistent with χ_τ and Bayes's rule for each $\tau \in \{1,2,\ldots,\}$. This is uniquely defined for each $\tau \in \{1,2,\ldots\}$ and satisfies the following: $\mu_\tau^* = 1$ for all $\tau \in \{1,2,\ldots\}$.

[14] The structure of the payoffs in (2), along with the assumption that $\phi_0 > 0$, imply that $\Pr[\theta = 1] > 0$ in any Bayes Nash equilibrium of this model, so that Bayes's rule implies that $\mu(1) = 0$ in any Bayes Nash equilibrium.

assume throughout that $c_L < w < c_H$, so that $q = 1$ is strictly dominated if $c = c_H$ but not strictly dominated when $c = c_L$ (this does not imply that $q = 1$ is a best response for the worker when $c = c_L$).

Assumption 1. *Qualification is strictly dominated for W conditional on $c = c_H$ and costly, but not strictly dominated, conditional on $c = c_L$:*

$$0 < c_L < w < c_H.$$

With Assumption 1 in hand, we simplify notation and write W's strategy simply as $\chi \equiv \Pr[q = 1 \mid c = c_L]$ (i.e., $\chi(c_H) = 0$ in all equilibria). We begin with E's sequentially rational hiring decision conditional on χ and $\theta = 2$. E is willing to hire (i.e., to set $\eta > 0$) only if

$$\Pr[q = 1 \mid \theta = 2, \chi, p, \phi] = \frac{(1-\phi_1)p\chi}{(1-\phi_1)p\chi + (1-\phi_0)(1-p+p(1-\chi))} \geq \frac{w}{B}. \tag{5}$$

Similarly, conditional on $c = c_L$ and the strategy η by E, it is incentive compatible for W to play strategy $\chi > 0$ only if

$$\Pr[h = 1 \mid q = 1, \eta, \phi] = \frac{1}{\phi_1 + \eta(\phi_0 - \phi_1)} \leq \frac{w}{c_L}. \tag{6}$$

Eqs. (5) and (6) give us two cases to consider, distinguished by the employer's sequentially rational decision conditional on $\theta = 2$ when E believes $\chi = 1$. If

$$\frac{(1-\phi_1)p}{(1-\phi_1)p + (1-\phi_0)(1-p)} \geq \frac{w}{B},$$

then E is willing to hire upon observing $\theta = 2$ if he or she believes that $\chi = 1$. Accordingly, in this case there exists an equilibrium with $\chi = 1$ if and only if

$$w \geq \frac{c_L}{\phi_0}. \tag{7}$$

On the other hand, if inequality (5) fails to hold with $\chi = 1$, then E is unwilling to hire conditional on $\theta = 2$ regardless of χ. In this case there exists an equilibrium with $\chi = 1$ if and only if w is sufficiently high and/or $\theta = 3$ is sufficiently likely, conditional on $q = 1$:

$$w \geq \frac{c_L}{\phi_1}. \tag{8}$$

Finally, it may be the case that the Pareto efficient equilibrium is a mixed strategy equilibrium, with W using a nondegenerate mixed strategy conditional upon $c = c_L$, and E using a nondegenerate mixed strategy conditional upon $\theta = 2$. In this case Eqs. (5) and (6) must hold with equality, implying

1. the test is more accurate conditional on being qualified than not: $\phi_0 < \phi_1$,
2. the wage is sufficiently high to sustain positive qualification: $w \geq \frac{c_L}{\phi_1}$, and
3. the players' equilibrium strategies are described by the following:

$$\eta_M \equiv \eta_M(w, \phi, c_L) = \frac{w\phi_1 - c_L}{w(\phi_1 - \phi_0)}, \tag{9}$$

$$\chi_M(p) \equiv \chi_M(p, B, w, \phi, c_L) = \frac{w(1-\phi_0)}{p(B(1-\phi_1) + w(\phi_1 - \phi_0))}. \tag{10}$$

We now define the employer's **hiring threshold**, denoted by p_E^*, as the probability of qualification that makes E indifferent about hiring W after observing a garbled test result ($\theta = 2$) conditional on W becoming qualified if and only if W's cost of qualification is $c = c_L$ (i.e., $\chi = 1$):[15]

$$p_E^* \equiv \frac{w(1-\phi_0)}{B(1-\phi_1) + w(\phi_1 - \phi_0)}. \tag{11}$$

Putting Eqs. (5), (9), and (10) together, we can characterize six equilibrium regions, depending on w, ϕ_0, ϕ_1, and c_L. In order to better characterize these regions, we first describe the types of equilibria that can emerge in our framework.

Types of Equilibria: In terms of the worker's strategy, χ, our model admits three qualitative types of equilibria:

- In a full qualification equilibrium (FQE), all low-cost workers get qualified: $\chi = 1$.
- In a zero qualification equilibrium (ZQE), no workers get qualified: $\chi = 0$.
- In a mixed strategy equilibrium (MSE), some low-cost workers get qualified and some don't: $\chi \in (0, 1)$.[16]

With the three classes of equilibria in hand, the following proposition characterizes all equilibria. It also demonstrates that, when multiple equilibria exist, the worker and employer share the same preferences over these equilibria.

Proposition 1. *Table 1 characterizes all equilibria of the model. When multiple equilibria exist, they are strictly Pareto ranked as follows: the FQE dominates the MSE, which dominates the ZQE.*

[15] Note that term p_E^* defined in (11) is simply a rearrangement of Eq. (5).

[16] Note that the full qualification equilibrium and zero qualification equilibria, when they exist, are otherwise independent of the parameters of the model. The mixed strategy equilibrium, when it exists, on the other hand, is sensitive to the exact values of these parameters.

Table 1: Equilibria of the single-group case.

