# Strategic Discrimination in Hierarchies*

Dominik Duell† and Dimitri Landa‡

Monday 30th September, 2019

**Abstract**

In a laboratory setting, we explore strategic discrimination in principal-agent relationships, which arises from mutually re-enforcing expectations of identity-contingent choices. Our experimental design isolates the influence of the strategic environment from effects of other sources of discrimination, including statistical differences between sub-populations and outright prejudice. We find that, in a strategic setting, principals who reward agents based on outcomes more readily attribute high performance to effort when they share the agent's group identity. No such bias exists either for principals whose reward decisions are outcome-independent or for principals in a non-strategic environment. Agents in the strategic setting tend to anticipate higher demands from out-group principals, and condition their effort choice on that expectation. Because they under-appreciate this conditionality, principals tend to underestimate the effort from out-group agents.

Keywords: Bureaucracy, strategic discrimination, principal-agent relationship, reciprocity
JEL: J7, J15, J24, D83, D84

*The research presented in this paper was supported by NSF Grant $SES-1124265. Support through the ANR - Labex IAST is also gratefully acknowledged. We are grateful to Alessandra Casella, Catherine Hafer, Sanford Gordon, Tom Clark, anonymous reviewers, and audiences at Columbia University, the Max Planck Institute in Bonn, University of Oxford, and NYU for their helpful comments.

†University of Essex

‡New York University

# 1 Introduction

Considerable evidence from increasingly sophisticated controlled studies set in Western liberal democracies, including the U.S., suggests the persistence of discrimination.[1] This evidence, frequently obtained in the context of underlying principal-agent relationships, is buttressed by the systematic average pay and promotion differentials across sexes and racial and ethnic groups.[2] Yet, despite such aggregate-level evidence, discrimination is notoriously difficult to prove at the individual-case level – owing, in part, to the sometimes subtle nature of discriminatory practices, and in part, to the de-facto institutional discouragement of individual redress. Our aim in this paper is to provide an experimental analysis of one critical source of discriminatory practices which helps account for this disjunction: the strategic calculus of mutual expectations at the core of many principal-agent relationships.

Consider the following example. Alice's team supervisor, Bob, will decide which members of the team to promote in order to encourage good performance. Bob does not directly observe how much effort they put in and will base his judgment on his interpretation of their individual performance outcomes – noisy measures of the effort levels underlying them. Alice, who is concerned about the possibility that Bob's decisions will favor team members who are like himself over those who are like her is pessimistic about her chances for promotion and therefore considers whether it might be wiser to re-allocate some of her time elsewhere, or to increase her effort in the hope of impressing Bob. Bob, who suspects that Alice is choosing to under-invest, is less likely to attribute a good outcome from her to her effort, and more likely to her good luck. In effect, then, the quality of outcome Alice needs to generate to obtain a promotion is higher than the quality of outcome needed for other similarly situated team members.

If, realizing this, Alice is discouraged and chooses to invest less, Bob's suspicions are confirmed; his interpretation of outcomes and Alice's expectation of a tougher standard would be both correct and consistent with each other and with the actions supporting them. Yet, the state of affairs

[1]These studies include audit studies (Bendick, 2007; Bertrand and Mullainathan, 2004; Goldin and Rouse, 2000), "hit-rate" analyses (Knowles et al., 2001; Persico, 2002, 2009), and experiments implementing behavioral games (Falk and Zehnder, 2007).

[2]See, e.g., Wright et al. (1995); Altonji and Blank (1999); Western and Pettit (2005).

would be clearly discriminatory, and arguably no less insidious in this case, where it is a result of parties' higher-order beliefs (i.e., beliefs about beliefs), than when discrimination is a response to statistically or psychologically sustained asymmetric group-based generalizations. Of course, Alice may, instead, choose to invest more, not less, effort, in which case Bob's judgment would be both discriminatory *and* incorrect.

*Strategic discrimination* as exemplified in Bob's response to Alice in the above example need not be identified with specific discriminatory institutional features, but it may be reinforced by institutional details that, though not discriminatory in and of themselves, enable discriminatory practices by severely limiting the prospect of legal remediation. An example of such institutional details is the legal and administrative framework for addressing allegations of workplace discrimination in the U.S. In almost all cases, employees alleging workplace discrimination must pursue legal or administrative remedies on their own.[3] Compensatory damages are capped at a maximum of $300,000 for large companies, and considerably lower for smaller ones. The burden of proof for awarding punitive damages is such that those awards are rare even conditional on employee's prevailing at trial (Captain, 2017). The low odds of an employee winning a discrimination-based legal dispute and the low upside of such a victory overwhelm the legal fees and other investments into a protracted administrative and/or legal process demanded of the complainants.

The situation that our example of Alice and Bob captures is generic, and the identity dimension in question could have nothing whatsoever to do with the team tasks at hand. It could be gender, or race, or ethnicity. But Alice and Bob could also be bureaucrats belonging to different political parties, the supervisor suspecting his underling of "deep state" preferences that would lead to an under-investment of effort into an ideologically charged task. The latter possibility underscores two immediate senses in which the analysis of discrimination in question is of political significance: it applies to discrimination within a government bureaucracy and can help shed light on politically driven purges of bureaucracy that often dominate the news cycle (Lewis, 2011; Gordon, 2009).[4]

[3]The U.S. Equal Employment Opportunity Commission (EEOC), which is charged with certifying discrimination complaints under the existing non-discrimination statutes, brings the cases on behalf of the employees very rarely – filing, for example, only 86 lawsuits alleging discrimination in fiscal year 2016 (Captain, 2017).

[4]While in the U.S., civil servants are protected against discrimination on the grounds of political

Ironically, protections against workplace discrimination in U.S. federal and state governments – by far the largest employers in the country, and encompassing countless principal-agent relationships that hold the potential for discriminatory behavior – are sometimes weaker than in the private sector. Not only are complainants against discrimination within the federal bureaucracy barred from seeking punitive damages, but both federal and state bureaucracies are exempt from some of the anti-discrimination protections that bind on private employers.[5] Recent empirical work provides evidence of discriminatory behavior by federal government employees (Giulietti et al., 2017), hinting at the culture of discrimination.[6] The bottom line is the familiar discriminatory pattern and the society lacking a critical element of effective social order – equal treatment of citizens and legitimacy

affiliation, those protections are not constitutional, and the expansive interpretations of the Article II of the U.S. Constitution, which have become increasingly influential, especially within the current administration, clearly undermine their force (Huq and Ginsburg, 2018). Even within the context of the existing protections, the leaders' leeway for selection for special tasks and re-assignment to other duties is considerable, and the comparatively large share of "political appointments" creates an altogether unprotected class.

[5]Both federal and state governments are exempt from the Americans with Disabilities Act (ADA), and the state bureaucracies are, further, exempt from the Age Discrimination in Employment Act (ADEA) and from the state employment-discrimination laws based on sexual orientation. With the Trump administration's reversal of the Obama-era policy that the Civil Rights Act implied protections against sexual orientation-based discrimination in the workplace, the federal bureaucracy is no longer subject to those constraints as well.

[6]Although we are not aware of the systematic studies of discrimination across federal bureaucracy, the gaps in salary and promotion for government employees across demographic groups are suggestive. Per U.S. Federal Government Office of Personnel Management FedScope Federal Human Resource Data (accessed January 8, 2018), in 2017, the average salary of female members of the federal bureaucracy was about $4800 lower than of the male ones. African Americans and Hispanics earned $6200 and $5300 less than whites, respectively. When separating by occupation group, women earned less than men in 85% of different job categories. The corresponding numbers for African Americans and Hispanics are, respectively, 86% and 76%. Despite accounting for 43% of all federal employees, women hold only 33% of the supervisory and leadership positions.

of social and political institutions.

The first line of attack on discrimination is often through the legal system, seeking to enforce equal treatment laws and non-discrimination statues. But when it is traceable to determinants that fall outside the letter of the statute or when the legal enforcement process lacks adequate tools or will to address it, discrimination is, once again, a fundamentally political problem. For reasons indicated above, these conditions describe strategic discrimination. Because it is, most proximately, a function of individuals' higher-order beliefs, rather than of specific, clearly discriminatory, institutions, it can be self-reinforcing – yielding observable behavioral patterns that, in effect, belie the true extent of discrimination and often "slip" through the statutory net. And to prevent such outcomes, one needs mechanisms for shifting the social norms that clearly go beyond the inadequate enforcement framework described above.

The search for understanding the persistence of discriminatory patterns, the social and political inequality they entail and reinforce, as well as for institutional solutions to these problems must start with an improved understanding of the mechanisms underlying discrimination and their behavioral attributes. The primary aim of this paper is to contribute to the latter by experimentally isolating the distinctly strategic effects of individuals' responses to sharing a group identity in a principal-agent environment.

In psychology, the phenomenon of prejudicial judgment is grounded in a psychological disposition to a bias known as *the ultimate attribution error* (Pettigrew, 1979; Hewstone, 1990). The bias concerns differences in how observers account for identical levels of performance from individuals who do or do not share the observer's relevant social identity. Thus, for example, when observing good outcomes from individuals with shared group identity (e.g., a male team leader's male underling in gender-salient environments), the team leader/principal will be more inclined to attribute those outcomes to factors (e.g., effort) that are controllable by the underling/agent than when that principal sees those outcomes coming from *out-group* actors (e.g., a female team member in the same environment). By the same token, the principals will be marginally more likely to associate good outcomes from the out-group agent with factors, such as e.g., favorable circumstances, that are not in the agent's control. The relationships described here, however, are fundamentally strategic in that outcomes will depend not only on the actions by those supervised but also on their expectations of the feedback from their supervisors. When we observe asymmetric attribution in these settings, it may be pure prejudice, but it may also reflect correct, while clearly regrettable, beliefs about differences in performance arising from strategic responses to the asymmetric beliefs and choices of others. While the economic theory of principal-agent relationships and statistical discrimination has understood that it does not take a psychologically driven misattribution to create and sustain stereotypes (Phelps, 1972; Arrow, 1973; Spence, 1973; Coate and Loury, 1993), there has been a gap between the recognition of the different contributors to discrimination and their empirical evaluation in either political or economic contexts. Similarly, there has been little or no uptake of this issue in the political economy literature, which has, otherwise, yielded considerable work on agent choice and oversight in principal-agent relationships in hierarchies (Ting, 2002; Miller, 2005; Besley, 2006; Ting, 2011; Gailmard and Patty, 2012; Bueno de Mesquita and Landa, 2015).

