Determinants of Discrimination in Strategic Settings[*]

Dominik Duell[†] and Dimitri Landa[‡]

Draft, October 7, 2014

Abstract

In a laboratory experiment in a strategic principal-agent setting, we analyze how principals' elicited assessments of the causes of agent performance vary depending on whether they share the social identity with the agents. We isolate the effect on subjects' beliefs and choices of the strategic environment as such and provide a direct test of the strategic theory of statistical discrimination. We find that when principals use the sanctioning tools at their disposal in an outcome-contingent way, principals' and agents' choices sustain a pattern of beliefs that is observationally equivalent to "ultimate attribution error": upon observing good outcomes, principals attribute them more readily to their agents' effort and reward their agents more frequently when they share a social identity; and in turn, agents who share a social identity with their principals tend to invest more into effort in expectation of principals' reward choices. However, when principals do not use the sanctioning tools outcome-contingently, and when they do not have access to sanctioning tools, they do not hold such beliefs. This and other evidence we report suggests that in strategic settings, principals' prejudicial treatment of agents may be a product of the strategic nature of the environment, rooted in asymmetric but correct beliefs. Prejudice and discrimination may owe more to the strategic nature of the environment than previously recognized.

[*] The research presented in this paper was supported by NSF Grant #SES-1124265. Previous versions of this paper have been presented at the Political Economy Seminar at Columbia University, September 2014, New York, NY; Annual MPSA Conference, April 2013, Chicago, IL; EPSA General Conference, June 2013, Barcelona; Center of Experimental Social Science Colloquium, Nuffield College, Oxford; and North-American Economic Science Association Conference, October, 2013, Santa Cruz, CA.

[†] Research Fellow, Institute for Advanced Study in Toulouse

[‡] Associate Professor, Department of Politics, NYU

1 Introduction

Ethnic, racial, gender, and other forms of social identity influence how people inhabiting them respond to others and how others respond to them. At its most benign, this influence brings to the table special knowledge and difference that enriches the interactions, but more frequently, it gives rise to well-known empirical patterns that come to frame public and policy debates about social and political inequalities: the wage gap between men and women, and between whites and minorities (Altonji and Blank, 1999), the under-employment of blacks compared to whites (Chandra, 2000; Western and Pettit, 2005), the under-representation of women and minorities in legislative bodies of most Western democracies (Fox and Smith, 1998; Paxton, Kunovich and Hughes, 2007; Iversen and Rosenbluth, 2008; Griffin and Newman, 2007, 2008; Bird, Saalfeld and Wüst, 2010), and ethnic- and race-based voting in elections (Kaufmann, 2004; Chandra, 2004).

Consider the following example that encapsulates the insidious nature of discrimination is strategic principal-agent settings like those that underlie these empirical patterns. Alice works in a department managed by Bob, who has the power to recommend promotion for department employees and who will do so depending on his perception of their respective effort levels. However, Bob cannot observe effort levels directly and must base his judgment on his interpretation of the outcomes they individually generate – a noisy measure of the effort levels underlying them. Alice, who is pessimistic about her chances for promotion, is considering whether it is not wiser to re-allocate some of her time elsewhere. Bob, who suspects that Alice may be under-investing, is less likely to attribute a good outcome she may generate to effort and more to Alice's good luck. In effect, the outcome she needs to generate for promotion is higher than those other employees do. Realizing this, Alice is discouraged and chooses to invest less, confirming Bob's suspicions. Bob's interpretation of outcomes and Alice's expectation of a tougher standard are both correct and equilibrium-consistent with each other and the actions supporting them. But if the equilibrium is selected by and feeds into expectations of gender-specific behavior, they reinforce an identity-based discriminatory practice.

In psychology, the phenomenon of prejudicial judgment is grounded in a psychological disposition for *the ultimate attribution error* (Allport, 1954; Pettigrew, 1979; Tajfel, 1981; Kramer, 1994; Knippenberg, 2003). According to the logic of this bias, when observing good outcomes from actors with shared social identity – e.g., from Bob's male employees in gender-salient environments – individuals like Bob will be more inclined to attribute those outcomes to disposition, their fellow "in-group" members' "hard work;" in contrast, when those good outcomes come from "out-group" actors, like Alice, Bobs will be more inclined to associate success with a favorable circumstances and not with Alices' effort.

The settings of the employer-employee relationships and the voter-office candidates are, however, fundamentally strategic. When we observe asymmetric attribution in these settings, is it a consequence of a psychological disposition toward prejudice or of correct beliefs about difference in performance arising from strategic responses to the asymmetric beliefs and choices by others? We may justifiably infer discrimination when observing employers reward some of their employees but not others who performed similarly but differ in their social identity or group background or when observing voters' reluctance to support a more competent candidate from another social group. But saying that what underlies these phenomena is prejudice and measuring it without accounting for selection effects is problematic and can lead away from appropriate policy remedies. While economic theory of principal-agent relationships has understood that it does not take a psychologically driven misattribution to create and sustain stereotypes (Phelps, 1972; Arrow, 1973; Spence, 1973; Loury, 1976; Coate and Loury, 1993), there has been a gap between the recognition of the different contributors to discrimination and their empirical evaluation. As (Moro, 2009) puts it in a recent review: "the main problem is to find ways to identify, using available data, to what extent group differences are caused by prejudicial attitudes, or by asymmetric beliefs (self-confirming or otherwise) and incentives."

The primary aims of this paper is to address the challenge to identifying the sources of discrimination with a laboratory experiment designed to identify the distinctly strategic effects of individuals' responses to sharing social identity in a principal-agent environment – effects that are independent of asymmetric group-based generalizations, either rationally or psychologically sustained – and to investigate implications of social identities for agents' and principals' actions and beliefs that sustain discriminatory outcomes in strategic settings.

Our findings suggest that patterns of beliefs associated with ultimate attribution error may emerge as a fundamentally strategic phenomenon: the principals' asymmetric attribution can be rationalized by the properties of agents' choices, which, in turn, are rationalizable by the principals' choices. In particular: on average, principals set lower thresholds for rewarding the agents when the agents matched with them share membership in the same salient social identity group (in-group matches) than when the matched agents do not share principals' salient social identity (out-group matches); Conditional on setting lower threshold in in-group matches, principals develop (correct) beliefs that are observationally akin to the presence of the ultimate attribution error; agents who share a social identity with the principals tend to invest more into effort than those who have a different identity than their principals; And that higher investment tends to come from those agents in in-group matches who anticipate having to meet lower thresholds for receiving a reward from the principal, justifying the principals beliefs about determinants of agents' performance, conditional on principals' reward choices.

The differences in beliefs of principals playing distinct types of rewarding strategies and the contingency of principals' asymmetric beliefs on the access to sanctioning tools further reinforce the "strategic" interpretation of prejudice and discrimination in our data. More broadly, the analysis of the subject behavior we present suggests that the strategically induced attribution asymmetries may be *a*, if not *the*, first-order phenomenon when it comes to accounting for prejudicial choices by principals; thinking about the roots of prejudice in (purely) psychological terms may be missing the biggest culprit.

Our paper is organized as follows. Section 2 introduces the key elements of our identification strategy against the background of prior theoretical and empirical work that has sought to characterize and identify varieties of discrimination. Section 3 provides the description of our model of the principal-agent relationship. Section 4 describes the setup of the experiment. Section 5 and 5.4 presents empirical results. We discuss those results in the light of our aim to identify Arrovian (endogenous) statistical discrimination in Section 6 and conclude in Section 7.

2 Discrimination: Variety and Identification

2.1 The Different Effects of Social Identity

Discrimination refers to a situation in which persons who perform equally in a physical or material sense are treated unequally in a way that is related to an observable characteristic such as race, ethnicity, or gender.[1] Prejudice – understood as a key determinant of discrimination – is a faulty or inflexible generalization about members of a group (Allport, 1954). Unlike discrimination, which may be rationalizable with a set of potentially correct beliefs, prejudice, as defined above, necessarily entails a mistake.

We analyze discrimination and prejudice in the relationship of delegation found naturally in the contexts of political accountability and the labor-management relations. The corresponding political economy and labor economics literatures include explorations of patterns of discrimination, prejudice, and individuals' feedback to experienced and expected biases for or against them.

A recent wave of studies seeking to detect discrimination have focused on employer responses to auditions, audits, and resumes from candidates from different social groups. In an audition-type study, Goldin and Rouse (2000) demonstrate the existence of discrimination against female musicians auditioning for orchestras by comparing employers' behavior in blind auditions to behavior when gender is observable. Audit studies usually provide evidence for employment discrimination against black job candidates (Bendick, 2007). Bertrand and Mullainathan (2004) find discrimination against black candidates by comparing interview re-calls in response to resumes that vary the name of the applicant from stereotypical white names to black names, holding everything else constant. While these studies document discrimination or prejudice by controlling for selection effects and many confounding variables, they do not model the mechanism by which discrimination, prejudice, and feedback to experiencing the two interact; the effect measure may, thus, be a partial equilibrium one. This concern is moot in the context of the "outcome-based approach" to characterizing the possible presence of racial bias in policing assuming the possibility of strategic responses by the agents (Knowles, Persico and Todd, 2001; Persico, 2002; Persico and Todd, 2006; Persico, 2009; Coviello and Persico, 2013). In the studies developing this approach, the police is inferred to be biased against a particular social group when the "hit-rate" against the members of that group, e.g. the rate by which stopped individuals from that group are arrested, is significantly lower than the "hit-rate" for other groups. This approach offers a clever way of evaluating whether a policy or an institution creates "too much discrimination" in equilibrium relative to an egalitarian social

[1] See Altonji and Blank (1999); Holzer and Neumark (2000) for a more detailed elaboration of this definition in a labor market context. Note that as defined above, discrimination may or may not be rationally sustained. Whether it is is one of the central questions of this study. Throughout, when referring to discrimination, we have in mind a concept as defined above, and use qualifiers to indicate its correspondence to rational action, lack thereof, etc.

optimum while sidestepping the question of what drives discrimination. This way of proceeding is attractive when the goal is to inform the decision on whether to support or rescind a policy like stop-and-frisk, but understanding what drives discrimination in principal-agent settings in, say, labor market, is essential for choosing optimal policy in responding to it.

An influential theoretical approach to analyzing the determinants of discrimination views it as resulting from a *taste for discriminating* against out-group members (Becker, 1973; Akerlof and Kranton, 2000, 2010). 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 a prejudice or a primal affect: positive when interacting with someone with a shared identity and negative when the identity is unshared. A somewhat different version of this mechanism can be found in the social psychology work that ties prejudice, and discrimination to which it may give rise, to a predisposition to a particular kind of bias known as the *ultimate attribution error* (Pettigrew, 1979). This error manifests when individuals are biased in their attribution of outcomes to a disposition or a situation when judging in-group members in contrast to out-group members. In particular, the claim is that individuals will attribute what they perceive to be a negative outcome from an out-group member more to disposition than they would a similar outcome from in-group members; in contrast, when they observe what they perceive to be a positive outcome from an out-group member, they will regard such an event as more likely a consequence of luck or special situation than they would if they saw the same outcome from a fellow in-group member. Given the connection between the ultimate attribution error and prejudice, it should be unsurprising that the attribution error manifests itself most strongly when group membership is more salient, or when there is a history of tense intergroup conflict (Hewstone, 1990).[2]

A profile of principals' beliefs consistent with a possibility of ultimate attribution error may be consistent with the choices made by agents in the strategic environment (see more on that below). In that sense, beliefs that look like prejudice may come from prejudice but they also may not. However, the absence of significant belief asymmetries on the part of the principal (or, in a stronger claim: the absence of the particular asymmetries described by the ultimate attribution error) is an indication that there is no significant evidence of the taste for discrimination. In the experiment we

[2] Landmark experimental psychological studies of in-group favoritism and discrimination include (Billig and Tajfel, 1973; Turner and Brown, 1978; Vaughan, Tajfel and Williams, 1981; Diehl, 1988; Klein and Azzi, 2001).

report, we elicit motivated beliefs from the principals to gage how they correspond to the ultimate attribution error as a necessary behavioral condition for prejudice.

The possibility of asymmetric beliefs and discrimination being consistent with the "facts on the ground" underlies a different theoretical approach to accounting for discrimination, known as *statistical discrimination.* This approach expects discrimination to occur "when rational, information-seeking decision makers use aggregate group characteristics to evaluate relevant personal characteristics of the individual with whom they interact" (Moro, 2009, 1). Statistical discrimination does not presuppose a prejudice or, indeed, any kind of unreflected psychological affect; it is grounded entirely in a rational inference. In an early paper raising the possibility of this kind of discrimination, Phelps (1972) ties it to exogenous variation in the relevant statistics of the demographic populations, which could reflect their distinct histories, experiences, etc. Arrow (1973) endogenizes group differences and argues that asymmetric beliefs about members of different groups can be self-confirming even when those groups are identical ex-ante.