Parameters (c_L, w, ϕ_0, ϕ_1)	Equilibria
Equilibria when $p > p_E^*$	
$\phi_0 > \frac{c_L}{w} > \phi_1$	FQE with $\chi^* = 1, \eta^* = 1,$, MSE with $\chi^* = \chi_M(p), \eta^* = \eta_M,$ ZQE with $\chi^* = 0, \eta^* = 0$
$\phi_0 > \frac{c_L}{w}$ and $\phi_1 > \frac{c_L}{w}$	FQE with $\chi^* = 1, \eta^* = 1$
$\phi_1 > \frac{c_L}{w} > \phi_0$	MSE with $\chi^* = \chi_M(p), \eta^* = \eta_M,$
$\frac{c_L}{w} > \phi_0$ and $\frac{c_L}{w} > \phi_1$	ZQE with $\chi^* = 0, \eta^* = 0$
Equilibria when $p < p_E^*$	
$\frac{c_L}{w} > \phi_1$	ZQE with $\chi^* = 0, \eta^* = 0$
$\phi_1 > \frac{c_L}{w}$	FQE with $\chi^* = 1, \eta^* = 0$

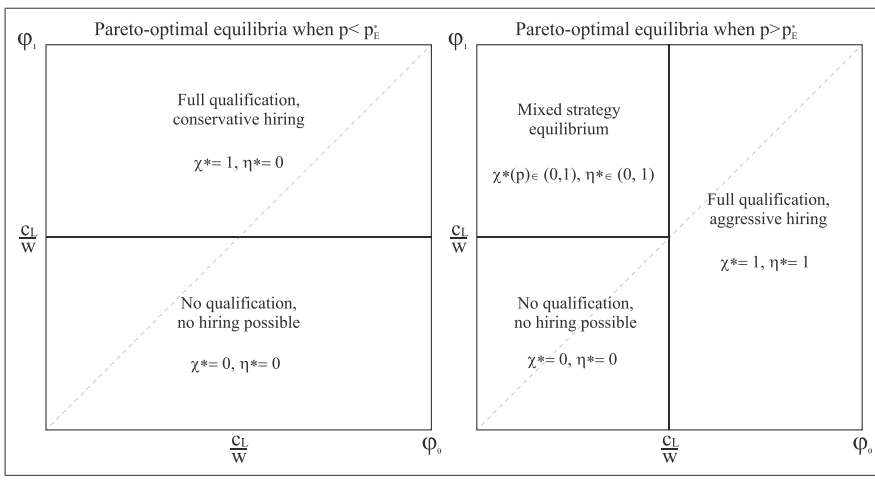

Figure 1: Pareto-optimal equilibria: single-group case.

Proof. Proofs of all numbered results other than corollaries are located in Online Appendix A. □

Table 1 is illustrated in Figure 1, which displays the (Pareto efficient) equilibrium regions with respect to the testing technology, $\phi = (\phi_0, \phi_1)$, and the group's potential, p, for a given, arbitrary pair of "low-cost level," c_L, and

wage, w. The principal point of the single-group analysis is to establish a baseline for examining the effects of labor market heterogeneity and differential information on qualification, employment, and welfare.

Note that, in both panes of Figure 1, the 45-degree dashed line represents the continuum of situations in which $\phi_0 = \phi_1$ and, as intuition would suggest, there are exactly two possible Pareto-optimal equilibria in these cases: when $\phi_0 = \phi_1$ is close enough to 0, then the unique equilibrium is a ZQE in which nobody gets qualified and nobody gets hired, because the moral hazard problem is "too severe" to sustain credible hiring in equilibrium. Otherwise the unique sequential equilibrium is an FQE in which all workers get qualified if and only if $c = c_L$. In such cases, the employer's hiring strategy is aggressive if potential (p) is sufficiently high and conservative otherwise.

With the equilibrium analysis of the single-group case in hand, we now extend the model to allow for two groups of workers, one of which has greater potential than the other.

Market Heterogeneity: Two Groups

In this section, we maintain the basic structure of the model analyzed in the section "The Model" while allowing workers to come from two different groups. Formally, the worker, W, now has a two dimensional type, $t = (g, c) \in T \equiv \{1, 2\} \times \{c_L, c_H\}$, with $0 < c_L < c_H$, where $g \in \{1, 2\}$ denotes the worker's group. The worker's type, $t \in T$, is assumed to be distributed as follows:

$$\begin{aligned} \Pr[g = 1] &= \gamma, \\ \Pr[c = c_L \mid g] &= p_g, \\ \Pr[c = c_H \mid g] &= 1 - p_g, \end{aligned}$$

with $\gamma \in (0, 1)$ representing the proportion of workers who are members of group 1, and $1 > p_1 \geq p_2 > 0$ representing the potentials of groups 1 and 2, respectively.

As in the single-group case analyzed in the section "The Model", c represents the cost to W of obtaining qualification. After observing his or her type, $t = (g, c)$, W chooses $q \in \{0, 1\}$, where $q = 1$ represents a decision to become qualified. E then observes both W's group, $g \in \{1, 2\}$, and his or her test result, $\theta \in \{1, 2, 3\}$, distributed conditional on q as described in the single-group case analysis (Eq. (1)).[17] After observing the worker's group membership and test results, (g, θ), E again chooses to hire W $(h = 1)$ or not $(h = 0)$, the game

[17] As in that section, we assume that both $\phi_q \in (0, 1)$ are common knowledge. Note that this implies that the testing technology is equally informative about q conditional on true qualification, q, for workers from both groups.

concludes, and the players' payoffs are as defined in (2) for the single-group case. As with the single-group case, the Employer's sequentially rational hiring decisions is always $h = 0$ conditional on $\theta = 1$ and $h = 1$ if $\theta = 3$, regardless of the worker's group, g. Accordingly, we extend the hiring strategy notation used in that section as follows: $\eta(g) \equiv \Pr[h = 1 \mid \theta_{=}2, g]$ for both groups $g \in \{1, 2\}$.

Equilibrium Analysis "With the Box"

The two-group model represents the situation facing the worker and employer when the employer can observe the worker's group and condition his or her hiring decision on it. In other words, E can directly observe g, and (importantly) W knows that E can observe g. Formally, the employers' set of information sets (which was Θ in the single-group case) in the two-group case *when the box is present* is:

$$\mathcal{I} = \{1, 2\} \times \Theta = \{1, 2\} \times \{1, 2, 3\}.$$

Equilibrium analysis when the box is present involves simply applying the analysis of the single-group case in the section "The Model" to each group separately. Accordingly, we omit a fuller recounting of this analysis and instead turn to consider the effects of "banning the box," or removing the employer's ability to directly condition his or her hiring decision on the worker's group membership.