Our analysis yields a number of novel results that help close this gap, some of which we highlight here. First, our findings suggest that the patterns of beliefs associated with the ultimate attribution error may emerge as a fundamentally strategic phenomenon (though, as we will see shortly, not necessarily fully consistent with equilibrium play). In strategic environments, principals who reward their agents contingent on the outcomes tend to attribute good outcomes, on average, more readily to their agents' effort when they share a group identity and reward those agents more frequently. When principals' and agents' choices are not strategically co-dependent, the attribution asymmetries disappear along with the possibility of (asymmetric) rewards.

Second, the agents' choices suggest the presence of an important subtlety, which we identify both theoretically and in our experimental data, and which does not neatly match up with the principals' revealed expectations. We show, in particular, that agents' choices are subject to two effects that sometimes push in opposite directions. The first, *the expected bias effect* manifests in the agents' effort choices increasing in the expectation of the principals' in-group bias in rewards. The second, *the expected demand effect*, is the agents' effort choice increasing with their expectations of the demands from the principals. While the expected bias effect reinforces the principals' asymmetric attribution, the expected demand effect runs counter to it precisely because the principals' higher demands tend to occur in out-group matches. This helps explain another of our findings: that principals tend to do better at anticipating the choices of in-group than of out-group agents – they underestimate the possibility that agents in out-group matches increase their effort in response to their expectations of higher demands from the principals.

## 2 Discrimination: Variety and Identification

We analyze discrimination and prejudice in the relationship of delegation found naturally in the contexts of hierarchical relationships in bureaucracies. Discrimination, in a textbook account, refers to a practice of treating persons who perform *equally* in a physical or material sense *unequally* in a way that is related to an observable characteristic such as race, ethnicity, or gender.[7] Thus defined, discrimination is useful for operationalizing an anti-discriminatory policy, but if we take seriously the effect of the expectation of discriminatory treatment on the agents' choices, then observed unequal treatment may not be the full story of discrimination. This idea is at the core of the present study. A discriminatory impulse is distinguishable from prejudice – a faulty or inflexible generalization about members of a group (Allport, 1954), which is often a key psychological determinant of discrimination. Unlike discrimination, which may be rationalizable with a set of potentially correct beliefs, prejudice necessarily entails a mistake.

An influential theoretical approach to analyzing the determinants of discrimination views it as resulting from a *taste for discriminating* against out-group members (Becker, 1971; Akerlof and Kranton, 2000). The mechanism underlying this kind of discrimination is, in the first place, psychological: the differential treatment it envisions is not a product of a rational response, but rather of prejudice.[8]

In contrast to the taste for discrimination, *statistical discrimination* does not presuppose a prejudice; it is grounded in a rational inference about the likely features of group members given the relevant statistics of the demographic populations (Phelps, 1972). But those statistics, of course, reflect groups' distinct histories, experiences, etc. – factors which could arise endogenously from others' treatment of them – and, as Arrow (1973) points out, they can be, in that sense, self-confirming. When the relevant choices are sufficiently close to each other in time, asymmetric beliefs about members of different groups can, through mutual strategic feedback, become self-confirming

[7]See Holzer and Neumark (2000) for a more detailed elaboration.

[8]A somewhat different version of this mechanism, the *ultimate attribution error* (Pettigrew, 1979) manifests when individuals are biased – tracking shared vs. unshared salient social identity – in their attribution of outcomes to the contributing factors controlled by the outcome-generating agent rather than factors not controlled by her (Hewstone, 1990).

even when those groups are identical ex-ante.[9] This latter idea, that discriminatory behavior may rest on higher-order beliefs about strategic feedback, suggests the possibility of discrimination that is a specifically *strategic phenomenon*. (To be sure, even when that is the case, it may or may not be an *equilibrium* phenomenon – depending on whether the higher-order beliefs and actions are jointly consistent.)

Strategic discrimination, which is sometimes described as the Arrovian version of statistical discrimination, has informed a number of important debates about both public policy and politics. As some scholars have argued, policy interventions such as affirmative action programs can induce differences in principals' beliefs about how much effort members of different social groups exert, prompting the principal to discriminate; the resulting discrimination reduces incentives for members of the disadvantaged group to invest, creating a self-fulfilling prophecy (Loury, 1976; Coate and Loury, 1993).[10] With respect to supply-side behavior that is consistent with the strategic expectations at the core of the Arrovian approach, Niederle and Vesterlund (2007) and Kanthak and Woon (2015) find that women are less likely to select into competitive environments and pre-labor market discrimination has been shown to affect career choices by minorities and women (Benabou, 1996; Neumark and McLennan, 1995). These studies go considerable distance in distinguishing the taste for discrimination and the statistical discrimination mechanisms. However, because differences in group statistics are typically part of the specific principal-agent interaction analyzed, a controlled laboratory environment holds particular promise for getting at the distinctly strategic determinants of the principals' responses.

Several previous laboratory studies have made steps in that direction. Fershtman and Gneezy (2001) provide evidence of differences in attribution in interactions with a strategic component (modeled as a trust game) and without it (modeled as a dictator game), but find that senders' stereotype-driven beliefs in the trust game are inconsistent with the return decisions, which do not

[9] While the controlled observational studies cited above document discrimination, they leave aside the question of what drives discrimination in principal-agent settings.

[10] In the context of electoral representation, the conclusion may be the opposite: voters may be better off with an out-group candidate because she will work to earn the electoral support that an in-group candidate will take for granted (Swain, 1993; Landa and Duell, 2015).

vary with the group identity conditions. The result could be explained by the fact that the receiver has no affirmative (motivated) reason in the experiment to act on the stereotypes, whether the senders' or her own, because the payoffs to the receiver's choice in the trust game are independent of whatever beliefs she may have about the sender. To capture the effect on subjects' beliefs and choices that models strategic discrimination, our experimental design implements strategic feedback both before and after the receivers' choices. The important experiments in Fryer et al. (2005) and Haan et al. (2015) simulate both principals' hiring decisions and agents' choices whether to invest into education, and report evidence of strategically reinforced discrimination in settings that contain both strategic feedbacks. In both studies, the publicity of the initial asymmetries is key, but makes it difficult to identify the extent to which the strategic actions they report remain anchored in the seeded population statistics, or, to put it differently, in the distinctly Phelpsian framework of statistical discrimination. The design of our study obviates this concern by avoiding the seeding of discrimination with either the asymmetric group-level parameters or the asymmetries in the history of play. Further, the Fryer et al. and Hahn et al. studies do not endow principals with distinct group identities, while our study assigned potentially differing group identities to agents and principals. This allows interpretations of outcomes to arise endogenously entirely in response to beliefs about the consequences of shared vs. unshared social identities – the mechanism at the core of Arrovian strategic statistical discrimination.

**Identifying Arrovian Statistical Discrimination in a Principal-Agent Environment** We highlight three features of our experimental design that allow for the identification of strategic discrimination: First, to separate the strategic effect from the psychological one, we create counterfactual environments. (1) We compare the beliefs of principals whose reward strategies are constant in outcome to those of *incentivizing* principals whose reward decisions vary with the observed outcomes; and (2) we compare the principals' beliefs in a treatment that implements a strategic to those in a corresponding non-strategic environment. The strategic environment has two-sided feedback, allowing the agents to condition their effort choices on their expectations of the principals' reward decisions and the principals to condition their reports of beliefs about agents' choices on their expectations of how agents likely evaluated their own expectations of being rewarded. This creates the possibility of strategically reinforced identity-contingent incentivized beliefs. In the non-strategic environment, whatever asymmetry in beliefs is observed must be due to psychological, taste-for-discrimination factors like the ultimate attribution error. Using that behavior as a baseline, we can interpret the behavioral differences between strategic and non-strategic environments as explainable by the specifically strategic aspects of the interaction.

Second, to further separate strategically driven belief asymmetries from the non-strategic (Phelpsian) statistical belief asymmetries, we adopt a design that does not pre-treat subjects with reputations of social groups. In particular, we induce artificial group identities in a treatment related to the "minimal group paradigm" (Tajfel and Turner, 1986) – an approach to inducing a (weak) notion of identity that is seemingly unrelated to the behavior of interest – and provide minimal feedback to subjects in the course of play. This approach advances our overall goal of isolating the beliefs-driven determinants of strategic discrimination from the influence of other elements of the social environment that may also affect willingness to discriminate, e.g., reputation costs for discrimination.

Third, to avoid the possibility that principals may rationally use their reward instruments to elicit different behaviors from different types of agents to effect a type separation in equilibrium, we tie the principal's payoffs to her beliefs about the realization of agent's underlying type vs. effort, but not to the principal's decision whether to reward the agent.

# 3 A simple model of principal-agent relationships

## 3.1 Set-up

We capture the underlying strategic principal-agent relationship in a simple model of incomplete contracting. The principal faces an agent with privately known type $t \in \{1, 2, 3\}$. The principal's commonly known prior is assumed to be uniform on that support. The agent chooses her effort level, $e \in \{1, 2, 3\}$, which is costly to herself, with $\alpha$ denoting the marginal cost. The outcome $F$ is given by $F = t + e + \omega$, where the noise, $\omega$, is a random draw from a uniform distribution on $\{-1, 0, 1\}$. Thus, $F \in \{1, 2, 3, 4, 5, 6, 7\}$.