A number of studies, including laboratory experiments, report evidence consistent with statistical discrimination. Anderson and Haupert (1999) exogenously assign productivity to artificial candidates that subjects in the role of employers may hire. They find statistical discrimination in an environment in which employers know group membership of the individual job candidate and average productivity of the group. Falk and Zehnder (2007) elicit differences in trust across social groups (populations of different districts of Zurich) and find that subjects' investments decisions in a trust game are rooted in subjects' knowledge about the social structure of the respective districts.

The idea of discrimination as a specifically strategic equilibrium phenomenon – the Arrovian version of statistical discrimination – has informed a considerable body of theoretical and empirical work, some speaking directly to the debates regarding the desirability of policy interventions such as affirmative action programs. Such policy interventions can induce differences in employers' beliefs about effort exerted by members of different social identity groups and result in discrimination, which, in turn, reduces incentives for members of the disadvantaged group to invest and creating a self-fulfilling prophecy (Loury, 1976; Coate and Loury, 1993). In this sense, employers allocating members of the on average higher-qualified social group to higher-skill jobs (Lundberg and Startz, 1983; Lundberg, 1991) may be making ex ante optimal economic decisions. An opposite conclusion associated with strategically induced asymmetric beliefs about behavior of agents from different social identity groups has been suggested in the context of electoral representation, where expected discrimination by voters is linked to representatives' effort on behalf of their constituents. In that context, an argument that has received some political traction holds that voters should, all else equal, prefer a candidate with unshared identity because she will work to earn the electoral support that a candidate with a shared identity will take for granted (Swain, 1993; Landa and Duell, 2014).

The Arrovian theory's prediction of the supply-side labor market behavior has also received some empirical support. Pre-labor market discrimination has been shown to affect human capital investment of future generations and, in so doing, arguably solidify segregation (Coate and Loury, 1993; Benabou, 1996; Bowles, Loury and Sethi, 2009). Other studies have shown that women who reported discrimination in the work place are subsequently more likely to change employer, have children, and marry (Neumark and McLennan, 1995).[3]

While these studies go some distance in separating the taste for discrimination from the statistical discrimination mechanisms, they do not offer a clean test of the Arrovian strategic theory of statistical discrimination. The Arrovian theory is, in the first place, an explanation of what drives discrimination by the principals. Indeed, even the agents' strategic responses to being discriminated are consistent with the possibility that what drives principals' discriminatory choices is psychological, taste-for-discrimination, factors that may have little to do with statistical discrimination, let alone an Arrovian version of it.

Evaluating the prediction of the principals' (employers/managers, voters) strategic response is trickier than that of the agents because such a response needs to be distinguished from the response to the differences in the population statistics that typically form an empirical background of the specific principal-agent interaction analyzed. When empirically evaluating the behavior of principals, the predictions of the Phelpsian and the Arrovian versions of statistical discrimination are particularly hard to tell apart.

Two laboratory studies reporting results that speak to this distinction are particularly relevant to the present analysis. Fershtman and Gneezy (2001) provide some evidence of differences in attribution in interactions with a strategic component and the interactions without it. They pair Israelis with either Ashkenazic or "Eastern" names and report increased mistrust of men with

[3] (Niederle and Vesterlund, 2007) find that women are less likely to select into competitive environments and also less likely to believe that they meet the criteria to qualify for public office; the number of women running for office trails far behind the number of men (Fox and Lawless, 2010, 2011).

"Eastern" origin in a trust game (a game with a strategic component), but no effect of ethnicity in a dictator game (which has no strategic component). The subjects' inferences from the stereotypes are, though, mistaken: groups in their trust game do not differ in how much they return to the sender. Fershtman and Gneezy argue that these results indicate that subject behavior is driven by ethnic stereotypes but not by the taste for discrimination, but also that subjects' incorrect beliefs imply that the discrimination is not statistical, even if it can be rationalized by the mistaken beliefs. The contrast between their results and our findings that strategic discrimination by principals exist and is based on correct beliefs about agents' effort choices is instructive and suggests boundaries for the scope of strategically driven statistical discrimination. In the trust game, beliefs about group-differences in performance can affect play only at the moment when the sender makes her allocation decision. Because the game does not provide for a strategic feedback after the receiver's choice, the receiver has no affirmative reason to act on the stereotypes, whether senders' or her own – at least not in the artificial environment of the laboratory. In contrast, the principal-agent interaction in our experiment allows for such a strategic feedback after agents' effort choices, when principals evaluate the observed outcome and decide whether to award a bonus. This gives agents in our experiment an incentive to follow their beliefs about the stereotypes held by principals – an incentives that does not exist for the receivers in the canonical trust game.

Fryer, Goeree and Holt (2005) simulate both principals' hiring decisions and agents' choices whether to invest into education. They find that principals discriminate against the group of agents that was randomly but publicly chosen to be disadvantaged (having higher costs of investment), and that discrimination persists, albeit, non-linearly and non-monotonically, even after the differential in investment costs is removed. While the initial discrimination rests on exogenously given differences between groups, subjects' behavior after the removal of the cost differential implicates effects that go beyond the Phelpsian version of statistical discrimination – including, potentially, a strategic equilibrium effect sustained by the persistence of principals' beliefs about agents' group-level behavioral differences. The difficulties of accounting for the behavioral variation in these effects in the data point, though, to a potentially complex, and hard to parse, mix of underlying factors, including stickiness of subjects' action choices over the session, which is common in lab settings and has little to do with discrimination. This suggests the value of eliciting principals' beliefs to help account for their choices and of avoiding the protocols that build-in history-dependence – conclusions that motivate some of the key elements of our experimental design. Another distinction between our setup and the setup in Fryer, Goeree and Holt (2005) is the sources of discrimination analyzed. Because in their study the principals are not endowed with distinct social identities, the discrimination observed in principals' choices is not driven by their attitudes to identity-driven relationships. Because we are interested in analyzing the consequences of discrimination driven by those relationships, we endow all subjects in our experiment, principals and agents, with social identities.

These distinctions notwithstanding, the methodological insights of the artificial identity construction that is not directly correlated with material differences in players' utilities (in stage 2 of the treatment in Fryer, Goeree and Holt (2005)) and the comparison between settings with strategic and without strategic choice (Fershtman and Gneezy, 2001) are important elements of our own strategy for identifying strategic discrimination, as described in the next section.

2.2 Identifying Arrovian Statistical Discrimination in a Principal-Agent Environment

Our goal is to understand how subjects respond to group identities in principal-agent settings. In this subsection, we draw attention to some of the elements of our experimental design as it relates to the identification of the determinants of subjects' responses to group identities. As will be seen below, some of these elements are independent of and others directly related to the details of the particular principal-agent setting in which the instantiated interaction between subjects takes place.

To separate strategically driven belief asymmetries from the non-strategic (Phelpsian) statistical belief asymmetries, we adopt a design that avoids pre-treating subjects with different reputations of different 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. Both of these elements have the effect of freeing individual decision-making from history dependence. Our analysis of the dependence of subjects' choices on the history of interactions in the course of the sessions in the data confirms that this is carried through the experiment. The primary analysis, then, focuses on the presence and the principals' responsiveness to the existence of strategic incentives for asymmetric treatment of groups.

Our main measure of principals' beliefs about agents corresponds directly to the measurement of the ultimate attribution error. We elicit principals' incentivized explanations of observed outcomes associated with their agents' choices. When the principal is uncertain about the primary determinants of the outcome – agent's underlying type vs. agent's motivated effort – her explanation may be grounded not only in psychologically driven bias (prejudice) but also in expectations of strategic responses by the agent. This observation underscores two main challenges to interpreting principals' beliefs: (1) strategic responses from agents to principals' perceived bias may undermine the inferences about prejudice from the observables; (2) when agent's type is payoff-relevant for the principal, the principal's strategic incentives may be to induce type-revelation by the agent, which, in turn, could influence principal's beliefs' in a way that confounds the revelation of the attribution error.

To appreciate the first challenge, suppose the principal engages in causal attribution and upon observing a good outcome assigns higher probability to the event that an agent who shares a group identity with the principal has exerted a larger amount of effort than he actually had. Agents may anticipate such a belief and compensate by either reducing effort when in the in-group match or increasing effort when in the out-group match whenever there is a strategic incentive to please the principal. In this case, assuming that principals' beliefs about agents' strategic choices are broadly correct, attribution of symmetric weights to effort behind the observed outcomes in in-group and in out-group matches may obscure the presence of the attribution error. But, there is another possibility that leads to agents' choices that are the opposite of those that would create this challenge.

This possibility, more consistent with our data, is that agents in in-group matches may increase their effort if they perceive the marginal return on their effort as to increase above the marginal cost. In this case, the asymmetric attribution may be justified, in part, by the strategic response by the agents, but treating such an asymmetry as the clean measure of the attribution error would overstate the latter. It is clear that agents' strategic responses to the attribution error by the principals complicate the inferences from the observables.

To appreciate the second challenge, note that if the elicitation of principals' beliefs and principals' sanctioning decision are linked (and principals' beliefs are payoff-motivated), then principals may rationally use their reward/punishment instruments to effect a type separation in equilibrium either in order to ensure that they can identify the agent's type to maximize the post-election payoff (as in the classic adverse selection argument) and/or in order to increase precision of the beliefs about the causes of performance, if the revelation of such beliefs is incentivized. If the danger associated with the first challenge is that of over-stating the extent of psychological bias, failing to appreciate its strategic dimension, then the danger associated with the second is the opposite: principals' beliefs may be driven by identity-free strategic play "too much," potentially overshadowing the elicitation procedure and obscuring whatever bias the principal may have.

Our experiment models the interaction between agents and principals in which principals observe a noisy signal of outcomes but not their agents' type or effort and varies the existence and matching of the artificially induced group identities. To address the challenges described above, in the context of this setup, we make the following choices. First, we tie principal's payoffs to her beliefs about the realization of agent's underlying type vs. effort but not to the principal's decision whether or not to reward or punish the agent,[4] and elicit those beliefs without tying the elicitation procedure to principals' sanctioning tools. Second, we create incentives for the agents that lead to pooling effort choices – low-quality types invest highly into effort, high-quality types invest little.

Finally, as indicated above, the size of the difference in principals' beliefs between in- and out-group matches should not be thought of as a precise psychological measure of the attribution error because it is magnified by strategic incentives. To get a better handle on the size of this effect, we (1) create counterfactual environments that make it possible to separate the strategic effects of attribution asymmetries from the effects that are not, in part, driven by the expectations of strategic responses from the agents (i.e., from the psychologically driven effects). We do so by distinguishing the principals whose punishment/reward strategies are constant in outcome from those principals whose punishment/reward decisions vary with the observed outcomes and compare their attributions, and (2) comparing the principals' beliefs in the strategic treatment to those in the corresponding non-strategic treatment.

[4] In contrast, rewarding the agent with a second period in office, for instance, would ultimately affect principal's payoffs in that period.

3 A simple model of principal-agent relationships

3.1 Set-up

A principal faces an agent with privately known competence, modeled as her type $t \in \{1, 2, 3\}$; the principal's commonly known prior is assumed uniform on that support. The agent chooses her effort level, $e \in \{1, 2, 3\}$, that is costly to herself. The principal then observes the outcome F given by

$$F = t + e + \omega \, ,$$

where ω, a noise, is a random draw from a uniform distribution on $\{-1, 0, 1\}$. Having received this information, the principal decides whether to award the bonus, b, to the agent. The representative's payoff is given by

$$u_r = G(F, b, e),$$

where the function G is increasing in F and b and decreasing in e:

$$G(F, b, e) = \begin{cases} \beta\sqrt{F{+}1} \; - \alpha e \text{ if the bonus is awarded} \\ \beta\sqrt{F} \; - \alpha e \text{ if the bonus is not awarded.} \end{cases}$$

The principal also chooses whether she wants to double t or e (both unobserved). The principal's payoff, then, is computed accordingly as

$$F + De + (1 - D)t,$$

where $D = 1$ is the principal's decision to double e. The game ends when these payoffs are realized.

3.2 Equilibria and behavioral expectations

There are many Perfect Bayesian Equilibria of this game. We describe the classes of equilibria for the values of $\alpha = 1.95$ and $\beta = 6$ that are set in the experiment. In the equilibria with the highest expected welfare for the principal, the principal chooses to award bonus if and only if $F \geq z$, $z \in \{3, 4, 5\}$, and the agent chooses level of effort e^* such that $e^* {+} t = 4$. These are pooling equilibria, and the principal's beliefs in these equilibria are such that she is indifferent between choosing to double e or t. In these equilibria, 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. Note that there is no feasible response by the principal such that the agent of a given type would benefit from investing more than at these corresponding levels, i.e., no higher level of effort is rationalizable.