Banning the Box

When the box is banned, E cannot condition his hiring decision on g. We represent this formally by modifying the game form analyzed above such that the set of information sets for the employer is

$$\mathcal{I} = \Theta = \{1, 2, 3\}.$$

For notational simplicity, we will denote the employer's hiring strategy when the box is banned by $\eta(\emptyset) \equiv \Pr[h = 1 \mid \theta = 2]$ (so as to distinguish it from the single-group case analyzed at the outset).[18] A key point of the analysis is that removing the box has ambiguous welfare effects. We are also able to identify some key determinants of the direction, and size, of this effect. We denote the

[18]Note that, in equilibrium when the box is banned, E will learn something about W's group from $\theta = 1$ or $\theta = 3$. However, because we have assumed that E does not have a taste for discrimination (E cares only about q, not g *per se*), this is irrelevant for our purposes in this article.

Table 2: Typologies of group potentials and testing structures.

Group potential	Parameter region
Uniformly high	$p_1 > p_2 \geq p_E^*$
Uniformly low	$p_E^* > p_1 > p_2$
Statistically distinct	$p_1 > p_E^* > p_2$

Testing structure	Parameter region
Uniformly informative	$\min[\phi_0, \phi_1] > \frac{c_L}{w}$
Positively informative	$\phi_1 \geq \frac{c_L}{w} > \phi_0$
Negatively informative	$\phi_0 \geq \frac{c_L}{w} > \phi_1$
Uninformative	$\max[\phi_0, \phi_1] < \frac{c_L}{w}$

unconditional probability of an individual having low costs of qualification by $\overline{p}$, which we refer to as the **population potential**:

$$\overline{p} \equiv \gamma p_1 + (1 - \gamma) p_2.$$

We will refer to the population potential as **high** when $\overline{p} \geq p_E^*$ and **low** otherwise.

To describe our results, we label regions of the parameter space. Table 2 describes the groups' potentials relative to the employer's hiring threshold, p_E^*.[19] Table 2 also describes four different types of testing structures.[20]

With this terminology in hand, we now discuss the impact of the box by working through four qualitative cases.[21]

Situations in which the Box Has No Effect

We begin by identifying settings in which the box has no effect on equilibrium behavior. In the interest of space, we relegate much of the discussion of this case to Online Appendix B.

Statistically Non-Distinct Group Potentials: A fundamental factor in determining whether the box can have an effect is the structure of the employer's potential beliefs in equilibrium, which revolves around the employer's hiring threshold, p_E^* (which is independent of W's group). This leads to the following corollary of Proposition 1.

Corollary 1. *Banning the box can affect Pareto efficient equilibrium behavior and/or welfare only if the groups have statistically distinct potentials*: $p_1 \geq p_E^* > p_2$.

[19] Note that the case of $p_1 = p_2$ is omitted from Table 2. This case is equivalent to the single-group case and BTB has no effect on equilibrium behavior.

[20] Note that the case of $\phi_0 = \phi_1$ is omitted from Table 2. This case is discussed above on page 526.

[21] Proposition 1 provides a road map for our analysis of the equilibrium effects of BTB.

Uninformative Testing Structures: Corollary 1 identifies only a necessary condition for BTB to have an effect on equilibrium behavior. It is not sufficient — Figure 1 also depicts situations in which the groups have statistically distinct potentials, but BTB still has no effect on equilibrium behavior. This occurs when the test is uninformative, as stated in the following corollary of Proposition 1.

Corollary 2. *When the test is uninformative* ($\max[\phi_0, \phi_1] < \frac{c_L}{w}$), *BTB has no effect on equilibrium behavior or welfare.*

We now turn to situations in which BTB has an impact on equilibrium outcomes, focusing first on those in which BTB affects *only* E's equilibrium hiring strategy, η.

Situations in which the Box Affects Only Employer Behavior

By Corollary 1 we know that, for BTB to have an impact, it must be the case that the groups have statistically distinct potentials ($p_1 \geq p_E^* > p_2$). Figure 1 illustrates that when $\phi_1 \geq \frac{c_L}{w}$ and $\phi_0 \geq \frac{c_L}{w}$, an FQE exists when the box is present. Consequently, regardless of whether E hires aggressively or conservatively, all low-cost workers are incentivized to obtain qualification. However, BTB may affect E's equilibrium hiring strategy. Because of this, the welfare effects of BTB are ambiguous, as summarized in the following corollary.

Corollary 3. *When the groups are statistically distinct* ($p_1 \geq p_E^* > p_2$) *and the test is uniformly informative* ($\min[\phi_0, \phi_1] \geq \frac{c_L}{w}$), *employer behavior is affected by the box but worker behavior is not, and BTB has ambiguous welfare effects*:

Population potential	*Employer hiring strategy*	*Welfare effects of BTB*		
		Group 1	*Group 2*	*Employer*
Low ($\overline{p} < p_E^*$)	*Conservative*	*Harmed*	*No Effect*	*Harmed*
High ($\overline{p} \geq p_E^*$)	*Aggressive*	*No Effect*	*Helped*	*Harmed*

In addition to illustrating that the workers and employer might have opposed preferences about the presence of the box, Corollary 3 illustrates the central role of statistical discrimination in our theory by highlighting the importance of the *population potential* for the workers' induced preferences about the box's presence. When the population at large has high potential, then BTB helps workers in the disadvantaged group and, conversely, when the population has low potential, BTB hurts workers in the advantaged group.

Situations in which the Box Affects Workers' Incentives

The final set of cases represent the only situations in which the presence of the box affects worker incentives to obtain qualification. In these cases, E may strictly prefer to ban the box if doing so can stimulate a greater number of workers to become qualified in equilibrium. In these situations where E prefers to ban the box it may also be the case that W prefers to ban the box too, and BTB can represent a Pareto improvement. It can also be the case that BTB stimulates worker qualification and benefits E while hurting W. And finally, it can be the case that BTB can reduce worker incentives to become qualified, leading to losses by both E and W. This final case represents a situation in which observing group labels in the hiring decision is Pareto superior to BTB. We begin with the case of positively informative test structures.