The payoffs of both the principal and the agent depend on $F$, though in different ways. The principal observes $F$ (but not $t, e$, or $\omega$), and then makes two (simultaneous) choices. The first, the decision on whether to give a bonus, $b$, to the agent, has no direct effect on the principal's utility (but does affect it indirectly through the agent's effort level in expectation of the principal's bonus-awarding rule). The second choice has a direct effect on the principal's utility, and reveals her beliefs about the determinants of the agent's performance. That decision is the choice of whether to double the $t$ or $e$ component in her payoff, which the principal must make without directly observing $t$ or $e$ (i.e., just with her knowledge of $F$ and, as we will see below, the common knowledge between her and the agent of their respective group identities). The principal's payoff, then, is computed as $F + De + (1 - D)t$, where $D \in \{0, 1\}$ is the indicator variable such that $D = 1$ if the principal decides to double $e$ and $D = 0$ if she doubles $t$. $D$, thus, may be interpreted as the principal's belief whether the observed value of the outcome can be attributed more to the agent's effort or her type.

The agent's payoff is given by $G(F, b, e)$, where

$$G(F,b,e) = \begin{cases} \beta\sqrt{F+b} - \alpha e \text{ if the bonus is awarded} \\ \beta\sqrt{F} - \alpha e \text{ if the bonus is not awarded.} \end{cases}$$

$G(\cdot)$ is, thus, increasing in $F$ and $b$ and decreasing in $e$. The game ends when payoffs are realized.

### 3.2 Best Responses and Equilibria

There are many Perfect Bayesian Equilibria of this game, since any reward rule by the principal can be sustained in equilibrium. To focus our analysis, we restrict attention to two types of reward rules: constant in the outcomes that the principal observes, and monotonically increasing in those outcomes. We will refer to the equilibria corresponding to the second type of rule as the *outcome-contingent-play (OCP) equilibria*, and to the equilibria with the first type of rule as the *outcome-noncontingent-play (ONCP) equilibria.*

Intuitively, in ONCP equilibrium, the agents choose minimal levels of effort, inducing partial separation through outcomes, and the principal will always prefer to double type. In contrast, in OCP equilibria, principals' strategies may create incentives for forward-looking agents to invest into effort. We will call principals who are playing cutpoint strategies which call for rewarding outcomes that meet a given threshold and not rewarding otherwise as *incentivizing principals* and their strategies as *incentivizing strategies.*

The parameter values we use in the experiment are $b = 1, \alpha = 1.95$, and $\beta = 6$, and we next provide more specific predictions for OCP equilibrium play under those parameters. Here, the incentivizing principals reward agents upon observing outcomes $F \geq \hat{F}$, $\hat{F} \in \{2, 3, 4, 5, 6, 7\}$ and do not reward otherwise. In the cutpoint equilibria with the highest expected welfare for the principal, which are the standard predictions in such games (Persson and Tabellini, 2000; Bueno de Mesquita and Landa, 2015), the principal chooses an incentivizing strategy that calls for rewarding if and only if $F \geq z$, $z \in \{3, 4, 5\}$, and the agent chooses a level of effort $e^*$ such that $e^* + t = 4$. Thus, the agent of type 1 chooses effort 3, agent of type 2 chooses effort level 2, and agent of type 3 chooses effort level 1.[11] These are pooling equilibria, and in these equilibria, the principal's beliefs are such that she is indifferent between choosing to double $e$ or $t$.

One can construct equilibria in which the threshold for receiving a bonus is $z \in \{1, 2, 6, 7\}$. Those equilibria are semi-separating, in that the principal's posterior beliefs about the agent's type are not uniform, and there is a critical value in the $\hat{F}$ space such that the principal will double type for $F > \hat{F}$ and double effort for $F < \hat{F}$.

Given the payoff function, the principal will always prefer the pooling OCP equilibria – the equilibria with highest expected outcomes – to the equilibria with semi-separation, whether they are OCP or ONCP equilibria. That is, given the payoff structure, the principal always prefers to obtain the highest possible expected outcome $F$, in spite of the greater uncertainty about attribution that that entails, than to play an equilibrium in which it is easier to make a correct attribution but at the cost of a lower expected outcome $F$.

The multiplicity of equilibria creates a strategic coordination problem for the players. The presence of this problem is an intentional feature of our design. The rationale is two-fold. First, contractual uncertainty of reward and promotion expectations is a wide-spread feature of empirical environments with incomplete contracts, and, in particular, of environments in which discrimina-

[11] As is standard, these effort predictions are for agents endowed with the model payoffs in the experiment. In the implemented game, however, subjects face two kinds of uncertainty: about the realized noise draw and the strategic uncertainty about principals' critical outcome thresholds for rewarding the agents. This means that the actual choices of our subjects in the role of agents may be contingent on their expectations of outcomes and rewards, and reflect their underlying (unmodeled) risk preferences. We will examine the effects of risk preferences below.

tion is typically reported. One of our primary goals is to understand how the players behave in environments of precisely this kind. Second, allowing the players to take auxiliary actions that can reduce uncertainty over mutual expectations (e.g., making cheap-talk announcements before or in the middle of play) can have a separate psychological self-committing effect that is distinct from the purely informational coordination effect, altering what we think is the standard baseline behavior in such settings.

To get a handle on the expectations of agents' behavior in this setting, consider the following best-response analysis. Suppose the agent knows that the principal's reward rule is of the form "award a bonus iff $F > \hat{F}$," but is uncertain of $\hat{F}$. Let $p(\hat{F})$ be expected probability of bonus for $F = \hat{F}$, where $1 \geq p(7) \geq p(6) \geq p(5) \geq ... \geq p(1) \geq 0$. For each $t$, each choice $e$, there are three possible values of $F$: $t + e - 1$, $t + e$, $t + e + 1$. We can write the expected payoff for an agent of type $t$ from the effort choice $e$ given the realization of noise $\omega$ as

$$\frac{1}{3}E[u_A(t, e, \omega; p(\cdot))|\omega = -1] + \frac{1}{3}E[u_A(t, e, \omega; p(\cdot))|\omega = 0] + \frac{1}{3}E[u_A(t, e, \omega; p(\cdot))|\omega = 1],$$

where $p(\cdot)$ is given by $p(\hat{F})$ evaluated at the values of outcome given by the corresponding $(t, e, \omega)$.

Comparing this expectation at $e$ to one evaluated at $e = e+1$ and simplifying, we obtain that the expected payoff for $e + 1$ is higher than for $e$ if and only if

$$\begin{aligned} & (t+e+2)^{\frac{1}{2}} - (t+e-1)^{\frac{1}{2}} \\ & +p(t+e+2)\left((t+e+3)^{\frac{1}{2}} - (t+e+1)^{\frac{1}{2}}\right) \\ & -p(t+e-1)\left((t+e)^{\frac{1}{2}} - (t+e-1)^{\frac{1}{2}}\right) \\ \geq \quad & 3\frac{\alpha}{\beta} = 0.975. \end{aligned} \tag{1}$$

The agent's best response is, then, to choose $e = 1$ if inequality (1) fails at $e = 1$, choose $e = 2$ if inequality (1) holds at $e = 1$ but fails at $e = 2$ and choose $e = 3$ if inequality (1) holds at $e = 2$.

Note that the inequality includes two terms reflecting the agent's beliefs about the principal: $p(t+e+2)$ and $p(t+e-1)$. The first is the probability that the principal awards the bonus for the outcome that would be just out of the agent's reach at a given effort level $e$ – i.e., for the outcome that is one greater than what the agent could obtain with the luckiest noise draw at that value of effort. The second term is the probability that the principal would award the bonus for the outcome that the agent would assure at a given effort level $e$ even if the noise draw should turn out to be most unlucky. Inequality (1) is easier to satisfy when the former is larger (it enters the left-hand side with a positive sign) and when the latter is smaller (it enters with a negative sign). In what follows, we will refer to the conditions on these quantities implied by inequality (1) as the agents' *participation constraints*. We will refer to the lowest value of outcome that can earn a bonus as the *principal's demand* and to the agents' expectations of that value, which we modeled by $p(\cdot)$, as their *expectations of the principal's demand*. Assuming that the participation constraints hold, a distribution $p(\cdot)$ that sets a higher $p(t+e+2)$ and lower $p(t+e-1)$ than does another distribution describes a principal the agent believes will reward less for relatively low levels of performance yet reward more for performance levels that are relatively high – in other words, a principal who makes a stronger demand for high effort. Inequality (1) implies that *given that agents' participation constraints hold, agents choose higher effort when they expect the principal to be more demanding.* Of course, when the promise of reward becomes too remote ($p(t+e+2)$ becomes sufficiently low), the principal is "too demanding": for a given $t$, the incentives created for the agent may be such that the optimal effort actually drops.

Note that the baseline game described above does not assign identities to the players. In the identity treatments of the experiment, we prime and reveal to subjects their group identities by fixing labels to principals and agents and making them common knowledge within the pairs, but we do not alter the payoff structure described above. Because the payoff structure does not depend on these identities, one equilibrium behavioral expectation is that identity has no effect on behavior.

However, because players observe social identity matches, they may choose identity-contingent strategies leading to different equilibrium profiles being played in different identity matches (e.g., an OCP equilibrium profile with higher (lower) threshold for reward in in-group matches and an OCP equilibrium profile with lower (higher) threshold for reward in out-group matches). In this way, identity matches could matter as selectors of different equilibrium profiles. This role of identity is encapsulated in the hypotheses (below) concerning principals' (implicit) identity-contingent demands for outcomes necessary for receiving a bonus, the agents' identity-contingent expectations of principals' demands, and the agents' own identity-contingent (effort) choices.