One can construct equilibria in which the threshold for receiving bonus is $z \in \{0, 1, 2, 6, 7\}$. Those equilibria are semi-separating, in that the principal's posterior beliefs about the agent's type will not be uniform, and there will be a critical value in the $\hat{F}$ space such that the principal will prefer to double type for $F > \hat{F}$ and effort for $F < \hat{F}$. We will refer to these semi-separating equilibria and to the pooling equilibria described above as the outcome-contingent-play (OCP) equilibria.

In a different kind of equilibrium, with outcome-noncontingent-play (ONCP), the principal awards the bonus independently of outcome and the agents choose minimal levels of effort, inducing partial separation through outcomes. Here, the principal will always prefer to double type. Note that, given the payoff function, the principal will always prefer the pooling OCP – equilibria with highest expected outcomes – to the equilibria with semi-separation, whether those are OCP or ONCP equilibria. That is, given the payoff structure, the principal always prefers to obtain highest possible equilibrium outcome and live with ex post less correct attribution to playing an equilibrium in which it is easier to make a correct attribution but at the cost of a low expected equilibrium outcome.

Note that the traditional intuition for the highest expected principal's welfare equilibria – the as-if pre-play announcement of the reward/punishment rule by the principal who can freely commit to implementing it at the moment of accountability (Persson and Tabellini, 2000; de Mesquita and Landa, 2013) – makes those OCP equilibria most plausible as predictors of play in which the principal is, in fact, seeking to effect a higher outcomes with her punishment/reward instruments. We will return to this point below, when analyzing subject-level behavior in the experiment.

The environment described in our baseline game is "identity-free." When we prime and reveal to group members their social identities in the identity environment, we do so without altering this payoff structure. One equilibrium behavioral expectation is, thus, that identity has no effect on behavior. However, because players observe social identity matches and there are multiple identity-free equilibria, the game with the identity treatment also admits "meta" equilibria in which different equilibrium profiles are played in different identity matches (e.g., an OCP equilibrium profile with higher (lower) threshold for rewarding in in-group matches and an OCP equilibrium profile with lower (higher) threshold for rewarding in out-group matches). In this way, identity matches could matter as "switchers" between different equilibrium profiles.

4 Research design

4.1 Experimental design

The experiment models the interaction between a principal and an agent and elicits subject behavior by motivating reasoning with performance-based payments. We implement three treatment conditions with a total of 166 subjects and 3320 subject-round observations that simulate variants of the modeled interaction between principals and agents (*Principal-agent game*) presented in section 3. Additionally, in some of the treatments, at the beginning of each session, subjects self-select into artificial group identities (*Group identity inducement*).[5] We ran five sessions of our *identity-strategic treatment* (for ease of representation we will reference this treatment as the *main treatment*) that implements a principal-agent game closely following the model laid out previously with prior inducement of group identities.

Additionally, we implemented two supporting treatments. The first of these is referred to below as the *non-identity treatment.* It implements the baseline principal-agent game without inducement of group identities. This treatment helps us identify the effects of salient group membership controlling for the ID-invariant strategic play.

The second supporting treatment is referred to as the *non-strategic treatment.* Similar to the main treatment, it induces group identities but replaces principals' sanctioning tool with exogenously given incentives to the agents. Comparing behavior in this treatment against that in the main treatment aims to uncover the influence of a non-strategic environment in contrast to a strategic principal-agent interaction on attribution behavior. In this treatment, agents' payoffs are given by $u_r = G(F, e)$ where the function G is increasing in F and decreasing in e:

$$G(F, e) = \beta\sqrt{F} - e,$$

[5] At the beginning of each experimental session, we elicit risk-attitudes in a non-hypothetical, small stakes setting following the design presented by Holt and Laury (2002); since we elicit risk preferences in each session and treatment condition in the experiment, treatment effects should not be affected.

with $\beta = 4$ and F and e representing the outcome and the level of effort, respectively. The functional form of the payoffs and the parametrization were chosen so as to be as close as possible to those in the main treatment and to induce the optimal choices for agents, conditional on their type, that are identical to the optimal choices in the maximal principal welfare (3-4-5 threshold) outcome-contingent play equilibria in the main treatment game. In particular, in this game, 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. The incentives, thus, are set so as to create observational equivalence between agent choices in the main and in the non-strategic treatments and allow for the controlled comparison of principals' choices in these treatments to pinpoint the effect of strategic expectations.

Table 1 illustrates the manipulations that are varied across treatments.

Table 1: Components of experimental treatments

	Main treatment	Non-identity	Non-strategic
With sanctioning	✓	✓	
Without sanctioning			✓
Identity-inducement	✓		✓
Number of subjects	110	38	40
Number of observations	2200	760	800

After completing the description of the experimental treatments in the remainder of this section, we present our analysis of the average treatment effects in the main and the non-ID treatments in Section 5. We present results in the context of developing a comparison to the two counterfactual environments illustrated when we laid out our identification strategy. First, we parse behavior of principals whose punishment/reward strategies are constant in outcome from those principals whose punishment/reward decisions vary with the observed outcomes and compare their attributions in the strategic treatments (main treatment and non-identity treatment; Section 5.2). Then, second, we compare principals' beliefs in the main treatment to those in the non-strategic treatment (Section 5.4).

Sessions were carried out at the Center for Experimental Social Sciences/NYU. Each experimental session lasted 20 rounds with 14-20 participating subjects. Participants signed up via a web-based recruitment system that draws on a large, pre-existing pool of potential subjects. Subjects were not recruited from the authors' courses. The recruitment system contains a filter that blocked subjects from participating in more than one session of a given experiment. The subject pool consists almost entirely of undergraduates from around the university. Subjects interacted anonymously via networked computers. The experiments were programmed and conducted with the software z-Tree (Fischbacher, 2007).

After giving informed consent according to standard human subjects protocols, subjects received written instructions that were subsequently read aloud in order to promote understanding and induce common knowledge of the experimental protocol. No deception was employed at any point in the experiment, in accordance with the long-standing norms of the lab in which the experiment was carried out. Before the principal-agent game stage commenced, subjects were asked three questions concerning their understanding of the payoff tables provided to them in the instructions. 90% of participating subjects answered those questions correctly. At the end of the experiment, an exit survey was conducted. 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 round.

4.1.1 Group identity inducement

At the beginning of each session of the main and the non-strategic treatments, subjects were shown 5 pairings of paintings, one from Paul Klee and one from Vassily Kandinsky, and were asked which one they prefer. Based on which painter a subject preferred most of the time, he/she was assigned to be a *Klee* or a *Kandinsky*.[6]

Once identities were assigned, subjects participated in an activity aimed at strengthening the attachment to the new identities. In particular, they were given a quiz in which they were asked to identify the painter (Klee or Kandinsky) of five further paintings. In answering the question about each of those paintings, subjects gave initial guesses which were made available to other subjects in

[6] See Tajfel and Billig (1974); Chen and Li (2009); Landa and Duell (2014) for the use of painter-preferences to induce identities in Social Psychology, Economics, and Political Science.

the same identity group before everyone was asked for their final answer. Subjects within a group received $1 if the majority of members of their group named the correct painter in the final answer. Additionally, they received another $1 when members of their group gave at least as many correct final answers on all five quizzes as members of the other group (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.

Considerable experimental literature using the minimal group paradigm has shown its effectiveness in inducing the patters of responses to identity, including in-group favoritism and inter-group competition (Tajfel and Turner, 1986), that resemble those usually observed outside the laboratory with naturally occurring group identities. In particular, utilizing the minimal group paradigm approach, our experiment allows for subjects' inclinations to behave in a group-biased way to reveal themselves but also makes it possible for them consciously to control and correct their biases (something that is more difficult to do with subtle identity priming). Inducing identities in a way that enables the analysis of these real-world phenomena (Nosek et al., 2007) makes it possible to draw rich inferences about the different effects of social identities in a principal-agent relation between voters and representatives. The experimental literature provides evidence that "weak" induced identities significantly affect subject behavior with respect to individual shirking and free-riding (Eckel and Grossman, 2005), cooperation, and willingness to reward or punish (Chen and Li, 2009; Goette, Huffman and Meier, 2006; Bernhard, Fehr and Fischbacher, 2006; McLeish and Oxoby, 2007). Behavior driven by weak identities, thus, can be seen to approximate behavior that we would generally ascribe to the effects of strong, more contextualized or previously existing, group identifications. Eckel and Grossman (2005) and Goette, Huffman and Meier (2012) provide evidence that the effects of identity being induced are monotone in the strength of that identity, i.e., the weakness of identity inducement does not bias results in the wrong direction. Note that effect of artificially induced weak identities increases with salience (Eckel and Grossman, 2005; Charness, Rigotti and Rustichini, 2007; Chen and Chen, 2011); operationally, a key factor that raises such salience is interactions with fellow group members in performing joint tasks, such as group quizzes we administer as a part of our identity treatment.

Social identities enter our subjects' reasoning in principal-agent matches in a way that is not driven by the explicit payoffs assigned to them. As indicated above, this allows us to elicit effects of identity, including subjects' responses to identity, without "feeding" them to the subjects by directly and exogenously tying identity biases to their payoffs.

4.1.2 Principal-agent game

At the beginning of each round of the experiment, subjects are assigned to the role of either an *agent* or a *principal* and matched into pairs of one agent and one principal. They are randomly re-matched into pairs at the beginning of the next round but keep their role assignments throughout the course of the experiment.

The structure of our experiment approximates the principal-agent relationship between a voter and a representative or an employer and an employee in the laboratory. By monetarily incentivizing subjects in the role of agents, we create concerns about outcomes because agents value not being negatively sanctioned by principals. Subjects in the role of principals benefit from high outcomes and thus may want to incentivize agents to exert high investment into effort. These elements of the experiment represent core aspects of the empirical relationships between principals and their agents.

The game simulated in the lab mirrors exactly the structure and payoffs as laid out in Section 3. The sequence of moves is as follows: Agents are randomly assigned a *type* and informed privately about that draw. After that the agent makes a choice of the level of *effort* he wants to exert, and *noise* as well as *outcome* are realized. That outcome is then shown to the principal. The principal next chooses whether she wants to double the contribution of agent's type or effort to the outcome,[7] and whether to give the agent a bonus. The outcome is, then, augmented in accordance with the principal's choice and the round payoffs are realized.

In communicating the game to the subjects we referred to type as "Special Number," to noise as "Random Bump," to outcome as the "Choice Outcome", to subjects in the role of agents as "Player 1," and to subjects in the role of principals as "Player 2"; the value generated by principal's

[7] Given that we do not interpret the absolute magnitude of the the beliefs inferred from these choices but only their relative values in different treatments, not having an option to hedge does not bias our results. Allowing subjects to opt for a non-choice would likely miss such subtle considerations as attribution in the artificial environment of the lab.

decision whether to double type or effort in the outcome-function was termed "Increased Outcome." Subjects did not see agent's payoff function but received a table of all possible payoffs given type, effort, and noise, and principal's bonus decision, and in the instructions were told:

> "When you are participating in the role of Player 1, your payoff in a given round will depend on the *choice outcome* in that round (and so indirectly, on your *special number*, your *effort* level, and the realized *random bump*) but also directly on the chosen level of *effort* and on the decision of Player 2 you are matched with whether to give you a *bonus*."

Additionally, all subjects, those in the role of an agent and those in the role of a principal, were instructed that agents would be given payoff information on the screen whenever they are making their choice of effort, with the corresponding screen shot included in the instructions, along with examples of how payoffs are computed and how the provided payoff table is best utilized (see Appendix D). During the experiment, agents were asked on one of the screens,"What minimal outcome do you think Player 2 will demand to give you a bonus?" and shown all potential payoffs to them sorted by feasible choice of effort and probability of occurrence (dependent on the realization of noise).[8]

In summary, the sequence of moves in each round in this stage of the experiment is as follows:

1. Subjects are randomly assigned to the role of either agent or principal (termed "Player 1" and "Player 2", respectively).
2. Agents are assigned a type and privately informed about it's realization (1, 2, or 3).
3. Agents choose a level of effort (1, 2, or 3).
4. 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 (-1, 0, or 1).
5. Principals learn the value of outcome.
6. Principals choose whether to double type or effort and whether to give the agent a bonus (The latter choice is omitted in the non-strategic treatment)[9]

[8] This screen is part of the main treatment and the non-identity treatment but, of course, not of the non-strategic treatment.

[9] On the same screen, we also asked principals whether they thought type or effort was the higher quantity. The correlation between decision whether to double type or effort and the guess whether type or effort is higher is .74 (p=.00).