BTB with a Positively Informative Test: $\boldsymbol{\phi_1 \geq \frac{c_L}{w} > \phi_0}$

When the test structure is positively informative, the test result is more precise for workers who are qualified ($q = 1$) than for workers who are unqualified ($q = 0$). In the Pareto efficient equilibrium in this case, workers in the disadvantaged group ($g = 2$) obtain full qualification ($\chi^* = 1$) and E hires conservatively from this group ($\eta^*(2) = 0$). On the other hand, the employer E and workers in the advantaged group ($g = 1$) are playing mixed strategies in the Pareto efficient equilibrium, as characterized by Eqs. (9) and (10) (substituting the term p_1 for p in those equations). E would like to hire aggressively from group 1, but doing so would eliminate W's incentive to obtain qualification because the low ϕ_0 means that it is likely an unqualified person will send a signal of 2.

In such cases, BTB has two potential effects, depending on whether the potential of the population at large, $\overline{p}$, is high or low. The next proposition establishes that the worker's- and employer's-induced preferences regarding BTB are opposed when the test is positively informative and population potential is low. In this case, there is not a pure strategy equilibrium for workers in group 1 when the box is present. Banning the box enables E to credibly commit to hiring group 1 conservatively, thus stimulating full qualification by group 1 workers.

Proposition 2. *If the test is positively informative* ($\phi_1 \geq \frac{c_L}{w} > \phi_0$), *the groups are statistically distinct* ($p_1 \geq p_E^* > p_2$), *and population potential is low* ($\overline{p} < p_E^*$), *then* W *and* E *have opposed preferences over the box:* E *prefers that the box be banned, group* 1 *workers prefer the box to be present, and group* 2 *workers are indifferent.*

On the other hand, in contrast with Proposition 2, *the worker's- and employer's-induced preferences regarding BTB are aligned in favor of BTB*

when the test is positively informative and population potential is high. Using the phrase *BTB Pareto dominates the Box* to describe any situation in which there is a Pareto efficient equilibrium without the box that offers both players strictly higher (expected) payoffs than any equilibrium when the box is present, this is stated formally in the following proposition.[22]

Proposition 3. *If the test is positively informative* ($\phi_1 \geq \frac{c_L}{w} > \phi_0$), *the groups are statistically distinct* ($p_1 \geq p_E^* > p_2$), *and population potential is high* ($\overline{p} > p_E^*$), *then "BTB Pareto dominates the Box," strictly benefiting E and group* 2 *workers, and leaving the payoffs of group* 1 *workers unchanged.*

Proposition 3 is one of the key results of our analysis, but we defer discussion of it until the section "Discussion: Welfare, Testing, and Empirical Implications". Prior to that, we complete our analysis by considering the case of a negatively informative test.

BTB with a Negatively Informative Test: $\phi_0 > \frac{c_L}{w} > \phi_1$

For negatively informative test structures, the test result is more precise for unqualified workers than it is for qualified workers. In the Pareto efficient equilibrium with the box, workers in the disadvantaged group ($g = 2$) obtain no qualification ($q = 0$). E would hire conservatively from this group ($\eta^*(2) = 0$), and consequently the return to investment on qualification is too low to make qualification profitable. No one in the disadvantaged group becomes qualified, and no one is hired. Workers in the advantaged group ($g = 1$) obtain full qualification, and E hires from this group aggressively ($\eta^*(1) = 1$). With the box, the payoff for all workers in group 1 is strictly positive:

$$w - c_L > 0 \quad \text{if } c = c_L \text{ and}$$
$$w(1 - \phi_0) > 0 \quad \text{if } c = c_H.$$

The expected payoff for E in this case is

$$\gamma(p_1(B - w) - w(1 - p_1)(1 - \phi_0)) > 0.$$

This payoff is strictly positive in this case because $p_1 \geq p_E^*$ (otherwise BTB has no effect on equilibrium behavior): E receives a strictly positive payoff from hiring individuals from group 1 receiving a test score of $\theta = 3$ and a non-negative payoff for hiring individuals receiving a $\theta = 2$.

Our first result in this case is that BTB hurts both workers and the employer when population potential is low.

[22] By alluding to *expected* payoffs, we are referring to the worker's expected payoff prior to learning which group he or she is a member of (and, of course, prior to knowing the test result, θ).

Proposition 4. *When the test is negatively informative* ($\phi_0 > \frac{c_L}{w} > \phi_1$), *the groups are statistically distinct* ($p_1 \geq p_E^* > p_2$), *and population potential is low* ($\overline{p} < p_E^*$), *BTB is Pareto inefficient.*

Proposition 4 is informative: BTB will reduce employment in equilibrium when the population has low potential. This can occur for one or more of three reasons: (1) the advantaged workers have moderately high potential ($p_1 \approx p_E^*$), (2) the disadvantaged workers have sufficiently low potential (p_2 is too close to zero), and/or (3) the advantaged group is not particularly large (γ is too low). While of course group potentials might vary across different types of jobs, we believe that the third category is the most interesting. This is because γ reflects the proportion of *applicants* for the position in question who come from the advantaged group. It is well-documented that gender, racial, and ethnic compositions of the workforce vary — sometimes widely — across different types of jobs. Unfortunately, with this in mind, Proposition 4 suggests that BTB may not be as effective at promoting increased employment in sectors that are already disproportionately applied for by citizens from relatively disadvantaged groups.[23]

Our second, complementary, result in this case is that BTB benefits workers when population potential is high.

Proposition 5. *When the test is negatively informative* ($\phi_0 > \frac{c_L}{w} > \phi_1$), *the groups are statistically distinct* ($p_1 \geq p_E^* > p_2$), *and population potential is high* ($\overline{p} \geq p_E^*$), *BTB strictly benefits group* 2 *workers and leaves the payoffs of group* 1 *workers unchanged.*

Finally, when $\overline{p} \geq p_E^*$ the effect of BTB is ambiguous for the employer, and depends on whether E receives a positive or negative expected payoff from hiring individuals from group 2 aggressively. We have assumed that $p_2 < p_E^*$, and so it is not sequentially rational for E to hire an individual from group 2 receiving $\theta = 2$. This leads to no qualification by group 2 when E can observe group identity. However, E may strictly benefit from committing to aggressively hire from this group, because doing so stimulates a full qualification equilibrium. BTB can serve as a commitment device for E to hire aggressively, when such commitment would not be possible if group identity were observed. The following proposition details when this commitment benefits E.