### 3.3 Hypotheses

The hypotheses below derive from three sources of theoretical expectations: (1) the analysis of OCP and ONCP equilibria above; (2) the expectation of identity-contingent play in OCP equilibria, based on the analysis above and on the expectations of identity-contingent (and in particular of own-identity-favoring) play reported in the literature; and, (3) psychologically driven expectations that comport with theoretical and empirical results reported in the extant literature. While the behavioral comparison of the OCP and the ONCP play is an important element of our analysis, our primary focus is on the OCP play, which is the context where we expect to see the identity-driven effects, and so, in particular, on the OCP play that favors members of one's own group.

While, as we explained above, there are multiple equilibria in this setting, including equilibria indexed by different degrees of identity-contingent bias, we formulate the following hypotheses as descriptions of what we expect to be the average tendency on the part of the subjects. Our first three hypotheses are motivated by the expectation of sustained own-identity favoring play (Chen and Li, 2009; Landa and Duell, 2015), which, as mentioned, is consistent with OCP equilibrium play in our environment. The first hypothesis concerns what we referred to as "the principals' demands."

**Hypothesis 1** *(Principals' in-group bias in rewards): principals have lower demands for outcome from in-group agents than from out-group agents.*

The next hypothesis restates the expectation, but now as corresponding to the agents' own beliefs about the principals:

**Hypothesis 2** *(Agents' expectations of principals' in-group bias in rewards): Agents expect principals to have lower demands in in-group matches than in out-group matches.*

Our analysis of the agents' best responses in the OCP play yields the following hypothesis on the effect of a shift in the agent's expectation of the principal's demands, which we refer to as the *expected demand effect*.[12]

[12]Apart from the evidence on this effect, the experimental results we describe will also speak to

**Hypothesis 3** (expected demand effect): *Assuming the agents' participation constraints hold, agents' effort choices increase with their expectations of higher demands by the principals.*

The hypothesis requires that the agents' participation constraints (discussed in detail above) hold – that, in effect, the agents perceive the principals' demands to be such that they are worth trying to meet. We state the remaining hypothesis focusing on the case where these constraints indeed hold. (As the discussion of the results below suggests, though there is some evidence that this assumption fails for a small subset of the subjects, it is borne out in the bulk of our data.)

The agent's best response to lower demands from the principal is lower effort, and, to a higher demand, a higher effort. Given the expectation of lower demands in the own-identity-favoring OCP equilibria, in-group agents should, then, choose a lower effort, and out-group agents should choose a higher one. The agents' expectations of the principals' in-group bias in rewards, thus, condition the following hypothesis:

**Hypothesis 4a** *(Agents' equilibrium best response in own-identity-favoring OCP equilibria) Agents with higher expectations of principals' in-group bias in rewards choose higher levels of effort in out-group than in in-group matches.*

Hypothesis 4a is, notably, contrary to the expectation of own-identity-favoring behavioral bias on the part of the agents.[13] We next formulate that expectation, which were refer to as the *Agents' in-group bias effect*, as a rival hypothesis:

**Hypothesis 4b** *(Agents' in-group bias effect) All else equal, agents with expectations of higher principals' in-group bias in rewards choose higher levels of effort in in-group than in out-group matches.*

Linking principals' attribution decisions to agents' expected choices suggested by the previous two hypotheses, we can generate predictions regarding principals' attribution decisions. If agents condition on beliefs suggested by Hypothesis 2, then we should expect agents to choose *higher* effort

other predictions of OCP equilibria, though they are secondary to our focus in this paper.

[13]The contradiction would disappear if the participation constraints were to fail for out-group agents but hold for in-group agents.

in out-group than in-group matches, suggesting the following hypothesis[14]:

**Hypothesis 5a** *(Principals' own-identity-favoring OCP equilibrium attribution): Principals' propensity to attribute a given outcome to effort rather than type is lower in in-group than in out-group matches.*

However, if the expectation of agents' in-group-bias effect (Hypothesis 4b) is correct and dominates the expected demand effect, the correct expectation for the attribution by the principals would be the following rival hypothesis, which is behaviorally consistent with the prediction of the ultimate attribution bias:

**Hypothesis 5b** *(Principals' in-group-bias-effect attribution): Principals' propensity to attribute a given outcome to effort rather than type is higher in in-group than in out-group matches.*

Our last two hypotheses are meant to isolate the effect of strategically driven expectations on attribution decisions. The first of these hypotheses concerns the attribution choices of principals who are playing outcome-non-contingent reward strategies. We anticipate these principals' attribution choices by asking how they would affect agents' best responses. As explained above, in contrast to the principals who are playing outcome-contingent strategies and who may have a bias in reward decisions, we should expect these principals' attribution choices to be symmetric.

**Hypothesis 6** *Principals' asymmetric attribution exists only for principals in outcome-contingent-play equilibria.*

The last hypothesis is that asymmetric attributions are driven by the mutual expectations

[14]Note that this hypothesis depends on the assumption that subjects in out-group matches satisfy the participation constraints not too much worse than the subjects in the in-group matches – the assumption that, in effect, holds across the OCP equilibria that maximize the principal's welfare. When that assumption fails, the claim of the hypothesis may not hold.

that are set in motion by the strategic feedback, from rewards to the effort choice in expectation of rewards. When such expectations are irrelevant, we should see no attribution asymmetries.

**Hypothesis 7** *The asymmetric attribution effect disappears in the absence of strategic incentives.*

# 4 Experimental design

The structure of our laboratory experiment approximates the principal-agent relationship in a hierarchical bureaucracy, which we explore with two experimental treatments. The STRATEGIC treatment features the opportunity to reward the agent with a bonus (henceforth, referred to as the availability of the sanctioning device), following closely the model described above. The NON-STRATEGIC treatment removes the sanctioning device.[15] The experiment included 110 subjects in the STRATEGIC treatment (2200 subject-round observations) and 38 subjects in the NON-STRATEGIC treatment (760 subject-round observations).

Prior to the principal-agent game in each session, subjects' identities are induced as described in detail below.[16] Then, subjects are assigned to the role of either an *agent* (called "Player 1")

[15]In order to ascertain the relative power of incentives created in our treatments, we conducted additional exploratory experimental analysis in the standard principal-agent settings with no identity-inducement (all treatments are described in detail in Section A in the appendix). We measure this by comparing the average responsiveness of agent's effort to her type with and without identity and find no significant difference (Section B.5 in the appendix). No further treatments were conducted. The analysis reported in the main text and the appendix describes the full set of observations. The experiment was not pre-registered.

[16]At the beginning of each experimental session, right before the identity inducement, we elicit risk-attitudes in a non-hypothetical, small stakes setting following the design presented by Holt and Laury (2002). We evaluate how risk preferences affect agents' choices in Section B.3.3 in the appendix. We show that the magnitudes of the expected bias and the expected demand effects are importantly contingent on agents' risk preferences, pointing to an important under-explored factor in accounting for individuals' responses to discrimination.

or a *principal* ("Player 2") and remain in that role for the duration of the experiment. They are randomly re-matched into pairs of one agent and one principal in each of 20 rounds of a session. The implemented random matching protocol is the perfect stranger matching for the first (number of subjects in the session)/2-rounds of each session, followed, in subsequent rounds, by subjects meeting previous matches again in random order once.[17] Subjects received a show-up fee of $7 and performance-based payments of on average $23. Payments from the principal-agent game where taken from the two highest round-payoffs from three randomly selected rounds.[18]

**Group identity inducement** At the beginning of each session of both the STRATEGIC and the NON-STRATEGIC treatments, subjects were assigned to groups according to their stated preferences for either *Klee* or *Kandinsky* paintings and performed in a quiz collaboratively with their new fellow painter group members. Members of both groups, Klees and Kandinskys, in all treatments performed approximately equally well. In the subsequent principal-agent game part of the experimental session, the identities of both subjects within a matched pair were displayed for them on the screen along with icon-sized paintings by the corresponding artists. In this way, subjects learn whether they are in an *in-group* or *out-group* match.[19]

[17]Matching protocol and anonymized interaction between subjects precludes direct exchanges, and thus provides us approximately with as many independent observations as subjects in the experiment. We report standard errors clustered by subject throughout. We find no robust evidence for learning effects. Our main results are robust to accounting for the history of play at the subject-level and, except for our finding on bias in attribution decisions, all results are stable when comparing first and second half of the experiment (see Section B.3.5 in the appendix).

[18]This helps avoid endowment effects and hedging in lottery-like choices under uncertainty (Charness et al., 2016).

[19]See Tajfel and Billig (1974), Chen and Li (2009), and Landa and Duell (2015) for the use of painter-preferences to induce identities. Considerable experimental literature has shown the effectiveness of minimal groups in inducing responses to identity that resemble those observed outside the laboratory with naturally occurring group identities and the monotonicity of identity effect in identity strength (Eckel and Grossman, 2005).

**STRATEGIC treatment: Principal-agent game with sanctioning device** The game simulated in the STRATEGIC treatment mirrors the structure and payoffs laid out in Section 3. By monetarily incentivizing subjects in the role of agents, we create concerns about outcomes because agents value receiving a bonus from the principals. Subjects in the role of principals benefit from high outcomes. While the principals do not bear a direct cost of awarding the bonus, the agents' choices respond to the principals' bonus-awarding strategy. Because those choices affect principals' payoffs, they create a benefit to the principals of a bonus-rewarding strategy that induces higher choices by the agents, as is standard in moral hazard settings. As will become apparent in our analysis, agents in our experiment clearly respond to their expectations of principals' demands.

All subjects, agents and principals, were instructed that agents would be given payoff information on the terminal screens whenever they are making their choice of effort. Before agents make their investment decision but after they observe their randomly assigned type, they are asked: "What minimal outcome do you think Player 2 will demand to give you a bonus?" They are shown payoffs, contingent on their answer type, as a function of the level of effort they may choose and the possible values of noise. Agents may click through all possible values of outcome in any order, may choose to go back and forth between values, or not select to see any potential payoffs. Inputting their expected minimal rewarded outcomes to generate contingent payoffs enables the agents to obtain a more highly rewarded choice, thus creating a monetarily incentivized revelation of their belief.[20] All subjects (principals and agents) were shown the agents' decisions screen and extensive examples of principals' applying incentivizing strategies in the instructions as well as in the pre-play comprehension quiz (all these examples are identity-blind).