4.2 Variables and quantities of interest

Our primary interest is in three quantities: principals' *doubling decisions* (1=effort, 0=type) represents principals' motivated beliefs about the composition of outcomes; principals' *bonus decisions* (1=bonus, 0=no bonus), representing principals' decision whether to reward the agent; and agents' *effort* choices (1=low, 2=medium, 3=high). Further, we give agents the opportunity to compare payoff consequences of their choice options given which outcome they expect their matched principals will demand from them to award a bonus. From this comparison, we derive agents' beliefs about expectations towards their performance. Consistent with the convention introduced above, when principals and agents share a painter-group identity, we call it an *in-group match*, when they differ in their group identity we call it an *out-group match*.

Summary statistics for the variables of interested in identity inducement stage and principal agent game for all treatments are given in Appendix A.

5 Results

As noted above, we first focus our presentation of results on the main treatment and the non-identity treatment (Subsection 5.1 through 5.3). We consider average behavioral trends in relation to the equilibrium predictions in the first part of this section (Subsection 5.1). These findings, however, are subject to interpretive challenges that require the assessment of subjects' choices in reltion to their beliefs about their matches. To meet these challenges, we investigate the relationships between principals' and agents' actions and beliefs in Subsection 5.2. Finally, Subsection 5.3 demonstrates the robustness of behavioral patterns to history of play showing that subjects bring their motivation to treat in- and out-group matches differently into the laboratory. We discuss the non-strategic treatment in Section 5.4.

5.1 Average Behavioral Trends and Equilibrium Predictions

Considering subject-round observations, we find:

Result 1: average levels agents' effort are decreasing in agents' assigned type in all treatment conditions;

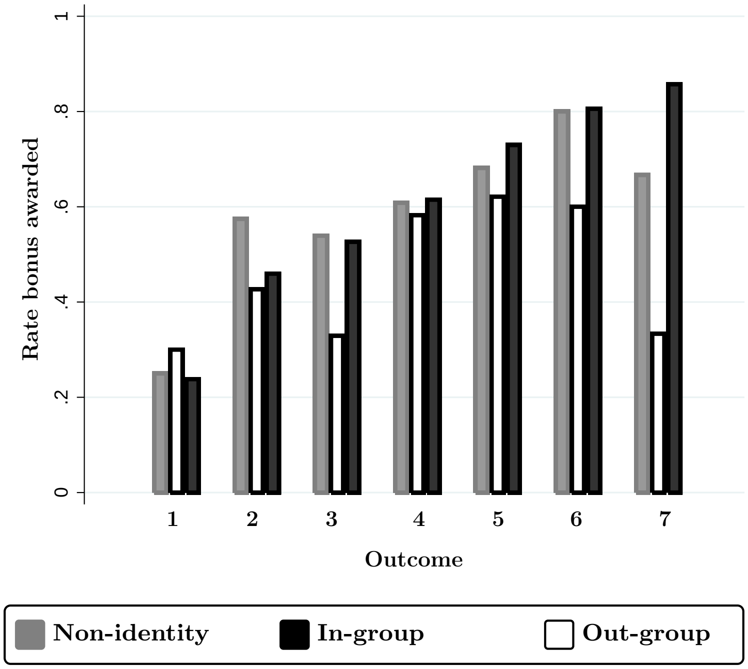
Result 2: average levels of agents' effort are not systematically different across treatment conditions;

Result 3: average rate at which bonus is awarded is higher in in-group matches of the main treatment than in out-group matches or in the non-identity treatment;

Result 4: average rate at which bonus is awarded is increasing in outcome in all treatment conditions.

Recall that in equilibria with the highest expected welfare for the principal were those in which the principal chooses to award bonus if and only the outcome is at least either 3, 4, or 5, and agents are incentivized to make effort choices that are decreasing in type. With respect to principals' reward-giving behavior, our average results show both that principals tend to choose incentivizing strategies: *on average, the bonus-giving decisions are increasing in the outcomes in all treatment conditions* (see Figure 1, Result 4), and that those strategies are identity-contingent: *the rate of bonus awarded in the main treatment is higher in the in-group than in the out-group matches, with the non-identity treatment splitting the difference* (Result 3). Differences between in- and out-group matches are substantial at outcomes lower than 4 (.53 vs. .38, difference of .15 (.10, .21)) and outcomes higher than 4 (.81 vs. .65, difference of .16 (.09, .27)), and not systematic at an outcome of 4 .05 $(-.5, .17)$. Those differences also result in significantly higher rates of granting a reward at high outcomes in in-group matches and lower rates at low outcomes in out-group matches than at the corresponding outcome levels in the non-identity treatment.

Figure 1: Rate at which principals award a bonus



Considering agents' choices, we find that average *effort is strictly decreasing in type in in- and out-group matches of the main treatment and in the non-identity treatment* (Result 1); the marginal effect of a unit increase in type is hereby a decrease in effort by .16 (.08,.25), .19 (.11,.27), and -.28 (-.18,.38), respectively – 95% confidence bounds are reported in parenthesis. This pooling behavior leads to a distribution of outcomes centered at 4 where 75% of observations in the main treatment and 76% in the non-identity treatment fall within the range of 3 and 5. However, despite the substantial identity-dependence of principals' average reward decisions, *agents' average effort levels do not differ systematically between in-group matches (1.79 (sd=.78)), out-group matches (1.76 (sd=.84)), and the non-identity treatment (1.74 (sd=.79)* Result 2). On its face, this result suggest that agents are not strategically responding to the identity-contingent asymmetric rewarding by the principals. But such a conclusion would be too hasty: to be convincing, we would want to anchor agents' choices in their beliefs about the principals. As it so happens, the average here is concealing a substantial variation in agents' beliefs and, in consequence, in their best responses to those beliefs. Indeed, as we will see below, in the subject-level analysis, those responses pull in opposite directions, and the lack of systematic differences in average agent choices is obscuring agents' robust, albeit different, systematic responses to what agents take to be principals' identity-contingent behavior.

A similar story can be told with respect to principals' attribution choices. We find that, on average, principals are more likely to attribute good performance to agents' effort when group identities exist than when they do not; the rate at which effort is doubled is .54 (.52,.57) in the main treatment but .46 (.41,.51) in the non-identity treatment. Interpreting these *average* doubling decisions is, however, somewhat difficult because the beliefs about effort they represent must depend on the incentives that the corresponding principal expects to be inducing with her bonus-rewarding strategy. To get at that properly, we need to look at the relationship between the principal's choices and beliefs at the subject-level.

5.2 Subject-level behavior and agents' beliefs

In section 5.1 we noted that in the aggregate, the data pattern suggests a divergence of principals' reward behavior in in- and out-group matches: principals in the in-group matches give a bonus more often. However, to get a precise sense of why principals are making such choices and to identify whether, conditional on reward decisions, there exists asymmetric attribution in our strategic principal-agent environment, we need to move to subject-level analysis and consider the rationalizability of observed behavioral profiles. We argue that differences in principals' behavior in in- and out-group matches result from diverging expectations on their part about their ability to incentivize agents' higher effort and from agents' responsiveness to their own expectations of the differences in principals' choices in in- and out-group matches.

In particular, we find that:

Result 4: Most principals use a bonus-rewarding strategy that incentivizes higher effort.

Result 5: Those principals who use such a strategy demand significantly higher outcomes for awarding the bonus in out-group matches than in in-group matches.

Result 6: Those principals who use such a strategy are significantly more likely to double effort than double type for higher outcomes in in-group matches and more likely to double type than double effort for higher outcomes in out-group matches.

Result 7: Conditional on principals not playing a bonus-rewarding strategy incentivizing higher effort, there are no appreciable differences in principals' attribution choices.

Result 8: The relationship between principals' bias in reward choices and the asymmetry in their attribution decisions is U-shaped: greatest and least relative leniency in reward decisions toward the in-group agents go with greater willingness to double effort for high outcomes in in-group than in out-group matches.

Result 9: While there is a variation in agents' expectations of bias in principals' reward decisions, agents tend to believe that they face systematically lower outcome demand for a bonus reward in in-group matches than in out-group matches.

Result 10: Agents' effort in in-group matches is increasing in their expectation of principals' in-group favoring bias in reward choices.

In this section, we first parse behavior that follows an outcome-contingent threshold strategy from choices that do not (Subsection 5.2.1). Then, we illustrate behavioral and attitudinal patterns that come with adopting such a strategy; Subsection 5.2.2 shows further that apparent discrimination and prejudice in favor of in-group matches can be accounted for by mutually reinforcing in-group favoritism based on correct asymmetric beliefs about each others choices by a sizable set of principals and agents. We also show that another set of principals holds asymmetric belief attributing good outcomes to effort more often in in-group matches than in out-group matches that go along with higher – and often fulfilled demands – for better outcomes from agents with whom they share a social identity.

5.2.1 Parsing rationalizable threshold-based strategies

We find that a substantial portion of principals employs a strategy best described as setting a threshold for observed outcome above which the agent is awarded a bonus (Result 4), i.e. strategy consistent with OCP equilibria described in Section 3.2. Identifying such thresholds provides a natural individual-specific definition of what outcomes a given principal perceives as good (at and above the threshold) vs. bad (below the threshold). To account for such behavior, we need to separate outcome-contingent from outcome-noncontingent play. We do so by separating principals into those who discriminate with their bonus decision by the outcome observed from those who always or never give a bonus and so are clearly not tying their bonus decisions to outcomes observed.

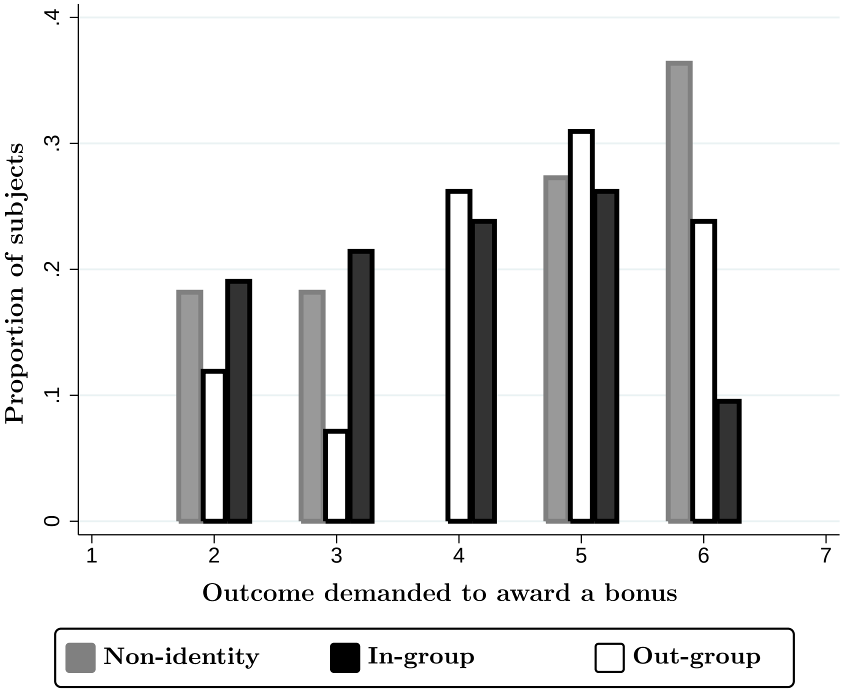
In our sample, 18 principals (7 in the non-identity and 11 in the main treatment) out of the 74 principals in the main and the non-identity treatments give a bonus in every round and 3 principals (1 in the non-identity and 2 in the main treatment) never give a bonus in any round. This leaves over two thirds of principals in our sample, 53 principals (42 in the main treatment and 11 in the non-identity treatment). For each of these principals, we compute her own threshold in the outcome-space as one that minimizes error in categorizing bonus decisions.[10] The inferred principal-specific thresholds vary from 2 to 7, with the average threshold being lowest (3.96) in in-group matches, higher in the non-identity treatment (4.45), and higher still in out-group matches (4.53). *We refer to these principals, whose rewarding choices are outcome-contingent, as "incentivizing principals," or as principals who are playing an incentivizing rewarding strategy* (Result 4).[11]

Figure 2 gives the full distribution of principals' thresholds.

[10] The average proportion of bonus decisions incorrectly classified with the error-minimizing threshold is .20 in the non-identity treatment, .23 in in-and out-group matches of the main treatment, and .22 in out-group matches. In Section B.2 of the Appendix we show that principals behave very consistently with their inferred individual threshold in deciding whether or not to award a bonus.

[11] A somewhat different approach to identifying such principals is to find those principals whose reward threshold, if known to the agent, would lead some type of agent to increase her effort level relative to her best response to a constant reward rule. All reward thresholds greater than 2 have this effect for agents of type 1, whose best response becomes to choose the effort level of 2. In our sample, there are 8 principals whose reward choices are outcome-contingent but whose inferred threshold is 2 or below. All comparisons of quantities presented in Section 5.2. are robust to excluding those principals as well as to excluding principals whose thresholds have higher than average error rates in categorizing bonus decisions.