Proposition 6. *When the test is negatively informative* ($\phi_0 > \frac{c_L}{w} > \phi_1$), *the groups are statistically distinct* ($p_1 \geq p_E^* > p_2$), *and population potential is high* ($\overline{p} \geq p_E^*$), *BTB Pareto dominates the Box if*

$$p_2 \in \left[\frac{w(1-\phi_0)}{B - w\phi_0}, p_E^*\right),$$

[23] Of course, there are many reasons for demographic variation across different jobs, including variation in wages. Our point here is meant only to be suggestive regarding the empirical implications of our analysis.

and E is hurt by BTB if

$$p_2 < \frac{w(1-\phi_0)}{B - w\phi_0}.$$

Again, we defer discussion of this result (along with its sibling, Proposition 3) in the section "Discussion: Welfare, Testing, & Empirical Implications". Prior to that, we briefly summarize and illustrate the equilibrium effects of BTB.

The Equilibrium Effects of (and Induced Preferences for) BTB: Mirroring Figure 1, Figure 2 depicts regions on which BTB can affect outcomes. Figure 2 illustrates that the effect of BTB depends critically on whether the population potential, $\overline{p} = \gamma p_1 + (1-\gamma)p_2$, is high or low. This figure is our roadmap for the impact of BTB on welfare when groups are statistically distinct.

Perhaps unsurprisingly, when population potential is low, BTB (weakly) hurts the advantaged group, and when population potential is high BTB (weakly) helps the disadvantaged group. However, the employer's preferences for BTB are less obvious, and are dependent on the testing structure. When the test is positively informative, the employer always prefers to BTB, whereas when the test is negatively informative the employer only prefers to BTB when population potential is high. We now turn to a discussion of this result, and of what the testing structure represents.

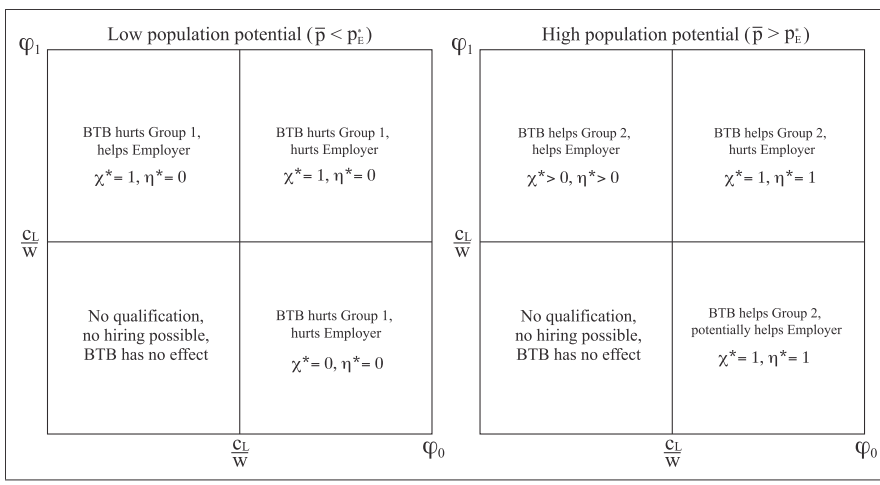

Figure 2: Equilibrium effects of BTB when groups are statistically distinct.

Discussion: Welfare, Testing, and Empirical Implications

The finding that BTB can be Pareto-dominant is arguably the most provocative of the conclusions we obtain from this framework. Along these lines, it is informative to contrast Propositions 3 and 6. These two results characterize the situations in which BTB can be Pareto-dominant, and are unified by their common reliance on population potential. When population potential ($\overline{p}$) is low, BTB can only hurt the workers by inducing E to hire everyone conservatively. Consequently, these results require that population potential be high. However, the two results differ in the exogenous nature of the testing technology and, in particular, whether the test is positively informative (Proposition 3) or negatively informative (Proposition 6).

Positively and negatively informative testing structures induce very different types of equilibrium behavior. When a test is positively informative it is, by definition, more likely that an unqualified individual will send a garbled signal (as $\frac{c_L}{w} > \phi_0$). This implies that, when the test is positively informative, *a full qualification equilibrium cannot be sustained in any equilibrium with aggressive hiring.* If E hires aggressively, then no individual can be induced to obtain qualification, as the likelihood that an unqualified worker will send a garbled signal is too high. In this case, if population potential is high, only MSE exist. Conversely, when the testing structure is negatively informative (requiring $\frac{c_L}{w} < \phi_0$) there always exists an FQE when population potential is high.

The Beneficial Role of BTB for the Employer: Although positively and negatively informative tests induce different types of equilibrium behavior, when the employer benefits from BTB in our framework, it is always because BTB is solving a **commitment problem** for the employer. This logic is most clear when considering Proposition 6 (where the test is negatively informative, and mixed strategies play no role). In this case, BTB commits E to hire group 2 aggressively, which induces full qualification by group 2, and strictly benefits both E and the group 2 workers. In Proposition 3, this logic is less obvious. In this case, BTB results in a shift from a FQE with conservative hiring for group 2 to an MSE for this group. Group 2 actually obtains *less* qualification after BTB, so how could this strictly benefit the employer?

The answer to the above question rests on the effect of BTB on group 1's qualification decision. Because the test is positively informative, both before and after BTB group 1 is in an MSE, with low-cost types obtaining qualification with some positive probability. However, this probability changes as a result of BTB. Since the total population potential $\overline{p}$ is less than group 1's potential, group 1 must obtain qualification at a higher rate in order to sustain the MSE. This improved rate of qualification for group 1 strictly benefits E (it leaves members of group 1 indifferent between BTB or maintaining the box).

Thus, by committing to disregard group labels, E can induce a higher rate of qualification from group 1.