After agents made their choice of effort, and outcomes are realized, principals are asked to double either the effort or the type component in their round payoff. In making that choice, principals are effectively stating their (motivated) belief on whether outcomes are more driven by the agent-controlled attribute (effort – an internal, dispositional attribute) or by agent-uncontrolled attribute (type – an external, situational attribute that is randomly assigned). The principal's doubling de-

[20]More specifically, we capture agents' beliefs by recording the mean expected demands of all clicks they make in each round. Section B.3.2 in the appendix gives more data on frequency and extent of agents' use of this tool.

cision, thus, models the choice situation that is at the core of the ultimate attribution error. In this way, principals' beliefs are elicited monetarily rewarding correctness in the attribution decision. For convenience, we will refer to the principal's decision to double effort upon observing a given outcome as "attributing the outcome to effort" and the decision to double type as "attributing the outcome to type."[21] Because agents' beliefs are elicited by a procedure in which the effort choices and the underlying beliefs are (procedurally) interdependent, we should expect the relationship between those variables in the data to be closer than would be otherwise. For reasons of external validity, variation in the responsiveness of agent actions to their beliefs is, therefore, not an appropriate focus for a study with this design. Our focus in characterizing strategic discrimination is, rather, on principals' attribution and reward decisions and on responsiveness of agents' beliefs/actions to the principals.

**NON-STRATEGIC treatment: principal-agent game without sanctioning device** The NON-STRATEGIC treatment replaces the principals' sanctioning tool with exogenously given incentives to the agents. In this treatment, agents' payoffs are given by $G(F, e) = \beta\sqrt{F} - e$, with $\beta = 4$. Note that, as in the STRATEGIC treatment, $G(\cdot)$ is increasing in outcome $F$ and decreasing in effort $e$. The functional form of the payoffs and the parametrization were chosen to be as close as possible to those in the STRATEGIC treatment and to induce optimal choices for agents, conditional on their type, that are identical to the optimal choices in the maximal principal welfare (3-4-5 threshold) OCP equilibria in the STRATEGIC treatment game. Principals observe the outcome and are asked to make their attribution decision, incentivized in the same way as in the STRATEGIC treatment.[22]

[21]On the screen where principals make reward and attribution decision, we also asked principals whether they thought type or effort was the higher quantity (with a strong correlation of .74 ($p = .00$) between subjects' guess and their attribution decision.

[22]A different way of designing the study to get at the difference between strategic and non-strategic settings would be to randomly assign probabilities of sanctioning device being available rather than exogenously adjusting payoffs. The downside of that approach in our setting is that a low probability of being rewarded (for the non-strategic setting) would imply that agents' effort would approach the minimal possible level, undermining the variation in the principal's beliefs.

**Summary of the experimental set-up** The sequence of moves in each round of the experiment is as follows (the principal's reward decision and elicitation of agent's beliefs is omitted in the NON-STRATEGIC treatment):

1. Agents are assigned a *type* and privately informed about its realization (1, 2, or 3).
2. Agents choose a level of *effort* (1, 2, or 3) and state their expectation about which minimal outcome principals demand to see to give a bonus (1-7, *expected demands* – agents' beliefs).
3. *Noise* and *outcome* are realized where the value of *outcome* is the sum of agent's *type* (1, 2, or 3), agent's chosen level of *effort* (1, 2, or 3), and a *noise* realization (-1, 0, or 1).
4. Principals learn the value of *outcome* (1-7).
5. Principals choose whether to attribute outcomes to *type* or *effort* (*attribution decision* – principals' beliefs) by doubling the payoff contribution of the *type* or *effort* component of *outcome* and whether to give the agent a bonus *(reward decision)*.
6. Round feedback: principals observe whether type or effort was higher and agents learn the principal's reward decision (where applicable).

# 5 Results

In Section 5.1, we first summarize and compare principals' choices across in- and out-group matches as well as in STRATEGIC and NON-STRATEGIC treatments. Consistent with the theoretical expectations set out above, we distinguish behavior of two sets of principals, incentivizing and non-incentivizing, whose strategies suggest very different best responses from the agents. We establish that incentivizing principals in the STRATEGIC treatment tend to be in-group biased in their rewards choices and to make identity-contingent attributions of outcomes. In contrast, non-incentivizing principals in the STRATEGIC treatment and principals in the NON-STRATEGIC treatment do not make attributions contingent on identity. In Section 5.2, we investigate agents' effort choices. While we find that in the aggregate, those choices are *not* identity-contingent, focusing on agents' effort choices in interaction with their expectations about principals' outcome demands (relevant to the STRATEGIC treatment) yields a more nuanced picture. We show, in particular, that agents respond in heterogeneous but identity-contingent way to their expectations of principals demands and that those responses are driven by their expectations of principals' demand and of principals' bias in rewards. In Section 5.3 we elaborate on the interpretation of identity-contingent choices and beliefs and provide evidence that principals fail to correctly anticipate the strength of the expected demand effect in out-group agents.[23]

### 5.1 Principals' choices and beliefs

To properly characterize the attribution decisions and develop the comparison of those decisions in the STRATEGIC and NON-STRATEGIC treatments, it is necessary to begin by distinguishing incentivizing and non-incentivizing principals in the STRATEGIC treatment (the distinction is moot in the NON-STRATEGIC treatment, since the principals in that treatment do not have a reward decision, but see below for some estimates).

In the STRATEGIC treatment, principals' behavior consistent with outcome-contingent play, following strategies associated with OCP equilibria, is clearly prevalent. As anticipated by those equilibria, the distribution of outcomes is centered at 4; 75% of observations fall within the range between 3 and 5. Principals' reward choices are systematically increasing in observed outcome. The marginal effect of outcome on rate rewarded is $.07\,(.03, .11)$ in in-group matches and $.10\,(.06, .13)$ in out-group matches.[24] Further characterizing outcome-contingent play, we distinguish between two distinct behavioral groups of principals: those whose bonus-awarding strategies are contingent on the received outcomes (incentivizing principals in the context of the OCP equilibria) and those whose strategies are not (non-incentivizing principals in the context of the ONCP equilibria). Incentivizing principals constitute 76% of the principals in the STRATEGIC treatment. For each of these principals, we compute the individual-specific threshold of outcome that minimize errors in categorizing their respective reward decisions.[25]

The inferred principal-specific reward thresholds, whose distribution is given in Figure 1, vary from 2 to 7. The average threshold in the STRATEGIC treatment is lower in in-group (3.93) than

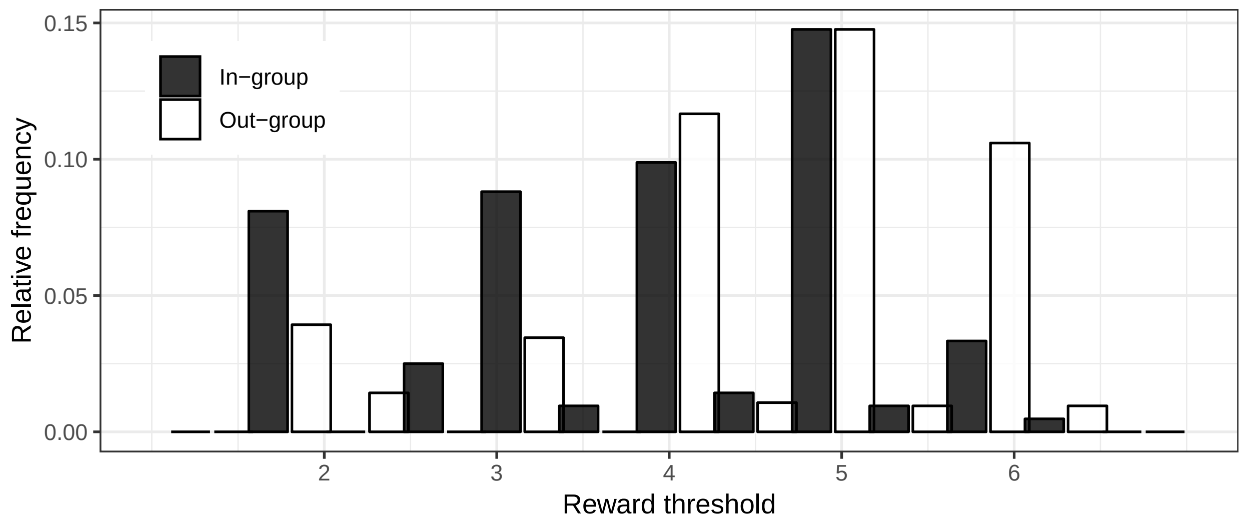
[23] Summary statistics for all variables are given in Section B of the appendix.

[24] Marginal effects are estimated from the regression of reward decision shown in Table B.5 in the appendix. 95% bootstrapped confidence intervals based on a subject-clustered bootstrap are reported in parentheses throughout.

[25] The average share of reward decisions incorrectly classified by the error-minimizing threshold is .19 suggesting that principals' reward decisions are largely consistent with their inferred individual thresholds.

in out-group matches (4.56), implying that incentivizing principals are less demanding in in-group than in out-group matches; the significant difference in means is $-.63\,(-1.07, -.14; p < .01)$.[26]

Figure 1: Incentivizing principals' reward thresholds by in-group status in the STRATEGIC treatment (rounded to steps of .5).