Figure 2: Distribution of principals' outcome thresholds above which they are willing to assign a bonus

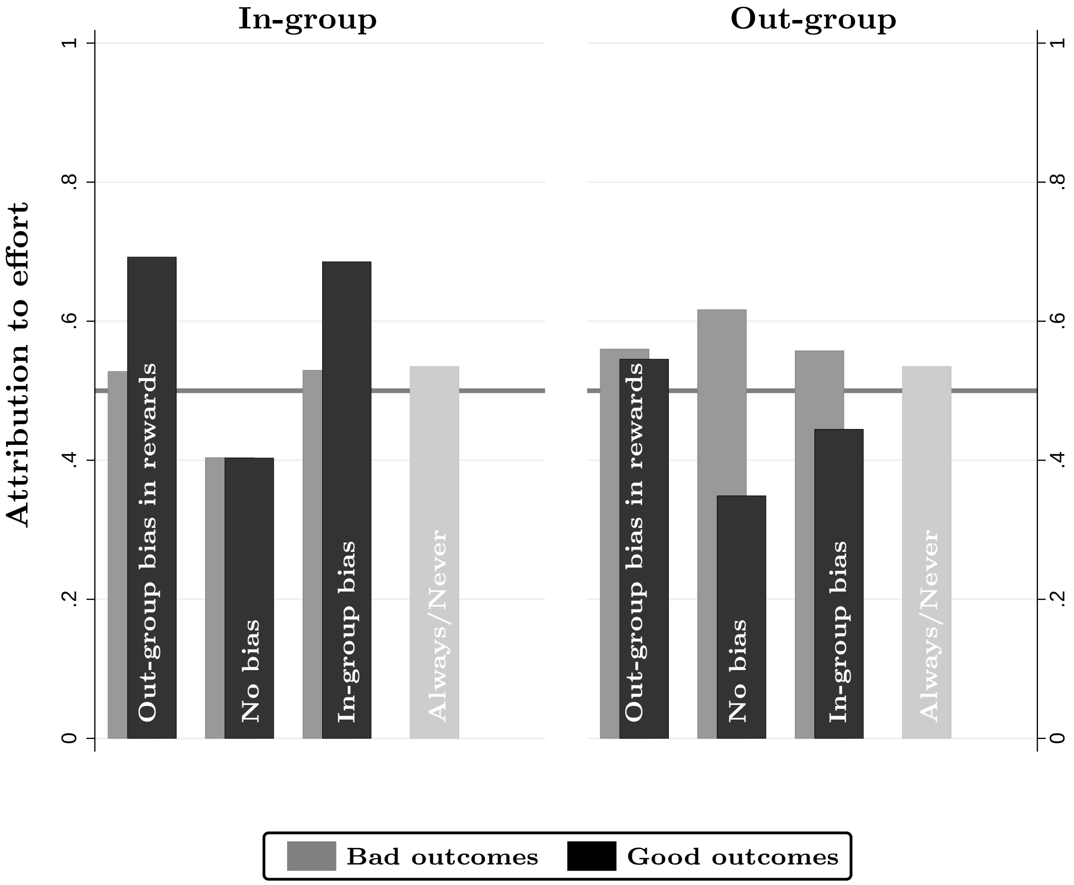


A casual glance at the figure makes clear that, relative to the in-group matches, the principal-specific thresholds in the out-group matches of the main treatment and in the non-identity treatment are shifted rightward. The difference-in-distribution test confirms this observation. In fact, *incentivizing principals in the in-group matches are less demanding than those in out-group matches ($p < .01$) and less demanding than the principals in the non-identity treatment ($p < .01$)* (Result 5). To award a bonus, 50% of incentivizing principals demand to observe lower outcomes associated with agents form their own group than those associated with out-group agents and only 20% ask agents in in-group matches for higher outcomes than those in out-group matches to award them a bonus. The remaining incentivizing principals do not differentiated between in- and out-group matches in their thresholds.

5.2.2 Motivational rationales behind asymmetric attribution

We next turn to describing the incentivizing principals' doubling choices relative to their inferred reward thresholds. This analysis allows us to get at the principals' equilibrium-specific beliefs about the agents since the notions of "good" and "bad" outcome thresholds are endogeneous to the strategy profile the principals believe themselves to be playing. In the discussion that follows, "good outcome" refers, thus, to the outcomes that are at or above the inferred reward threshold for a given principal in a given identity match (in or out), and "bad outcome" to the outcomes that are below that threshold. Figure 3 shows the rates at which the principals double effort relative to their bonus reward thresholds.

Figure 3: Rate at which principals double effort upon observing either *good outcomes* (outcomes above the individual threshold) or *bad outcomes* (outcomes below the individual threshold) in in- and out-group matches separated into principals that are more lenient towards in-group agents (*in-group bias in rewards*), principals that show *no bias* in reward decisions, and principals that are more demanding of in-group agents (*out-group bias in rewards*)



We find that principals playing the incentivizing bonus-rewarding strategy in the identity treatment, attribute good outcomes to effort more often in in-group than in out-group matches

– the difference in the rate at which effort is doubled is .15 $(.05, .28)$ – while they attribute bad outcomes to effort more often in out-group than in in-group matches – the difference is .09 $(.00, .19)$. The prevalence of attribution of good outcomes to effort in in-group matches and of attribution of bad outcomes to effort in out-group matches is nicely illustrated in Figure 3; the figure displays that higher levels of attribution to effort goes along with the observation of outcomes above the threshold in in-group matches (black bars are higher than gray bars for all levels of bias in rewards) but less so in out-group matches. Conversely, the observation of outcomes below the threshold in out-group matches leads principal to attribute effort in out-group matches (gray bars are higher than black bars) but less so in in-group matches. The asymmetric attribution, plays out more strongly at the high end of the outcome range – above the threshold – where the belief that effort is higher than type goes along with awarding a bonus.

Thus, *principals playing bonus-rewarding strategies are significantly more likely to attribute good outcomes to effort in in-group matches and more likely to attribute bad outcomes to effort in out-group matches* (Result 6), while *principals who are not playing an the incentivizing bonus-rewarding strategy show no difference in attribution* (Result 7). Principals who are always or never rewarding show no differences between attribution to effort in in- and out-group matches (.53 $(.48, .60)$ and .52 $(.45, .58)$, respectively).

Incentivizing principals in the non-identity treatment do not privilege type or effort as more likely explanations for outcome when that outcome is above their reward threshold (rate of effort doubled is .49); while they are somewhat less likely to attribute bad outcomes to effort (rate of effort doubled .39), with a difference of .10 $(-.04, .25)$ between those two values.

It is important to note that the in-group bias of attribution of good outcomes to effort is mainly driven by principals that demonstrate a bias in their reward decisions as well. Recall that a subset of incentivizing principals are more lenient towards in-group agents in their reward decisions and also demonstrate the largest in-group bias in the attribution of effort to good outcomes. The difference in rate at which effort is doubled in in- and out-group matches by those principals is .24 $(-.02, .38)$. Similarly, principals that are more demanding of in-group agents in their bonus decision are also largely in-group biased in their attribution of effort to good outcomes; the difference in rate at which those principals double effort upon observing good outcomes in in- and out-group matches is .15 $(.01, .28)$. These results establish that *there is a strong U-shaped relationship between* *subject-level bias in reward and asymmetry in attribution of good outcomes to effort* (Result 8).[12]

5.2.3 Agents' beliefs and discriminatory effort

Recall from our Result 1, that, on average, agents in the identity treatment invest slightly more into effort in in-group matches than in out-group matches but that this bias is far from obvious. However, we will show next that there is, in fact, systematic discrimination in effort choices in favor of in-group principals and that this behavioral pattern is a reaction to anticipated favorable reward decisions by the matched in-group principal. We will also show, however, that another sizable group of agents responds to such expectations of in-group favoritism on the part of the matched principals by decreasing their effort in in-group matches.

The setup of the experiment indirectly incentivizes agents monetarily to reveal their beliefs on principals' reward rules truthfully. Our analysis of these revealed beliefs shows that agents expect in-group principals to be more lenient in their award decisions and that, in turn, agents invest more when they share an identity with their principal than when they do not. Greater investment by the agents in in-group matches, thus, occurs as a response to the beliefs of facing a lower reward threshold.

Recall that before agents make their investment decision and 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?" Contingent on their answer and their type, they are given payoffs conditional on 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 at all.

71% of agents check at least one minimal outcome they expected to be demanded by their matched principals; the willingness to check stays constant throughout all 20 periods of the experiment. 22% of agents also investigate the payoff consequences of a second minimal outcome demanded and 13% a third value. In the modal case – in 26% of the agent-rounds – agents obtain information about payoffs for a minimally required outcome of 4, the next highest-frequency outcome value checked is 3 (22%). The distribution of checked outcomes is approximately normal,

[12] All comparisons of quantities presented are robust to excluding principals with very low thresholds (2 and below) or thresholds with high error rates in categorizing bonus decisions (.2 and larger).

centered around 4.[13]

Subjects in the role of an agent do not simply click through all potential outcomes. Most of them only check outcomes from the middle of the outcome range and tend to do so only once. If agents had clicked through all possible values of outcome, we would not be able to claim confidently they were checking the expected outcome that is most reasonable to them, given their match. Since agents are very specific in their expectation of the payoff information they want to obtain, and their behavior with respect to which expected outcome they check to obtain their potential payoffs does not change over the course of the experiment, their choices here indicate a targeted and reasoned attempt to learn payoffs at the expected outcome threshold. In short, agents' outcome-checking choices appear to elicit what they believe is the outcome principals are most likely to demand in order to award a bonus.

Figure 4: Distribution of agents' beliefs about in-group bias of principals

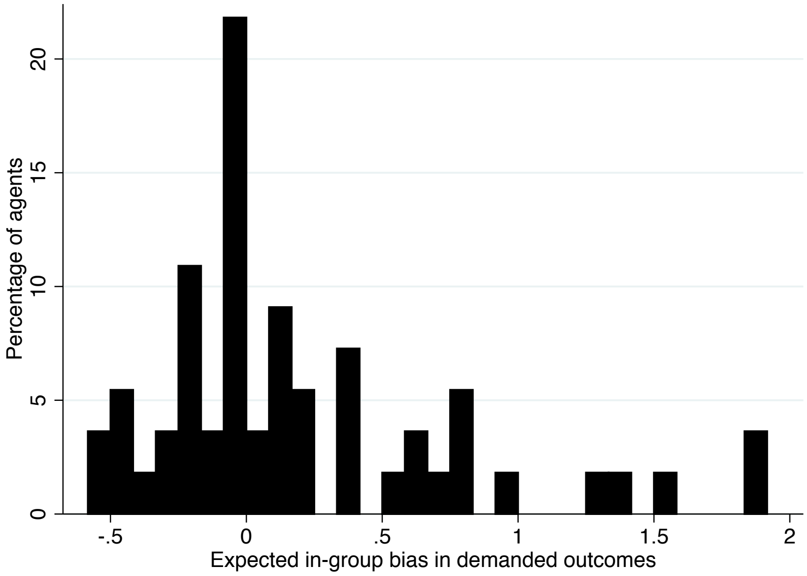


Figure 4 gives the distribution of agents' beliefs about principals' group-contingent reward biases constructed in this fashion.[14] As this figure shows, agents' beliefs about principals' biases are fundamentally asymmetric, tracking the direction of bias in principals' actual reward choices. More precisely *while there is a variation in agents' expectations of bias in principals' reward decisions,*

[13] Section A gives more data on frequency and extent of agents' use of this tool.

[14] We capture agents beliefs by recording the first expected demanded outcome they click. Defining this measure as average of all clicks by an agent does not change the results of our analysis.