This commitment logic also explains Proposition 2, in which BTB benefits the employer but not the worker. In this case the test is positively informative, but population potential is low. BTB enables E to credibly commit to hiring group 1 conservatively, thus inducing full qualification by that group.

Pareto Efficiency and Inequality: Even when BTB is Pareto-dominant in our framework, it nonetheless has uneven impacts on welfare. Comparisons between W and E are difficult for typical reasons (such a comparison depends on the exogenous parameters B and w), but a similar comparison between the two groups of workers is more straightforward. When BTB is Pareto-dominant, its adoption strictly benefits only *disadvantaged* workers. This mirrors much of the policy and scholarly discussion regarding BTB. At the same time, this one-sided nature of the welfare impact of BTB is essentially "built into" the model because of our assumption that the two groups of workers differ only in terms of potential (as opposed to, for example, the precision of the testing structure or the preferences of the employer), and because of our assumption that there is not competition between workers in the hiring process. We will return to these assumptions in our concluding remarks.

What the Testing Structure Represents: The distinction between positively and negatively informative tests raises the question of what these testing structures represent in substantive terms. In terms of robustness, it is important to note that our analysis does not rely on the assumption that a qualified worker ($q = 1$) can never receive a score of $\theta = 1$ or that an unqualified worker ($q = 0$) can never receive a score of $\theta = 3$: these probabilities can be positive, so long as they are not too large. With this in mind, one description of the distinction in employment situations is with respect to whether qualifications for the job in question are possible to directly demonstrate (either in one's record or during the hiring process).

- *Negatively Informative Tests:* For many entry-level positions, there are few objective indicators that an individual *is* specifically qualified for the position. On the other hand, there may be several indicators that an individual is *not* qualified for such a position.[24] For example, while some "general" credentials, such as a high school or college diploma, are relatively easy to verify, they may reflect more in their absence than in their presence. In our model, then, the absence of such a qualification represents $\theta = 1$, but a test result of $\theta = 3$ would be relatively rare,

[24]This asymmetry is due, in part, to the reality that an entry-level position typically does not require that one have held a similar job in the past. In the modern economy, many such jobs are in retail and customer service positions that do not depend heavily upon task-specific expertise.

requiring additional evidence (e.g., a credible and personalized recommendation from a teacher with personal knowledge of the applicant's abilities) that might, but need not, emerge from successfully completing the course of study leading to the diploma.

- *Positively Informative Tests:* As opposed to entry-level positions, more advanced positions often require task-specific experience and skills that can be more easily "directly" verified.[25] Similarly, for more advanced positions that require specific experience, it is reasonable to suppose that such performance might be gradated into more refined categories, ranging from "above the bar" ($\theta = 2$) to "clearly qualified" ($\theta = 3$). Our analysis requires only that it is rare for a person with the appropriate skills to be identified as "clearly unqualified."

Empirical Implications

The most general empirical implication of our theory is that the effects of BTB are sensitive to several factors (e.g., the proportion of advantaged workers and the nature of the testing technology) and, furthermore, these factors' effects are interdependent. For example, BTB can reduce employment among members of the advantaged group, especially when the population potential is low relative to the wage paid by the employer. The analysis above indicates that this is not always the case, of course, but a general implication of the model is that BTB can never increase the prospects of advantaged workers.[26] Of course, there are many possible goals one might seek to promote through BTB, including its effects on employment and qualification, per se, without necessarily needing to focus on a specific welfare criterion. Furthermore, welfare is not directly observable, so considerations of such observables might be the best one can do when evaluating the impact of BTB.

The predicted effects of BTB on qualification and employment for each of the parameter regions are displayed in Table 3.[27] The effects of BTB on employment and investment in qualification by each type of worker reflect the increase/decrease after BTB is imposed. The fact that this effect is sometimes

[25]For example, it is arguably easier to reliably infer that an applicant has knowledge of a specific programming language than that the applicant is generally unflappable in a wide array of customer service settings.

[26]This is partly because there are only two groups in our model. In a richer model with more than two groups that all have distinct group-specific potentials, this statement holds for the group with the *highest* potential — it is possible that *every* other group might benefit from BTB.

[27]Note that, when the test is negatively informative and population potential is high, there are three equilibria. As with our theoretical analysis, these comparative statics focus on the Pareto efficient FQE. There is a single row for the uninformative testing technology cases because neither statistical distinction nor population potential have any impact on the predictions in that case.

Table 3: Equilibrium effects of BTB on employment and qualification.

Testing Informativeness	Statistically Distinct?	Population Potential	Qualification		Employment (Pr[h=1])
			Advantaged	Disadvantaged	
Positive	**Yes**	**High**	↑	↓	**Ambiguous**
Positive	Yes	Low	↑	No Change	**Ambiguous**
Positive	No	High	↑	↓	↓
Positive	No	Low	No Change	No Change	No Change
Negative	**Yes**	**High**	**No Change**	↑	↑
Negative	Yes	Low	↓	No Change	↓
Negative	No	High	No Change	No Change	No Change
Negative	No	Low	No Change	No Change	No Change
Uniform	Yes	High	No Change	No Change	↑
Uniform	Yes	Low	No Change	No Change	↓
Uniform	No	High	No Change	No Change	No Change
Uniform	No	Low	No Change	No Change	No Change
Uninformative	———	———	No Change	No Change	No Change

Bold terms: Regions where BTB can be Pareto efficient.

negative (↓) and sometimes positive (↑) for each of the three outcomes provides some understanding of why empirical analyses of the effects of BTB to date have been mixed.

Two interesting conclusions from Table 3 emerge for the effect of BTB on employment when the groups are statistically distinct and the test is positively informative.[28] In each of these two regions, BTB might increase or decrease employment. Space precludes a more in-depth treatment of the determinants of the sign of this effect, but it represents a promising avenue for future research. Similarly, the table reminds us again that the effects of BTB on qualification will not only be conditional on the group identity of the worker, but also that in the positively informative testing case, these effects may very well move in opposite directions. Finally, each of these conclusions suggest that one might reach various normative/welfare conclusions when the testing structure is positively informative: for example, qualification and/or hiring might have spillover effects outside of the hiring situation considered here.