Our first result summarizes the preceding discussion:

**Result 1** *(Principals' in-group bias in rewards) The bulk of principals in the STRATEGIC treatment play incentivizing reward strategies. Among these incentivizing principals, significantly more demand higher outcomes for rewarding out-group than in-group agents [supporting Hypothesis 1].*

The comparison of the rates at which principals reward in in-group and in out-group matches reinforces this result: the marginal effect of in-group status on principals' rewards, holding outcome at its mean, is $.09\,(.00, .20)$. We find differences in incentivizing principals' reward rate in in- and out-group matches above the reward threshold (.69 vs .77) as well as below it (.17 vs .22) with a difference of $.08\,(.00, .17; p = .07)$ and $.05\,(-.01, .11; p = .08)$, respectively.

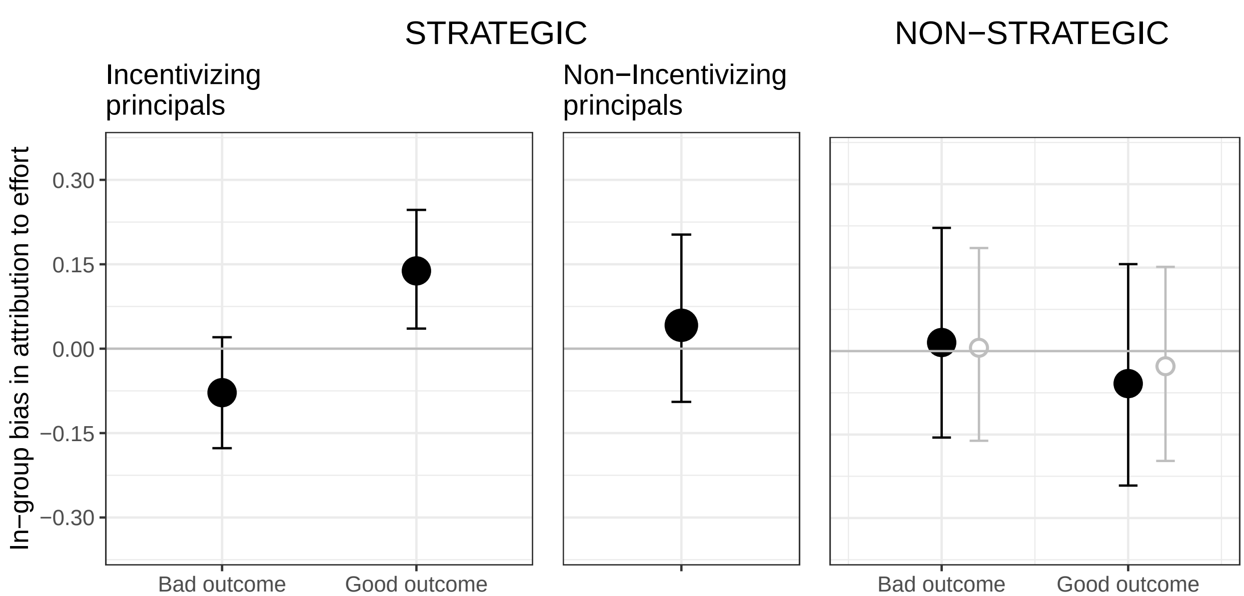
We next turn to describing the principals' attribution decisions in STRATEGIC and NON-STRATEGIC treatments to assess existence and degree of in-group bias in attribution to effort. To fix concepts, let the *in-group bias in attribution* at outcome $O$, $b(O)$ be the rate of attribution to effort in in-group matches at $O$ minus the rate of attribution to effort in out-group matches at $O$. Different outcome thresholds in the 3-5 range are consistent with OCP equilibria that, in our model,

[26]Figure B.1 provides the subject-level distribution of reward thresholds.

maximize the agents' effort. This means that restricting attention to these equilibria, principals' attribution decisions in the STRATEGIC treatment may be driven by attribution biases that would be "canceling" each other at any exogenously fixed level of performance in that range. To get a valid measure of attribution bias, we need to evaluate attributions at the thresholds of good/bad performance that are subject-specific. The reward threshold values computed above provide natural individual-specific definitions of what outcomes a given principal perceives as good performance (at and above the threshold) as opposed to bad performance (below the threshold).

The left and the middle panels of Figure 2 display in-group bias in attribution for in the STRATEGIC treatment. Incentivizing principals attribute outcomes to effort more often in out-group than in-group matches when the observed outcome is bad, at rates of .59 vs .51, respectively, with a difference of $.08\,(-.02, .18; p = .11)$, but more often to effort in in-group than out-group matches when the outcome is good, .56 vs .43, respectively, with a difference of $.14\,(.04, .25; p < .01)$. Non-incentivizing principals (who always or never reward) attribute outcomes to effort in both in-group and out-group matches at similar rates: .57 and .62 (the difference of $.04\,(-.09, .20)$ is not systematically different from zero).

Figure 2: In-group bias in attribution to effort by outcome and treatment. Gray marker in NON-STRATEGIC panel is conservative estimate.




Because of the nature of the NON-STRATEGIC treatment, we can identify neither who the incentivizing principals are nor, endogenously, what constitutes good vs. bad outcomes in principals' eyes. For this treatment, we estimate attribution bias drawing the line of "good" outcomes with respect to the NON-STRATEGIC treatment at 5 (just above the median) or above, and "bad outcomes" at 3 (just below the median) or below.[27] While this implies a limitation, the two sets of cases that this demarcation creates are outside of "grey area," and our confidence in the treatment comparison with respect to these cases is particularly high. The black markers in the right panel of Figure 2 show in-group bias in attribution, pooling all principals in the NON-STRATEGIC treatment. In contrast to the STRATEGIC treatment, we do not observe a significant in-group bias in attribution for the NON-STRATEGIC treatment, either when the principals observed bad outcomes, $.01\,(-.16, .19)$, or when they observed good outcomes, $-.03\,(-.20, .15)$. The absolute levels of attribution to effort in this treatment are .80 in the in-group and .79 in the out-group for good outcomes and .48 in the in-group and .50 in the out-group for bad outcomes.[28]

Given that we cannot distinguish incentivizing from non-incentivizing principals in the NON-STRATEGIC treatment, the estimate of in-group bias in attribution for the full set of principals is averaging across two types of principals who, as our analysis of the STRATEGIC treatment suggests, would behave differently in the strategic setting. Given this implicit averaging, it would be reasonable to expect the resulting estimate of in-group bias in attribution to be lower than the bias observed among incentivizing principals in the STRATEGIC treatment, simply due to "mixing-in" of the non-incentivizing types, rather than due to differences in behavioral implications of the two treatments.

The right sub-panel of Figure 2 shows a conservative estimate of in-group bias in attribution for this treatment in gray – an estimate that is biased against finding the average treatment effect. To arrive at this estimate, we look at the attribution decisions of the 76% most in-group biased principals (in attribution choices) in the NON-STRATEGIC treatment – the share of incentivizing principals among all principals in the STRATEGIC treatment. Strikingly, we find that the

[27]We have no clear expectation about whether a principal ought to perceive the median outcome of 4 as good or bad.

[28]The attribution of an outcome to effort at a rate below .50 means, in effect, that the principal was attributing the outcome to agent's type more than to her effort.

attribution bias at good outcomes among these (most biased) principals in the NON-STRATEGIC treatments is still smaller than that among the incentivizing principals in the STRATEGIC treatment.

We summarize the preceding analysis in the following two results:

**Result 2** *(Principals' own-identity-favoring attribution) In the STRATEGIC treatment, there exists a systematic attribution asymmetry between in- and out-group matches for the incentivizing principals and no asymmetry for non-incentivizing principals [supporting Hypotheses 5b and 6].*[29]

**Result 3** *(Strategic discrimination) Principals' in-group favoring choices and the accompanying asymmetric beliefs disappear in the NON-STRATEGIC environment [supporting hypothesis 7].*

### 5.2 Agents' choices and beliefs

Agents' effort choices are decreasing with type, suggesting that agents are playing a pooling strategy (consistent with our prior observation on the distribution of outcomes). The marginal effect of type on effort is $-.18\,(-.25, -.10)$ in the STRATEGIC and $-.52\,(-.70, -.34)$ in the NON-STRATEGIC treatment.[30] The evidence shows no in-group bias in effort (i.e., a higher level of effort when matched with an in-group principal than when matched with an out-group agent), on average. Agents in the STRATEGIC treatment invest slightly more into effort in in-group than in out-group matches but this average difference is small and not statistically significant: $.05\,(-.05, .15)$. This holds true at every level of expected demand. We do not find a significant difference in the NON-STRATEGIC treatment either, though the effort is somewhat lower in the in-group than in the out-group (the difference is $.06\,(-.09, .21)$). To summarize:

**Result 4** *In the aggregate, agents do not show in-group bias in effort either unconditionally nor conditional on expected demands in either STRATEGIC or NON-STRATEGIC treatment [contrary to Hypothesis 4b].*

This result may suggest that agents are not strategically responding to principals' identity-contingent asymmetric rewarding. However, interpreting these *average* effort decisions is difficult

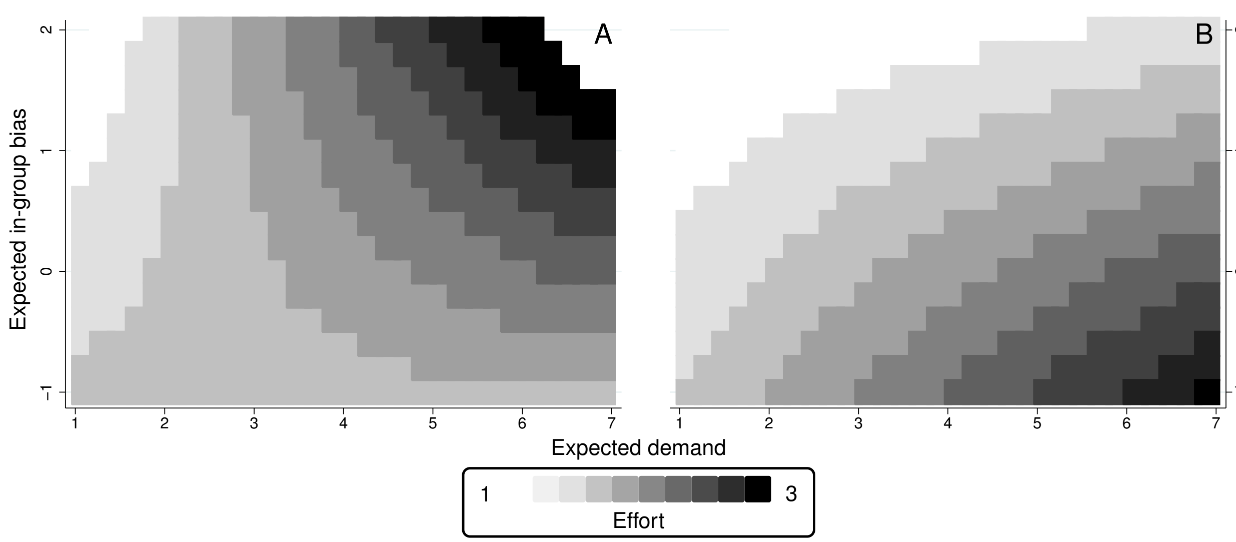
[29] Section B.3.1 in the appendix shows a more detailed discussion of the robustness of this result.