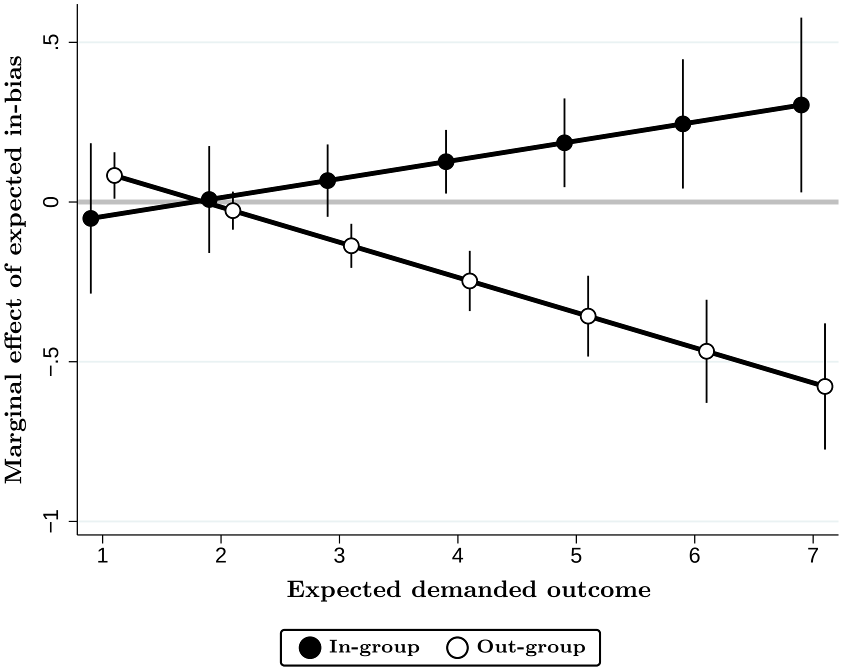
agents tend to believe that they face systematically lower outcome demand for a bonus reward in in-group matches than in out-group matches (Result 9).

We next turn to the consideration of the relationship between agents' beliefs and their effort choices. Recall from our Result 1 that agents' average levels of effort do not vary systematically between in-group and out-group matches. As we indicated above, this result conceals an important variation in agents' decision-making rationale. Agents who tend to believe that in-group principals are more lenient in their bonus reward choices make systematically different effort choices from agents who tend to believe the opposite. We find, in particular, that when agents believe that they are expected to deliver lower outcomes in in-group matches than in out-group matches to be awarded a bonus, their average effort in in-group matches is 1.87 but drops to 1.68 when they believe the demands are higher in in-group matches than in out-group matches (difference $= .19(-.12, .57)$). The corresponding effort levels in the out-group matches are 1.74 and 1.68 (difference $= .06(-.26, .37)$, respectively. While there is no discernible difference between out-group effort levels, we find that agents' effort in in-group matches is, on average, higher than in out-group matches regardless of whether they expect principals to show greater leniency toward in-group than out-group agents in their bonus award decisions, but it is highest if they expect the principals to be more lenient toward in-group agents. Putting the affirmative result differently, we find that *agents' effort in in-group matches is increasing in their expectation of principals' in-group favoring bias in reward choices* (Result 10).

Another way of approaching this conclusion is by looking at the marginal effects of a change in expected outcome on the agents' effort choice, averaging across types of expected reward bias from the principals, and of a change in expected reward bias on the agents' effort at particular values of expected demanded outcome. We compute both sets of marginal effects from the linear regression of agents' effort level on expected demanded outcome, expected difference in demanded outcomes (expected principals' bias), group status, agents' type, the interaction of those variables, and the round of play. The full table of the regression results is given in Table C.1 in the appendix. With respect to the first consideration, we find that for a one unit increase in expected demanded outcome, agents who expect leniency from in-group principals (i.e., that in-group principals' demand for outcome to grant a reward is lower for in-group agents than for out-group agents) increase their effort by 12.5% of the effort-range in in-group matches and by 5% of the effort-range in out-group matches. Similarly, for a one unit increase in expected demanded outcome, agents who believe that in-group principals will treat them more harshly than out-group principals (i.e., that in-group principals' demand for outcome to grant a reward is higher for in-group agents than for out-group agents) put in 10% and 9% more effort in in- and out-group matches, respectively. In short, while agents respond to expected increase in demand in general in a very predictable way – by raising effort – they do so in a way that is fundamentally contingent on identity and on their expectations of principals' bias.

Figure 5 plots the second set of marginal effects, of expected in-group bias in demanded outcome over the absolute level of expected demanded outcomes in in- and out-group matches. As the figure makes clear, a marginal increase in the agents' expectation of the in-group bias from the in-group principals leads them to choose increasingly higher effort in in-group matches and increasingly smaller effort in out-group matches with greater expected demanded outcomes.

Figure 5: Marginal effect of bias in expected outcome demand on level of agents' effort in in- and out-group matches plotted over the expected demanded outcome (agents' beliefs); predictions are based on a least squares regression of effort on expected outcome demanded, in-group status, whether in- or out-group principals are expected to be more lenient, assigned type, round of play, and the interaction of these variables.



5.3 Robustness of behavioral patterns to history of play

Because our research design seeks to separate non-strategic statistical responses from the strategically driven ones, it is important to rule out the possibility that the behavior we are characterizing is induced by learning while participating in the experiment. Recall that in the experiment, we do not give the subjects any group-level feedback; subjects only observe the payoff generated in their match. However, it may, in principle, be possible that subject behavior tracks different individual experiences in in- vs. out-group matches. If the latter is the case, our argument that the artificial identity does away with the possibility of a non-strategic statistical discrimination would be weakened. Given that principals' attribution is not related to the experience they had in previous rounds, we can dismiss this concern.

There is no correlation between the difference in level of outcomes observed in previous rounds between in- and out-group matches and the attribution decision in the current round ($\rho = -.03$, $p = .43$). Similarly, there is no correlation between differences in past outcomes and principals' asymmetric treatment of agents in the decision whether to award a bonus ($\rho = .01$, $p = .87$).

Modeling doubling decisions in more detail as function of the observed outcome, whether principals sanction based on a threshold, in-group status of the matched agent, and average observed outcome in in-group matches in the past, reveals no effect of such history of favorable experience with in-group agents on the decision whether to double effort or type. In contrast, the propensity to double effort is estimated to rise for principals that follow a threshold strategy with outcome when in in-group matches, a finding similar to the one from our analysis of the differences-in-means represented in Figure 3. In particular, the marginal effect of a one unit-increase in favorable in-group outcome-history for such principals decreases propensity to double effort by .07 units (p=.18), holding concurrently observed outcome at it's empirical mean. The marginal effect of being in an in-group match, in contrast to being in an out-group match, ranges from .04 (p=.44) at outcome 4 to .11 (p=.30) at outcome 7 with increasing but not systematically different from zero marginal effects at values in between.[15]

The absence of the history of play effects on principals' beliefs, their reward decisions, and agents' effort choices means that the asymmetric responses to identity matches that we observe

[15] These and other results in this subsection are from a model that assumes that the history variable completely captures the past. The directionality of marginal effects is robust to explicitly modeling temporal effects.

cannot be based on experienced statistical differences between groups. Asymmetric treatment of groups must arise from unreinforced asymmetric beliefs about what to expect from the behavior of the members of the two groups.

5.4 A non-strategic treatment

We next provide a measure of how much asymmetric attribution depends on a strategic environment in contrast to a non-strategic environment. In the latter environment, whatever asymmetry in beliefs is observed must be due to the 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. (By design, the possibility of Phelpsian statistical discrimination is precluded by the artificial identities that are not anchored in stereotypes or correlated with payoff-relevant factors.) The ideal non-strategic treatment would measure prejudice of otherwise strategically-minded individuals by exposing the subjects we identified as following a threshold strategy to a similar treatment but without strategic incentives. Because implementing such a treatment would almost surely create framing contamination for either strategic or the non-strategic part of the treatment, we do not conduct it. Instead, we implement an across-subject design with the non-strategic treatment which closely follows the main treatment discussed earlier. The only substantial difference between strategic and non-strategic treatment is that in the non-strategic treatment, principals cannot choose whether to give a bonus.[16]

Ultimate attribution error and strategically-driven asymmetric attribution should be most easily observed at low or high levels of outcome. While the notion of a good vs. bad outcome is well-defined in the strategic treatments because it is identified endogenously relative to the imputed threshold for awarding a reward, no such natural endogenous identifier exists in the non-strategic treatment. In what follows, then, we draw the line of "good" outcomes with respect to the non-strategic treatment at 5 or above, and "bad outcomes" at 3 or below. Since at medium outcomes (here: 4), we do not have a theoretically or empirically grounded predictions for how ultimate

[16] Behavior in the non-strategic treatment also provides a robustness check on whether agents are correctly incentivized by the differences in monetary payoffs offered. Subjects in the non-strategic treatment show almost perfect pooling behavior, 84% of observations on outcome are in the range of 3 to 5, suggesting that they are responsive to levels of differences in numeric payoffs assigned to them.

attribution error manifests itself, which we have at low or high outcomes, we restrict our attention in the comparison of behavior in strategic and non-strategic environments to low and high outcomes.[17]

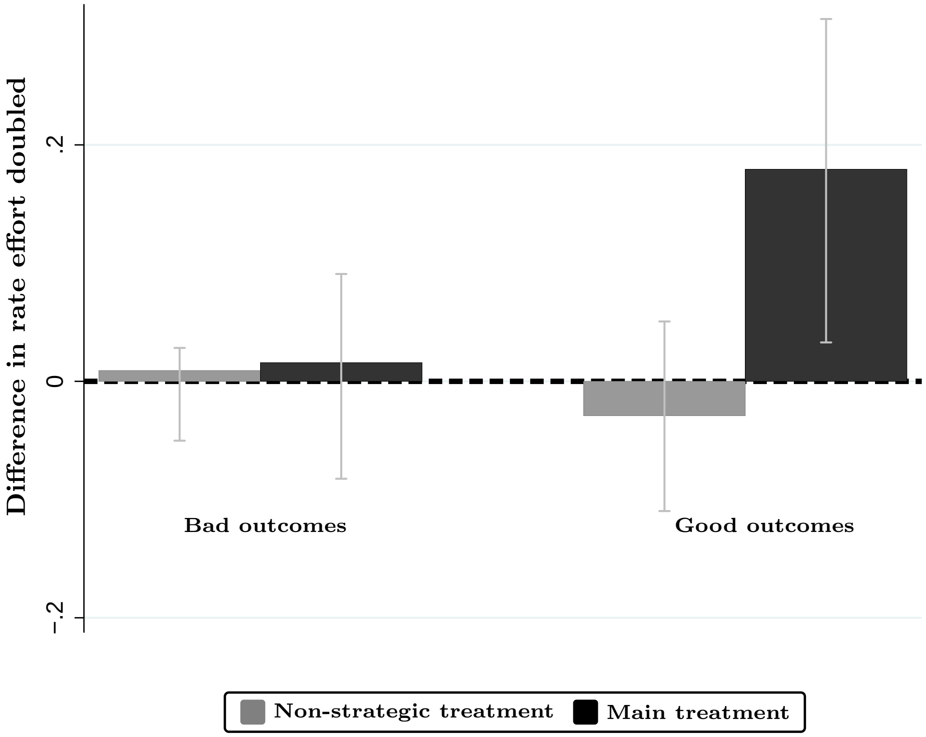
We find that the attribution asymmetries in our main strategic treatment are in fact, a consequence of the strategic environment. Figure 6 illustrates where strategic attribution asymmetry occurs on top of whatever ultimate attribution error exists as elicited in the non-strategic treatment. We provide differences in proportion of observations when effort is doubled between in-group and out-group matches for the non-strategic treatment and strategic main treatment.

We observe no asymmetries in attribution at either bad or good outcomes; the gray bars in Figure 6 show no significant difference in the proportion of effort doubled between in- and out-group matches in the non-strategic treatment. In contrast, the black bars demonstrate that at bad outcomes, threshold-strategy principals in the main treatment display a large in-group bias in their belief that effort is higher than type. Principals believe that effort is higher upon observing outcomes at a higher rate in in-group matches (.57) than in out-group matches (.33) producing the systematic in-group biased attribution.[18]

[17] The outcome of 4 in the non-strategic treatment corresponds to highly variable attribution choices. In the aggregate, those choices suggest in-group bias in attribution to effort; the different in proportion effort doubled between in- and out-group is .32 (p=.00). However, restricting attention to principals whose doubling choices are monotonic – defined by principals adhering to a threshold in the outcome space below which they interpret outcomes as bad and above which they see outcomes as good (i.e., when the set of outcomes the principal sees as good does not overlap with the set of outcomes s/he perceives as bad), this apparent bias disappears. It is exactly principals that show non-monotonicities in their choices that drive biased beliefs in the middle of the outcome range of the non-strategic treatment.

[18] Note that for many principal-specific thresholds, in-group biased attribution asymmetry already exists for medium values.

Figure 6: Difference in rate at which principals double effort in in-group matches minus rate in out-group matches (= in-group bias in attribution); bootstrapped 95%-confidence intervals are shown. Observations from the main treatment are shown for the 42 principals that follow a threshold-strategy



The rationale for these differences encapsulates the logic of strategically driven statistical discrimination. Principals who follow a threshold-strategy in the strategic treatment likely reason as follows. They trust that the high result, upon observing which they tend to give a bonus as the strategy requires, is more likely to be a product of effort in in-group matches than in out-group matches. In out-group matches, in contrast, they believe that noise or type were more favorable for the same high outcomes so that high outcomes are not a product of high effort but of agent's luck. Though they believe this, they nevertheless stick with the bonus strategy they are playing as the right bonus strategy because it is costless to do so and in the equilibrium they are playing, it incentivizes effort, however small, from the agents. Of course, because principals are less likely to give a bonus to out-group agents than to in-group agents at high outcomes, out-group agents will choose lower effort, which will reinforce the principals' bias. This reasoning is not present in the non-strategic setting, and so we do not observe divergence in attribution of high outcomes to effort between in-group and out-group matches at the high level.