[28]Note that Proposition 3 states that BTB is always Pareto efficient in this case when the population potential is high, so this Pareto efficiency is "caused" by BTB increasing employment. Similarly, Proposition 6 indicates that BTB *can*, but need not, be Pareto efficient in the negatively informative case, in spite of the fact that BTB always increases employment in this parameter region.

Extensions and Conclusion

Of course, our model has many avenues for extension. Before concluding, we briefly describe a few of these below.

Discrimination and Big Data. Our framework could be extended to incorporate noisy signals about group membership, allowing one to consider the employer's incentive to circumvent BTB with ancillary data about applicants. Particularly in the new age of algorithms and "big data," the data solicited for decision-making can have subtle and powerful impacts on outcomes (Kleinberg *et al.*, 2018; Patty and Penn, 2015). With massive, and often proprietary, data sets and algorithms, it is much more difficult to predict what information might ultimately serve as the basis for discrimination (Barocas and Selbst, 2016).[29]

Endogenous Wages. In this article, the analysis assumes that the employer must offer an exogenously determined wage when he or she hires a worker. Different settings to consider include allowing E to set the wage; allowing E to offer the groups different wages; and allowing the workers to set their own wages. Preliminary analysis of our model when the employer can offer the groups different wages suggests that the employer will not prefer BTB to the box, because the employer can eliminate his or her own credible commitment problem through the choice of w. Interestingly, for some parameter regions the employer will offer a higher wage to workers from the disadvantaged group, but hire those workers more conservatively.

Group-Specific Testing Accuracy: Our analysis above assumes that the only distinction between the two groups is their potential, a notion grounded in inherent opportunities available to the individuals in the two groups. A complementary analysis would consider the implications of the testing technology (i.e., the distribution of θ conditional on qualification, q) depending on the worker's group. Such an analysis would be interesting for several reasons, including raising the possibility that banning the test itself might be socially optimal. We provide a simple example demonstrating that our qualitative result that BTB can be Pareto improving carries through to a setting in which potentials are the same across groups, but the test is noisier for the disadvantaged group (Example 2 in Online Appendix C). The example illustrates that there are situations in which BTB is Pareto optimal because the test is sufficiently precise to support positive employment with one of the two groups, but not the other, but an FQE exists when the employer is blind to the employee's group membership. This is largely due to the fact that the accuracy of the test result plays a similar role in the employer's sequential rationality calculation to that played by the employer's prior beliefs.

[29]Beyond the scope of this article, but related, is the emerging topic of how algorithmic systems may produce *disparate mistreatment*: situations in which the algorithm's decisions are more accurate for one group than for another (Zafar *et al.*, 2017).

Taste-Based Discrimination: In terms of our chosen application, we have assumed that felons and nonfelons face different opportunities with respect to becoming qualified for the job in question. While the employer cares only about an applicant's qualification for the job, different group potentials induce the employer to statistically discriminate between the two groups. As above in the case of testing accuracy, an alternative assumption would be that the groups face the same opportunities for qualification, but the employer receives different rewards from hiring a qualified felon versus a qualified nonfelon (in the language of our model, we would set $B_1 \neq B_2$). In this different formulation, the employer discriminates between the two groups due to taste. We provide a simple example illustrating how our qualitative results carry through to this setting of taste-based discrimination (Example 1 in Online Appendix C). As long as the employer would prefer to hire a qualified applicant from either group (though perhaps having a preference for one group over the other), BTB can be Pareto optimal precisely to solve the employer's credible commitment problem, as in our main analysis above.[30]

Intersectionality: Our analysis focuses on the case in which there are two groups of workers. In reality, there are many relevant forms of group membership (e.g., race, ethnicity, citizenship, age, gender, and veteran status). Considering such an extension would enable us to explore the implications and challenges of issues of **intersectionality** — "the way in which various forms of inequality often operate together and exacerbate each other"[31] — when considering the impact of supplemental information on allocating scarce resources.

Voluntary Disclosure: Our analysis is centered on the effects of information about an individual's traits. In reality, this information is often solicited by the employer, as opposed to being directly observed. Accordingly, an important extension of the model would be to include voluntary provision/revelation of this information within the model itself.

Competitive Hiring: A final direction to extend the model is to incorporate the possibility that the employer will have a larger set of applicants to choose from than the number of positions he or she needs to hire. Such an extension will presumably exacerbate conflicts between advantaged and disadvantaged workers with respect to whether banning the box is beneficial.

Concluding Thoughts: Policies intended to eliminate discrimination often prohibit disparate treatment on the basis of group membership. The withholding of information about group membership is one commonly proposed solution to

[30] We thank Andrea Moro for pointing this out to us.

[31] Kimberlé Crenshaw, quoted in "She Coined the Term 'Intersectionality' Over 30 Years Ago. Here's What It Means to Her Today," by Katy Steinmetz, *TIME*, February 20, 2020. For a very recent formal contribution along these lines, see Stewart (2022).

the problem of disparate treatment.[32] We have presented a theory of how omitting such information — "banning the box" that provides this information — might affect both hiring and investment in qualification in the job market.

Policies such as BTB are often motivated by a desire to level the playing field for individuals from different backgrounds. Our analysis indicates some of the promises — and pitfalls — of such policies. In line with empirical evidence, the theory highlights the generally positive impact such policies will have on disadvantaged workers and the weakly negative impact they might have on advantaged workers. However, the theory also indicates, unsurprisingly, that such policies can sometimes harm employers, while at the same time offering (to us, at least) an unexpected conclusion: sometimes employers can strictly benefit from these policies if they are foreseen and reacted to by workers. Furthermore, the theory isolates one classic game theoretic reason for this potential salutary impact: the employers in some cases benefit from the "blindness" imposed on the employer by such policies because the concomitant lack of ability for the employer to discriminate between workers from the two groups can provide instrumental incentives to workers from such groups to make costly investments in qualification in the hopes of obtaining employment.