[30] Estimates are computed based on the regression reported in Table B.7 in the appendix.

without anchoring agents' choices in their beliefs about the principals similar to they way we anchored principals' choices in their beliefs about the agents. Indeed, the average of agents' effort choices here is concealing a substantial variation in expected demands and, in consequence, in their best responses to those beliefs. Agents' aggregate beliefs in the STRATEGIC treatment about principals' biases are asymmetric (identity-match contingent), consistent with the overall direction of bias in principals' actual reward choices. The average difference in expected demands between the in- and out-group matches, as elicited in the STRATEGIC treatment, is $.10\,(-.05, .24; p = .20)$ indicating that the distribution of expected in-group bias in principals' reward choices is somewhat skewed (though not rising to conventional significance levels). Thus, we have the following result:

**Result 5** *Agents tend to believe that they face systematically lower outcome demand for a bonus reward in in-group matches than in out-group matches [weakly supporting Hypothesis 2].*

Figure 3: Predicted levels of effort plotted over expected in-group bias and expected demands for in-group matches (Panel A) and out-group matches (Panel B).




We next consider how agents respond to their expectations of principals' demands, and provide evidence of both perceived identity bias and agents' identity-contingent responsiveness to that perception. We begin with the observation about the *expected demand effect*: in both in- and out-group matches, effort is increasing with expected demands. A one-unit increase in expected demands leads to an average increase in effort of $.20\,(.10, .29)$ in in-group and $.19\,(.06, .31)$ in out-group matches.

This phenomenon can be seen in Figure 3. Panels A and B of the figure present predicted values of effort as a function of agents' expected demand and agents' expectation of the principals' in-group bias in rewards for in-group and out-group matches, respectively. Reading the heat-plots from left to right, we clearly see the increase in effort (coloration becoming darker) with higher expected demand. However, as the expected bias increases (reading from bottom to top), the out-group agents decrease (and in-group agents increase) their effort especially in the far right zones of the maps, i.e., where the expected demands are highest.[31]

This is the evidence of agents' *expected bias effect*: the difference in effort between in- and out-group matches is increasing with agents' expectation of the principals' in-group bias in rewards. In particular, when agents believe that to be rewarded, they are expected to deliver lower outcomes in in- than out-group matches, their effort is predicted to be .14 (.02, .26) higher in in- than out-group matches; when they believe higher outcome is required for reward in in- than out-group matches, the difference estimate is $-.08\,(.04, -.20)$. Differences in effort choices are smallest for agents who do not expect identity-contingent differences in principals' demands. Note that the expected demand effect and the expected bias effect work, on average, in opposite directions. While according to the former, agents' expectations of lower demands from in- than out-group principals should induce higher effort in out- than in-group matches, such expectations lead, according to the latter, to higher effort in in- than out-group matches.[32]

It is also noteworthy that for lower types, the sufficiently high demand leads to a decrease in effort in out-group matches, pointing to a pattern of behavior that is illustrated by our motivating example of Alice and Bob: expecting Bob's demands may lead Alice to dis-invest if she perceives those demands to be very high (more likely to occur in the out-group matches), and so the probability of receiving a reward low, and expects the principals to be in-group biased. Indeed, agents who expect high demands of 5 and above and higher demands from out-group than in-group principals,

[31]More evidence supporting this claim is provided in Figure B.6 in the appendix. Local regressions show that the expected demand effect is significant and positive when estimated sub-setting agents by expected bias while the expected bias effect only is significant and positive for agents with expected demand 4 and above (See Tables B.9 and B.10).

[32]In Section B.3.2 of the appendix, we show that both the expected demand effect and the expected bias effect increase with agents' risk-aversion.

choose levels of effort that are .39 ($-.35, .96$) lower in out-group than in-group matches. However, for the bulk of the data, the expected demand in out-group matches is below such levels, and the average overall effect is the increase in effort, as depicted in the figure.

Properly accounting for the levels of (and differences in) expected demands, thus, both explains the aggregate-level finding of no difference between agents' behavior in in- and out-group matches and corrects the mistaken impression in may convey. The following result summarizes the above discussion and presents our key substantive conclusions on agents' effort choices:

**Result 6** *Agents' choices display an expected demand effect as well as an expected bias effect in in- and out-group matches [supporting Hypotheses 3 and 4b].*

### 5.3 Linking principals' and agents' choices and beliefs

Our key results show that principals' choices and judgments are systematically group-dependent and that agents anticipate and respond to that dependence. But are the principals' attributions ultimately correct in their assessments of the agents' decisions? As the evidence of a robust expected demand effect in agents' choices suggests, agents respond to higher expectation (in this case, in out-group matches) by increasing their effort to meet the demand (see Section 5.2). Even if agents' choices are subject to the expected bias effect, if they expect the demands in the out-group matches to be sufficiently high relative to the in-group demands, the expected demand effect may override the expected bias effect, producing a *higher*, not lower, effort in the out-group matches. Our regression-based estimate of agents' effort reinforces this conclusion. When the agents expect to be facing symmetric demands from in- and out-group principals, they choose higher effort in in-group matches (the difference is .14 ($-.09, .37$) in favor of the in-group), but the sign of the difference flips if the agents expect to meet higher demands in the out-group match: expecting that the principals' demand is two outcome points higher in out- than in-group matches increases the difference between average effort in out-group and in-group matches to .27 ($-.70, .17$).[33]

Whether the in-group bias in rewards and the in-group bias in attribution can be made strategically consistent is, thus, a function of the size of the reward bias and the assumptions we make about the agents' corresponding beliefs. A natural such assumption for the purposes of this as-

[33]Estimates are based on Model 4 in Table B.8 in the appendix.

sessment is that a principal's reward bias is (counterfactually) the object of a common conjecture with the agents. With this assumption, then, we can ask whether the principals' attribution decisions are correct if the agents correctly anticipate principals' reward biases. When the reward bias is relatively small, the two biases are mutually consistent. As the in-group rewards bias (and its expectation on the part of the agents) grows, the principals should be expecting the size of the expected demand effect increasingly to counter the size of the expected bias effect; as this occurs, the persistent in-group bias in attribution becomes evidence of the principals' under-appreciation of the force of the expected demand effect. Indeed, we find that, while agents' average effort clearly increases with expected demands by the principals (Figure 3), principals' attribution of good outcomes to effort, on average, does not increase with their demands. The marginal effect of principals' reward threshold on the attribution to effort is $-.05\,(-.14, .03)$ in in-group and $-.02\,(-.11, .07)$ in out-group matches for incentivizing principals who are in-group biased in their reward decision.[34]

We summarize the preceding in the following result:

**Result 7** *Principals' attribution decisions suggest a systematic underestimate of the positive influence of the expected demand effect on the out-group agents' effort choices.*

## 6 Discussion: interpreting the evidence

In the evidence on the discriminatory behavior we present, the principals' identity-contingent attribution choices reflect their expectations about agents' effort choices, which, in turn, are responding to expectations of the principals' reward choices. Consistent with the idea of strategic discrimination, the contrast between the STRATEGIC and NON-STRATEGIC treatments suggests that the effect of the strategic relationship in an identity-salient context is to create asymmetric behavioral expectations associated with the information entailed in the identity markers.

What is the source of that information? One plausible source is a norm of mutual reciprocity that may correspond to an equilibrium of a different game – played outside the lab – in which identity-indexed interactions are repeated and the mutual in-group favoritism (reciprocity) is the

[34]Estimates are taken from a regression of attribution to effort on outcome, in-group status of the matched principal, principals' individual reward threshold, the interaction of these variables, and round of play.

focal equilibrium. Such an equilibrium may motivate subjects' interpretations of the proper behavior in social identity contexts, and the principals' attribution choices would be understood as encapsulating the expectation that comes with that norm. This possibility would still be consistent with a distinctly strategic account of the evidence of discrimination we describe, even if it would be driven in the first place by the equilibrium beliefs induced outside, rather than inside, the lab.

Yet, it's important not to overweight the force of reciprocity as the explanatory account. The contrast between the attribution asymmetries in the STRATEGIC and the absence of such asymmetries in the NON-STRATEGIC treatment casts doubt on reciprocity, or at least on the reciprocity that is taken to be independent of the strategic properties of the proximate interactions (such as, for example, those that were instantiated in the lab). Even if the reward choices were somehow based on expectations of reciprocity, it is clear that attribution choices are responding to features of that proximate environment rather than being driven by considerations from outside the lab. The same evidence also suggests that the discriminatory behavior we report is unlikely to be driven by a "taste for discrimination," even one that may be entailed in internalized identity-contingent reciprocity. The "taste for discrimination" mechanism suggests more instinctive, less well-considered behavior than the behavior that is contingent on the presence of a strategic relationship.