The non-strategic treatment also provides evidence against another interpretation of principals' behavior. It may be that the principals who are discriminating in beliefs are just the kind of people who are particularly more likely to discriminate in rewards. In particular, they may just be the people who are simply more prone to discrimination. In that case, we are, arguably, not measuring strategically induced beliefs on the part of the principals, but rather identifying the types who are predisposed to discriminate (This doesn't affect the agent side, where this kind of criticism is less plausible.) The fact that there is no discrimination in principals' beliefs in the non-strategic treatment, however, argues against such a selection-based interpretation.

As a final consideration on this treatment, note that we do not know if the non-existence of attribution at low or high outcomes in the non-strategic treatment is a function of the weakness of induced identities in this experiments or of non-existence of ultimate attribution error in general. What we are able to infer from our observations is that inducing weak identities generates strategically-driven attribution asymmetries even if the full extent of ultimate attribution error may not be recoverable in the artificial environment of the laboratory. If, as seems reasonable, we assume that effects of identity are monotonic in its strength (Eckel and Grossman, 2005; Goette, Huffman and Meier, 2006), then even if with stronger identities, the ultimate attribution error is likely to emerge, we should interpret strategically-driven attribution asymmetries of the kind we characterize above as a first-order phenomenon in the settings with prima facie evidence of prejudice.

6 Discussion

The experiment described in this paper aims to provide a behavioral evaluation of the specifically Arrovian theory of discrimination. As such, our goal is to evaluate the presence of discriminatory behavior that is driven by the individuals' beliefs about strategic play in a principal-agent setting with social identities – beliefs that are distinct from pre-existing asymmetric, group-based generalizations, whether rationally or psychologically sustained.

Our key results, both at the sample and the individual level, show the existence of systematically asymmetric group-specific judgments by the principals. Given the average tendency of agents' choices toward greater effort investment in in-group matches, the principals' judgments can be rationalized by beliefs about the agents' choices that are, in fact, on average, consistent with those choices. Looking at the individual-level agent behavior more closely, though, we can distinguish two empirically prominent strategy profiles: (1) principals are more lenient toward in-group agents in their reward decisions, agents choose higher effort when matched with a principal of their own group, and principals attribute good outcomes to high effort more often in in-group matches than in out-group matches; and (2) principals demand higher outcomes of agents from their group, agents choose higher effort when they share a group identity with their principals, and principals attribute good outcomes to high effort more often in in-group matches than in out-group matches.[19]

The strategy profile (1) is inconsistent with equilibrium play in the explicitly instantiated game: conditional on expecting a lower threshold in in-group matches, agents' best response is to *lower*, not raise, their effort choice. However, it appears that a substantial proportion of agents are doing precisely the opposite: anticipating a lower threshold in in-group matches, they raise their effort choice. While in the context of the explicitly instantiated game this may appear puzzling, this profile can be readily interpreted as reciprocal in-group favoritism that may correspond to an equilibrium of a different game that may motivate subjects' interpretations of the proper behavior in social identity contexts. A principal who is favorably disposed toward the agents with the same group identity may set a lower threshold in the outcome space above which s/he is willing to award a bonus, attributing an occasional low performance to an unlucky noise draw. Similarly disposed agents respond to such favorable treatment with increased effort in in-group matches as a part of the reciprocal in-group favoritism. Showing marginal leniency towards agents from their own group, then, is effort-maximizing for the principals, who hold concurring asymmetric beliefs that attribute good outcomes to higher effort in in-group matches.

In profile (2), principals are *more* demanding of in-group agents for awarding a bonus, and agents respond by choosing higher effort when matched with a principal who shares their group identity. From the principal's standpoint, whether the agent does it anticipating a higher or, as in profile (1), a lower, threshold demanded is immaterial for sustaining the asymmetric beliefs attributing good outcomes to effort more often in in-group than in out-group matches. Such beliefs are consistent with both profiles. Despite their differences, thus, strategy profiles (1) and (2) have

[19] While some of our subjects in the role of agents choose lower effort in in-group matches, both the unconditional and conditional on the agent's beliefs expectations are that agents choose higher effort in in-group than in out-group matches.

in common a key feature: asymmetric reward and effort choices emerge alongside biased but *correct* beliefs about agent choices, which are induced by an anticipation of principals' reward choices. Whether the agents' beliefs about the principals' (average) reward strategies are correct, as they are in profile (1) or not, as they appear to be in profile (2), has no effect on this, though it bears noting that the fact that numerous principals play profile (1) strategies may justify agents' average beliefs and their corresponding actions we observe.

Our final piece of evidence reinforcing the interpretation of these asymmetries as evidence for the existence of the specifically strategic statistical discrimination is the relationship between principals' asymmetric attribution choices and their awareness of their own biases in beliefs and reward choices as indicated in their answers to the exit survey. We find that principals whose reward and doubling choices are asymmetric tend to be aware of it, and, further, that principals who are more likely to attribute to effort upon observing good outcomes when they share an identity with their agent, are the ones who are more aware of their asymmetric treatment of agents in their bonus award decisions.

In an exit survey, we ask subjects – who received the main treatment – whether their decision to award a bonus and their attribution of outcomes was affected by the identity of their agent-match. Subjects in the role of principals who indicate that they were influenced in their reward decision by group membership show a significantly larger in-group bias in awarding a bonus (.14) than subjects in the same role who claim group membership did not matter (.08; $p < .01$). Subjects who acknowledge that their agent's identity did matter for their own doubling decision as principal show significantly stronger asymmetric beliefs in attributing good outcomes to effort in in-group matches than those subjects who said group identities did not matter for their doubling choices (in-group biases in attributing good outcomes to effort are .10 and .01, respectively, with $p < .01$ for the difference-in-distributions test). Further, principals who are aware of their bias in reward decisions are also more likely to attribute high outcome to effort when in in-group matches than when in out-group matches (difference = .05, $p < .05$).

Agents who are playing profile (1) strategies – choosing higher effort in in-group than in out-group matches when believing that in-group principals are more lenient than out-group principals – are highly likely to be aware of their bias in effort choices; 71% of them acknowledge that their investment into effort was, at least partially, contingent on the identity of their matched principal.

In contrast, agents whose choices are consistent with the posited profile (2) strategies – higher effort in-group than in out-group matches when expecting higher outcome demands from in-group principals – are much less likely to acknowledge that their choices where driven by group identities (45%, differences between those two groups of agents are systematic with $p < .01$).[20]

Another indicator that supports our characterization of the two different strategy profiles is the look of the relationship between subjects' ability to understand the game presented to them and their beliefs and actions. Playing a strategy profile that sees agents' decrease their effort in in-group matches upon expecting lower levels of in-group bias in principals reward decisions, goes along with higher levels of strategic sophistication. The marginal effect of strategic sophistication on effort turns increasingly negative with an increase in expected principals' demands in in-group matches. Similarly, principals following the strategy profile that is characterized by higher demands to in-group agents in terms of outcome to be willing to award a bonus and subsequent attribution of good outcomes to effort are also more strategically sophisticated ($\rho = .55, p = .00$). We elicited strategy sophistication in a two-stage beauty contest game conducted in 3 out of the 6 sessions of the main treatment which allowed us to assess subjects' ability to respond optimally to their beliefs about other subjects' behavior.

Finally, the emergence of agents strategy profile that sees them choose lower effort in out-group matches than in in-group matches when expecting harsher treatment in out-group matches than in-group matches (equivalent to an increase in effort when expecting more leniency from in-group principals) is not driven by risk preferences. Agents risk attitudes are not correlated with their effort choices when conditioned on their beliefs about principals demands ($\rho = -.06, p = .22$).

7 Conclusion

Social identity relationships fundamentally affect mutual expectations of agents and principals, and through those, agents' performance and career prospects. The experiment we present in this paper analyzes those expectations in a strategic principal-agent setting that models relationships between voters and elected officials and employers and employees. Our treatments vary whether group identification, which is not payoff-incentivized, is specifically primed, as well as the in-group vs.

[20] 62% of agents who show no difference in their expectation of outcome demands by principals or their effort choices between in- and out-group claim that their choices where driven by their matches identity.

out-group matching of the principals and agents. We elicit from the principals their beliefs about agents' type and effort (monetarily incentivized), from the agents' their expectations about the outcome demanded by the principal to award a bonus, and determine how principals' explanations of agents' performance vary depending on whether they share the social identity with the agents.

Our key finding is that principals' and agents' choices may, through (correct) mutual expectations, sustain a pattern of beliefs that is observationally equivalent to those conforming to "ultimate attribution error" that is at the core of psychological accounts of prejudice and discrimination. Upon observing good outcomes, principals who reward agents in an outcome-contingent way tend to attribute those outcomes more readily to their agents' effort and to reward their agents more frequently when they share a social identity; and in turn, agents who share a social identity with their principals tend to invest more into effort in expectation of principals' award choices.

A direct implication of our analysis is that in strategic environments, discriminating behavior does not necessarily go together with prejudicial stereotyping. Principals' reward choices that seem to favor in-group agents reflect those agents' increased willingness to invest into effort – a circumstance that is correctly represented in principals' beliefs about the source of observed outcomes. In our setting, asymmetric social identity-contingent interpretations of agent performance that are observationally equivalent to prejudice are not based on a taste for unequal treatment or incorrect beliefs about differences in in-group agents' vs. out-group agents' performance but on correct anticipation of the opportunity to better incentivize those agents with whom principals share a group identity. These results lend clear support to the view that while attributing to an apparently "exogenous cause what is in fact an endogenous difference seems to be an important feature of how stereotypes work in practice, (. . .) an equilibrium with stereotypes does not require any such misattribution" (Coate and Loury, 1993, 1227).

An important related implication is that the existing measures of prejudice in the observational studies are based on partial equilibrium models and may be identifying a joint measure of prejudice and rational expectations associated with an equilibrium performance rather than prejudice alone. One of the benefits of the present experimental analysis is that it allows us to see that in a controlled setting and underscores the importance of the strategically driven asymmetric attributions as a first-order behavioral phenomenon. Attempts to measure prejudice in observational data settings should prioritize finding ways to account separately for the individuals' rational expectations.

References

Akerlof, George and Rachel Kranton. 2000. "Economics and Identity." *Quarterly Journal of Economics* 115(3):715–53.

Akerlof, George and Rachel Kranton. 2010. *Identity Economics.* Princeton: Princeton University Press.

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

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

Anderson, Donna M and Michael J Haupert. 1999. "Employment and statistical discrimination: A hands-on experiment." *The Journal of Economics* 25(1):85–102.

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

Becker, Gary S. 1973. *The economics of discrimination.* University of Chicago press.

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

Bendick, Marc. 2007. "Situation testing for employment discrimination in the United States of America." *Horizons stratégiques* (3):17–39.

Bernhard, Helen, Ernst Fehr and Urs Fischbacher. 2006. "Group Affiliation and Altruistic Norm Enforcement." *American Economic Review* 96(2):217–21.

Bertrand, Marianne and Sendhil 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.

Billig, Michael and Henri Tajfel. 1973. "Social categorization and similarity in intergroup behaviour." *European Journal of Social Psychology* 3(1):27–52.

Bird, Karen, Thomas Saalfeld and Andreas M Wüst. 2010. *The political representation of immigrants and minorities: Voters, parties and parliaments in liberal democracies.* Taylor & Francis.

Bowles, Samuel, Glenn Cartman Loury and Rajiv Sethi. 2009. Group Inequality. Technical report Institute for Advanced Study, School of Social Science.

Chandra, Amitabh. 2000. "Labor-market dropouts and the racial wage gap: 1940-1990." *The American Economic Review* 90(2):333–338.

Chandra, Kanchan. 2004. *Why Ethnic Parties Succeed.* New York: Cambridge University Press.

Charness, Gary, Luca Rigotti and Aldo Rustichini. 2007. "Individual Behavior and Group Membership." *American Economic Review* 97(4):1340–52.

Chen, Roy and Yan Chen. 2011. "The Potential of Social Identity for Equilibrium Selection." *American Economic Review* 101(6):2562–89.

Chen, Yan and Sherry Li. 2009. "Group Identity and Social Preferences." *American Economic Review* 99(1):431–57.

Coate, Stephen and Glenn C Loury. 1993. "Will affirmative-action policies eliminate negative stereotypes?" *The American Economic Review* pp. 1220–1240.