The highly stylized nature of our model arguably magnifies the impact of its conclusions: the ambiguity of the analysis even in such a constricted environment suggests that the ambiguities are in some ways fundamental to even more replete models of the labor market. In some cases, group information is crucial to sustaining the job market at all: when the conditional probability of an individual having the opportunity to obtain qualification is sufficiently sensitive to group membership, banning the box can induce an employer to raise the threshold for hiring anyone sufficiently high that no prospective worker will find it worthwhile to invest in qualification, leading to market collapse. On the other hand, there are also situations in which banning the box benefits both workers and the employer. In such cases, the employer benefits from a credible commitment to not treat workers differently based on the group membership. Interestingly, the model demonstrates that the employer can benefit from BTB through different causal mechanisms. In some cases, BTB benefits the employer because it incentivizes more effort from disadvantaged workers (Proposition 6), while in other situations it benefits E because it incentivizes more effort by *advantaged* workers (Proposition 3).

From a normative standpoint, the theory also suggests some key factors in determining whether banning the box may lead to a Pareto improvement in

[32]Clearly, for any given employment decision, certain group memberships are appropriate considerations for the employer (e.g., does the applicant have a high school diploma?). Accordingly, in practice, discrimination is legally barred only with respect to certain group memberships. For example, in the United States, federal antidiscrimination laws generally protect against discrimination on the basis of race, color, national origin, religion, sex, age, or disability.

any given setting. In addition, these factors can be interpreted in ways that are both strategic and substantive. First, the employer must have a reason to treat workers differently based on their group membership (the groups must be *statistically distinct* from one another), and the employers' information about a prospective worker's qualification must be *asymmetrically (positively or negatively) informative*. The first of these two necessary conditions revolves around the employer's sequential rationality constraints, and the second centers on the workers' incentive compatibility conditions. We suspect that the first of these two is more robust than the second. This is because the second condition follows from our assumptions about the workers' (1) uniform distaste for obtaining qualification and (2) uniform taste for being hired, while the first condition relies only on the presumption that the workers care about the employer's decision. In any event, the fact that these these two conditions bind on different sides of the market suggests to us that some of the conclusions from this model can be identified in (possibly quite) disparate social, political, and economic interactions.

References

Arrow, K. J. 1973. "The Theory of Discrimination". *Discrimination in Labor Markets*. Princeton, NJ: Princeton University Press, 3–33.

Autor, D. H. and D. Scarborough. 2008. "Does Job Testing Harm Minority Workers? Evidence from Retail Establishments". *The Quarterly Journal of Economics*. 123(1): 219–77.

Barocas, S. and A. D. Selbst. 2016. "Big Data's Disparate Impact". *California Law Review*. 104: 671–732.

Bartik, A. and S. Nelson. 2019. "Deleting a Signal: Evidence from Pre-Employment Credit Checks". *University of Chicago, Becker Friedman Institute for Economics Working Paper*. 2019(137).

Becker, G. S. 1971. *The Economics of Discrimination*. 2nd. Chicago, IL: University of Chicago Press.

Bertrand, M., D. Chugh, and S. Mullainathan. 2005. "Implicit Discrimination". *American Economic Review*. 95(2): 94–8.

Bjerk, D. 2008. "Glass Ceilings or Sticky Floors? Statistical Discrimination in a Dynamic Model of Hiring and Promotion". *The Economic Journal*. 118(530): 961–82.

Coate, S. and G. C. Loury. 1993. "Will Affirmative-action Policies Eliminate Negative Stereotypes?" *The American Economic Review*: 1220–40.

Doleac, J. L. and B. Hansen. 2020. "The Unintended Consequences of "Ban the Box": Statistical Discrimination and Employment Outcomes When Criminal Histories Are Hidden". *Journal of Labor Economics*. 38(2): 321–74.

Fang, H. and A. Moro. 2011. "Theories of Statistical Discrimination and Affirmative Action: A Survey". In: *Handbook of Social Economics*. Ed. J. Benhabib, A. Bisin, and M. O. Jackson. Vol. 1. Elsevier. Chap. 5, 133–200. http://andreamoro.net/assets/papers/survey-statdisc.pdf.

Fryer Jr, R. G. 2007. "Belief Flipping in a Dynamic Model of Statistical Discrimination". *Journal of Public Economics*. 91(5-6): 1151–66.

Kim, Y.-C. and G. C. Loury. 2019. "To Be, or Not to Be: Stereotypes, Identity Choice and Group Inequality". *Journal of Public Economics*. 174: 36–52.

Kleinberg, J., J. Ludwig, S. Mullainathan, and C. R. Sunstein. 2018. "Discrimination in the Age of Algorithms". *Journal of Legal Analysis*. 10.

Kreps, D. M. and R. Wilson. 1982. "Sequential Equilibria". *Econometrica*. 50(4): 863–94.

Lundberg, S. and R. Startz. 1998. "On the Persistence of Racial Inequality". *Journal of Labor Economics*. 16(2): 292–323.

Maturana, G., J. Nickerson, and S. Truffa. 2020. "Labor Market Effects of Deleting Delinquencies". *Available at SSRN*.

Moro, A. and P. Norman. 2004. "A General Equilibrium Model of Statistical Discrimination". *Journal of Economic Theory*. 114(1): 1–30.

Niu, M., S. Kannan, A. Roth, and R. Vohra. 2022. "Best vs. All: Equity and Accuracy of Standardized Test Score Reporting". In: *2022 ACM Conference on Fairness, Accountability, and Transparency*, 574–86.

Patty, J. W. and E. M. Penn. 2015. "Analyzing Big Data: Social Choice & Measurement". *PS: Political Science & Politics*. 48(1): 95–101.

Patty, J. W. and E. M. Penn. 2022. "Algorithmic Fairness and Statistical Discrimination". *arXiv preprint arXiv: 2208.08341*.

Phelps, E. S. 1972. "The Statistical Theory of Racism and Sexism". *American Economic Review*. 62(4): 659–61.

Raphael, S. 2020. "The Intended and Unintended Consequences of Ban the Box". *Annual Review of Criminology*. 4.

Stewart, R. T. 2022. "Identity and the Limits of Fair Assessment". *Journal of Theoretical Politics*. 34(3): 415–42.

Zafar, M. B., I. Valera, M. Gomez Rodriguez, and K. P. Gummadi. 2017. "Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification Without Disparate Mistreatment". In: *Proceedings of the 26th International Conference on World Wide Web*, 1171–80.