A different piece of evidence, from exit surveys following our STRATEGIC treatment, reinforces the view that the discriminatory judgments we document are well-considered and that their authors are self-aware. In the survey, we asked questions that allow us to evaluate the relationship between subjects' self-awareness and their choices in the experiment. In their responses, 44% of incentivizing principals indicate that they were influenced in their reward decision by the group membership of their matched agent, in contrast to no non-incentivizing principals' saying that group identities mattered. The contrast with respect to the attribution decision is less stark but still significant: 35% of incentivizing principals claimed to be influenced by group membership in their attribution choices compared to only 23% of principals who always or never rewarded. Further, within the set of incentivizing principals, awareness of one's own bias in reward decisions increases attribution of good outcomes to effort in in-group in contrast to out-group matches. For those who are aware of their reward biases, the in-group bias in the attribution of good outcomes to effort is $.27\,(.06, .49)$ in contrast to $.14\,(-.03, .30)$ for those who admit no such awareness. In sum, principals whose reward and attribution choices are asymmetric tend to be aware of it.

# 7 Conclusion

Our analysis has provided a behavioral evaluation of strategic discrimination – an important contributor to identity-based discrimination that has resisted clean identification and systematic analysis in previous work.

The results we presented have a number of implications. As a descriptive matter, the evidence of strategic discrimination suggests, first, that the existing measures of prejudice in the observational studies may be partial-equilibrium: they may be identifying a joint measure of prejudice and rational expectations associated with an equilibrium performance, rather than prejudice alone. But the measurement problem is subtle and points to the importance of laboratory-based research designs: given the strategic incentives we identify, prejudiced principals may be observationally indistinguishable from unprejudiced ones, and, facing either of them, agents who do not share their principals' salient social identity would be equally justified in expecting to be treated more harshly than their colleagues who do share it, and so, would also be justified in reducing effort in anticipation of the lower likelihood of receiving deserved recognition. And second, the disjunction between the aggregate-level evidence of discrimination and gaps in pay and promotion, on one hand, and of the rare successes of the complaints of discrimination at the individual-case level, on the other, may be indicative of discrimination as a strategic phenomenon. In such discrimination, principals' behavior may be consistent with actual differences in agents' contributions – yet those differences are endogenous to the expectation of discriminatory behavior from the principals, and, our evidence suggests, principals tend to under-appreciate the effort from the out-group agents (or, equivalently, out-group agents may be justified in reducing their effort still farther than they, in fact, do). The bottom line, though, is that strategically induced attribution asymmetries may be *a*, if not *the*, first-order phenomenon when it comes to accounting for discriminatory choices by principals, and, as such, need to be addressed in both positive studies of discrimination and in policy design.

With an eye toward normative considerations related to policy design, the most immediate observation is that, given the history of discrimination in the world outside the lab, the expected result of strategic discrimination is, probably, the persistence of the familiar asymmetric pattern, with the memes associated with strategically reinforced beliefs systematically undermining the historically underprivileged groups as surely as does well-ingrained prejudice. Recognizing the sources of discrimination may, however, help in properly calibrating anti-discrimination policies. A broad policy implication of our analysis is that reactive, discrimination-penalizing policies may be insufficient for defeating discrimination. More effective solutions should look to influence the formation of beliefs that support the asymmetric identity-based strategic responses – from affirmative-action policies at the managerial level (for which our analysis of strategic discrimination provides an efficiency-based rationale) to oversight schemes that suppress information about group identities (of agents, but no less importantly, also of principals), and reward agents strictly on observable measures of performance without conditioning on principals' beliefs of their causes. We leave the closer behavioral analysis of these and other institutional solutions to future work.

# References


Akerlof, G., and R. Kranton, 2000: Economics and identity. *Quarterly Journal of Economics*, **115 (3)**, 715–53.

Allport, G., 1954: *The Nature of Prejudice*. Reading: Addison-Wesley.

Altonji, J. G., and R. M. Blank, 1999: Race and gender in the labor market. *Handbook of labor economics*, **3**, 3143–3259.

Arrow, K., 1973: The theory of discrimination. *Discrimination in labor markets*, Vol. 3, Princeton: Princeton University Press.

Becker, G. S., 1971: *The economics of discrimination*. University of Chicago press.

Benabou, R., 1996: Equity and efficiency in human capital investment: the local connection. *The Review of Economic Studies*, **63 (2)**, 237–264.

Bendick, M., 2007: Situation testing for employment discrimination in the united states of america. *Horizons stratégiques*, **(3)**, 17–39.

Bertrand, M., and S. Mullainathan, 2004: Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American Economic Review*, **94 (4)**, 991–1013.

Besley, T., 2006: *Principled agents?: The political economy of good government*. Oxford University Press on Demand.

Bueno de Mesquita, E., and D. Landa, 2015: Political accountability and sequential policymaking. *Journal of Public Economics*, **132**, 95–108.

Captain, S., 2017: Workers win only 1% of federal civil rights lawsuits at trial. URL https://www.fastcompany.com/40440310/employees-win-very-few-civil-rights-lawsuits.

Charness, G., U. Gneezy, and B. Halladay, 2016: Experimental methods: Pay one or pay all. *Journal of Economic Behavior & Organization*, **131**, 141–150.

Chen, Y., and S. Li, 2009: Group identity and social preferences. *American Economic Review*, **99 (1)**, 431–57.

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

Eckel, C. C., and P. J. Grossman, 2005: Managing diversity by creating team identity. *Journal of Economic Behavior & Organization*, **58 (3)**, 371–392.

Falk, A., and C. Zehnder, 2007: Discrimination and in-group favoritism in a citywide trust experiment. Tech. rep., IZA Discussion Papers.

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

Fryer, R. G., J. K. Goeree, and C. A. Holt, 2005: Experience-based discrimination: Classroom games. *The Journal of Economic Education*, **36 (2)**, 160–170.

Gailmard, S., and J. W. Patty, 2012: Formal models of bureaucracy. *Annual Review of Political Science*, **15**, 353–377.

Giulietti, C., M. Tonin, and M. Vlassopoulos, 2017: Racial discrimination in local public services: A field experiment in the united states. *Journal of the European Economic Association*.

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

Gordon, S., 2009: Assessing partisan bias in federal public corruption prosecutions. *American Political Science Review*, **103**, 534–54.

Haan, T., T. Offerman, and R. Sloof, 2015: Discrimination in the labour market: The curse of competition between workers. *The Economic Journal*.

Hewstone, M., 1990: The ultimate attribution error? a review of the literature on intergroup causal attribution. *European Journal of Social Psychology*, **20**, 311–35.

Holt, C., and S. Laury, 2002: Risk aversion and incentive effects. *American Economic Review*, **92 (5)**, 1644–55.

Holzer, H., and D. Neumark, 2000: Assessing affirmative action. *Journal of Economic Literature*, **38 (3)**, 483–568.

Huq, A., and T. Ginsburg, 2018: How to lose a constitutional democracy. *UCLA L. Rev.*, **65**, 78.

Kanthak, K., and J. Woon, 2015: Women don't run? election aversion and candidate entry. *American Journal of Political Science*, **59 (3)**, 595–612.

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

Landa, D., and D. Duell, 2015: Social identity and electoral accountability. *American Journal of Political Science*, **59 (3)**, 671–89.

Lewis, D. E., 2011: Presidential appointments and personnel. *Annual Review of Political Science*, **14**, 47–66.

Loury, G. C., 1976: A dynamic theory of racial income differences, northwestern University, Center for Mathematical Studies in Economics and Management Science.

Miller, G. J., 2005: The political evolution of principal-agent models. *Annu. Rev. Polit. Sci.*, **8**, 203–225.

Neumark, D., and M. McLennan, 1995: Sex discrimination and women's labor market outcomes. *Journal of human resources*, 713–740.

Niederle, M., and L. Vesterlund, 2007: Do women shy away from competition? do men compete too much? *The Quarterly Journal of Economics*, **122 (3)**, 1067–1101.

Persico, N., 2002: Racial profiling, fairness, and effectiveness of policing. *The American Economic Review*, **92 (5)**, 1472–1497.

Persico, N., 2009: Racial profiling? detecting bias using statistical evidence. *Annu. Rev. Econ.*, **1 (1)**, 229–254.

Persson, T., and G. Tabellini, 2000: *Political Economics: Explaining Economic Policy.* Cambridge: MIT Press.

Pettigrew, T., 1979: The ultimate attribution error: Extending allport's cognitive analysis of prejudice. *Personality and Social Psychology Bulletin*, **5 (4)**, 461–76.

Phelps, E. S., 1972: The statistical theory of racism and sexism. *The american economic review*, **62 (4)**, 659–661.

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

Swain, C., 1993: *Black Faces, Black Interests: The Representation of African Americans in Congress.* Cambridge: Harvard University Press.

Tajfel, H., and M. Billig, 1974: Familiarity and categorization in intergroup behavior. *Journal of Experimental Social Psychology*, **10**, 159–70.

Tajfel, H., and J. Turner, 1986: The social identity theory of intergroup behavior. *The Psychology of Intergroup Relations*, S. Worchel, and W. Austin, Eds., Chicago: Nelson-Hall, 7–24.

Ting, M. M., 2002: A theory of jurisdictional assignments in bureaucracies. *American Journal of Political Science*, 364–378.

Ting, M. M., 2011: Organizational capacity. *The Journal of Law, Economics, & Organization*, **27 (2)**, 245–271.

Western, B., and B. Pettit, 2005: Black-white wage inequality, employment rates, and incarceration. *American Journal of Sociology*, **111 (2)**, 553–578.

Wright, E. O., J. Baxter, and G. E. Birkelund, 1995: The gender gap in workplace authority: A cross-national study. *American sociological review*, 407–435.