Coviello, Decio and Nicola Persico. 2013. An Economic Analysis of Black-White Disparities in NYPD's Stop and Frisk Program. Technical report National Bureau of Economic Research.

de Mesquita, Ethan Bueno and Dimitri Landa. 2013. "Moral Hazard with Sequential Policy Making." http://home.uchicago.edu/ bdm/PDF/whydfml.pdf.

Diehl, Michael. 1988. "Social identity and minimal groups: The effects of interpersonal and intergroup attitudinal similarity on intergroup discrimination." *British Journal of Social Psychology* 27(4):289–300.

Eckel, Catherine and Philip Grossman. 2005. "Managing Diversity by Creating Team Identity." *Journal of Economic Behavior Organization* 58:371–392.

Falk, Armin and Christian Zehnder. 2007. Discrimination and in-group favoritism in a citywide trust experiment. Technical report IZA Discussion Papers.

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

Fischbacher, Urs. 2007. "z-Tree: Zurich Toolbox for Ready-made Economic Experiments." *Experimental Economic* 10(2):171–178.

Fox, Richard and Eric Smith. 1998. "The Role of Candidate Sex in Voter Decision-Making." *Political Psychology* 19(2):405–419.

Fox, Richard L and Jennifer L Lawless. 2010. "If only they'd ask: Gender, recruitment, and political ambition." *Journal of Politics* 72(2):310–326.

Fox, Richard L and Jennifer L Lawless. 2011. "Gendered Perceptions and Political Candidacies: A Central Barrier to Women's Equality in Electoral Politics." *American Journal of Political Science* 55(1):59–73.

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

Goette, Lorenz, David Huffman and Stephan Meier. 2006. "The Impact of Group Membership on Cooperation and Norm Enforcement: Evidence Using Random Assignment to Real Social Groups." *American Economic Review* 96(2):212–6.

Goette, Lorenz, David Huffman and Stephan Meier. 2012. "The Impact of Social Ties on Group Interactions: Evidence from Minimal Groups and Randomly Assigned Real Groups." *American Economic Journal: Microeconomics* 4(1):101–15.

Goldin, Claudia and Cecilia Rouse. 2000. "Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians." *The American Economic Review* 90(4):715–741.

Griffin, John D and Brian Newman. 2007. "The Unequal Representation of Latinos and Whites." *Journal of Politics* 69(4):1032–1046.

Griffin, John D and Brian Newman. 2008. *Minority report: Evaluating political equality in America.* University of Chicago Press.

Hewstone, Miles. 1990. "The Ultimate Attribution Error? A review of the Literature on Intergroup Causal Attribution." *European Journal of Social Psychology* 20:311–35.

Holt, Charles and Susan Laury. 2002. "Risk Aversion and Incentive Effects." *American Economic Review* 92(5):1644–55.

Holzer, Harry and David Neumark. 2000. "Assessing Affirmative Action." *Journal of Economic Literature* 38(3):483–568.

Iversen, Torben and Frances Rosenbluth. 2008. "Work and power: The connection between female labor force participation and female political representation." *Annu. Rev. Polit. Sci.* 11:479–495.

Kaufmann, Karen. 2004. *Urban Voter: Group Conflict and Mayoral Voting Behavior in American Cities.* Ann Arber: University of Michigan Press.

Klein, Olivier and Assaad E Azzi. 2001. "The strategic confirmation of meta-stereotypes: How group members attempt to tailor an out-group's representation of themselves." *British Journal of Social Psychology* 40(2):279–293.

Knippenberg, Daan van. 2003. "Social Identity and Leadership Processes in Groups." *Advances in Experimental Social Psychology* 35:1–52.

Knowles, John, Nicola Persico and Petra Todd. 2001. "Racial Bias in Motor Vehicle Searches: Theory and Evidence." *Journal of Political Economy* 109(1).

Kramer, Roderick. 1994. "The Sinister Attribution Error: Paranoid Cognition and Collective Distrust in Organizations." *Motivation and Emotion* 18(2):199–230.

Landa, Dimitri and Dominik Duell. 2014. "Social Identity and Electoral Accountability." *American Journal of Political Science* forthcoming.

Loury, Glenn Cartman. 1976. "A Dynamic Theory of Racial Income Differences." Northwestern University, Center for Mathematical Studies in Economics and Management Science.

Lundberg, Shelly J. 1991. "The enforcement of equal opportunity laws under imperfect information: affirmative action and alternatives." *The Quarterly Journal of Economics* 106(1):309–326.

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

McLeish, Kendra and Robert Oxoby. 2007. "Identity, Cooperation, and Punishment." IZA Discussion Paper No. 2572.

Moro, Andrea. 2009. "Statistical Discrimination." *The New Palgrave Dictionary of Economics* .

Neumark, David and Michele McLennan. 1995. "Sex discrimination and women's labor market outcomes." *Journal of human resources* pp. 713–740.

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

Nosek, Brian, Frederick Smyth, Jeffrey Hansen, Thierry Devos, Nicole Lindner, Kate Ranganath, Colin Smith, Kristina Olson, Dolly Chugh, Anthony Greenwald et al. 2007. "Pervasiveness and Correlates of Implicit Attitudes and Stereotypes." *European Review of Social Psychology* 18(1):36–88.

Paxton, Pamela, Sheri Kunovich and Melanie M Hughes. 2007. "Gender in politics." *Annu. Rev. Sociol.* 33:263–284.

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

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

Persico, Nicola and Petra Todd. 2006. "Generalising the Hit Rates Test for Racial Bias in Law Enforcement, With an Application to Vehicle Searches in Wichita*." *The Economic Journal* 116(515):F351–F367.

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

Pettigrew, Thomas. 1979. "The Ultimate Attribution Error: Extending Allport's Cognitive Analysis of Prejudice." *Personality and Social Psychology Bulletin* 5(4):461–76.

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

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

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

Tajfel, Henri. 1981. *Human Groups and Social Categories.* Cambridge: Cambridge University Press.

Tajfel, Henri and John Turner. 1986. The Social Identity Theory of Intergroup Behavior. In *The Psychology of Intergroup Relations*, ed. Stephen Worchel and William Austin. Chicago: Nelson-Hall pp. 7–24.

Tajfel, Henri and Michael Billig. 1974. "Familiarity and Categorization in Intergroup Behavior." *Journal of Experimental Social Psychology* 10:159–70.

Turner, John C and Rupert Brown. 1978. "Social status, cognitive alternatives and intergroup relations." *Differentiation between social groups: Studies in the social psychology of intergroup relations* pp. 201–234.

Vaughan, Graham M, Henri Tajfel and Jennifer Williams. 1981. "Bias in reward allocation in an intergroup and an interpersonal context." *Social Psychology Quarterly* pp. 37–42.

Western, Bruce and Becky Pettit. 2005. "Black-White Wage Inequality, Employment Rates, and Incarceration." *American Journal of Sociology* 111(2):553–578.

A Statistical appendix

A.1 Session statistics

Table A.1: Number of subjects and number of observations by treatment.

Treatment		# of subjects	# of observations
	Klees	41	820
Main	Kandinskys	47	940
	Total	88	1760
Non-identity	Total	38	760
	Klees	21	420
Non-strategic	Kandinskys	19	380
	Total	40	800
		166	3320

The main treatment condition generated 41 Klees, subjects who preferred paintings by Paul Klee most of the time, and 47 Kandinskys, subjects who preferred those by Vassily Kandinsky most of the time. In the No-ST treatment we see 21 Klees and 19 Kandinskys. During the quiz, a majority of members in both groups gave correct answers in four out of five painting quizzes. Ultimately, all subjects received a payoff of $5 at this stage of the experiment. This positive group experience in a competitive environment is part of the intended group strengthening; we intentionally selected paintings whose authors are moderately easy to identify. Subjects were told how many correct answers their group gave and were notified that members of their group "gave at least as many correct answers" as members of the other group.

A.2 Summary statistics

A.2.1 Main treatment and non-identity treatment

Table A.2: Means (standard deviation), minimum, and maximum values of type, effort, outcome, doubling decision (0 = type doubled, 1 = effort doubled), and bonus decision (0 = no bonus awarded, 1 = bonus awarded) by treatment.

	Main treatment		**Non-identity**		
Variable	In-group	Out-group	**treatment**	Min	Max
Type	1.97 (.82)	2.00 (.79)	2.01 (.81)	1	3
Effort	1.88 (.78)	1.80 (.79)	1.76 (.84)	1	3
Outcome	3.77 (1.3)	3.73 (1.3)	3.81 (1.3)	1	7
Doubling	.563 (.50)	.526 (.50)	.455 (.50)	0	1
Bonus	.655 (.48)	.528 (.50)	.605 (.49)	0	1

Recall that our experimental design sought to incentivize pooling behavior among agents: when pooling occurs, principals gain no information about the agents' type from observing outcomes only. While pooling is imperfect in the data, the functional relationship between effort and type is systematically sloping downwards, as incentivized, suggesting broad congruence with the experimental design; the marginal effects of type on effort are negative in all treatments and run from a decrease of .19 (sd=.04) units of effort when type increases by one unit in in-group matches of the main treatment to .25 (sd=.06) units of effort in the non-identity treatment. Agents' behavior creates a distribution of outcomes centering at 4; 75% of observations in the main treatment and 76% of observations in the non-identity treatment have values of outcome in the interval from 3 to 5.

To ease interpretation, we mostly refer to results here by levels of outcome, defining *low outcome* to be outcomes smaller than 4, *medium outcome* to be outcomes of 4, and *high outcome* to be outcomes higher than 4. (All associated difference-in-means and difference-in-distribution tests, however, use the non-degenerate outcome variable to avoid erroneous conclusions based on aggregation effects.)

Table A.3: Proportion effort doubled (in contrast to type doubled) by level of outcome and treatment; low outcome is outcome $<$ 4, medium outcome is outcome $=$ 4, and high outcome is outcome $>$ 4, standard deviation in parentheses

	Main treatment		**Non-identity**
Outcome	In-group	Out-group	**treatment**
Low	.57 (.50)	.55 (.50)	.56 (.50)
Medium	.59 (.49)	.56 (.50)	.40 (.49)
High	.54 (.50)	.44 (.50)	.36 (.48)

Table A.4: Proportion bonus awarded by level of outcome and treatment; low outcome is outcome $<$ 4, medium outcome is outcome $=$ 4, and high outcome is outcome $>$ 4

	Main treatment		**Non-identity**
Outcome	In-group	Out-group	**treatment**
Low	.53 (.50)	.38 (.49)	.53 (.50)
Medium	.69 (.47)	.63 (.48)	.61 (.49)
High	.81 (.40)	.65 (.48)	.70 (.46)

A.2.2 Non-strategic treatment

Table A.5: Means (standard deviation), minimum, and maximum values of type, effort, outcome, doubling decision (0 = type doubled, 1 = effort doubled), and bonus decision (0 = no bonus awarded, 1 = bonus awarded) by treatment.

	Non-strategic treatment	
Variable	In-group	Out-group
Type	2.00 (.79)	2.05 (.79)
Effort	2.11 (.77)	2.17 (.76)
Outcome	4.03 (1.1)	4.14 (1.1)
Doubling	.655 (.48)	.570 (.50)

Table A.6: Proportion effort doubled (in contrast to type doubled) by level of outcome and treatment; low outcome is outcome < 4, medium outcome is outcome $= 4$, and high outcome is outcome > 4, standard deviation in parentheses

	Non-strategic treatment	
Outcome	In-group	Out-group
Low	.80 (.41)	.80 (.40)
Medium	.55 (.50)	.42 (.50)
High	.49 (.48)	.51 (.50)

B Robustness checks

B.1 Average treatment effects

Whenever we present average treatment effects we run appropriate difference-in-means and difference-in-distribution tests. Additionally, we conduct equivalent Fligner-Policello test, which relaxes both equal variance and approximately normal distribution, and report results as significant only when this tests delivers nearly identical p-values. Such a robustness check is valuable since bootstrapping the variance of effort for each (type,treatment)-sub-sample yields systematically unequal variance in in- and out-group. As a further robustness check for the Wilcoxon rank sum test, we simulated permutations of effort and ran the test on each generated sample. The created distribution of test-statistics, again, has to yield nearly identical results to the Wilcoxon test to make us claim that a difference is significantly different from zero. Throughout the paper, whenever we compare distributions associated with levels of effort or outcome, we run all three tests; p-values from a two-tailed test are reported. For the comparison of proportions (rate of effort doubled and rate of bonus awarded) we report 95%-confidence bounds from a session-clustered bootstrap.

B.2 Is the estimate of principals' threshold and thus the identification of an incentive-based strategy robust?

Figure B.1: Principals: Distribution of proportion of observations consistent with principal deploying their individual error-minimizing threshold in deciding whether to award a bonus

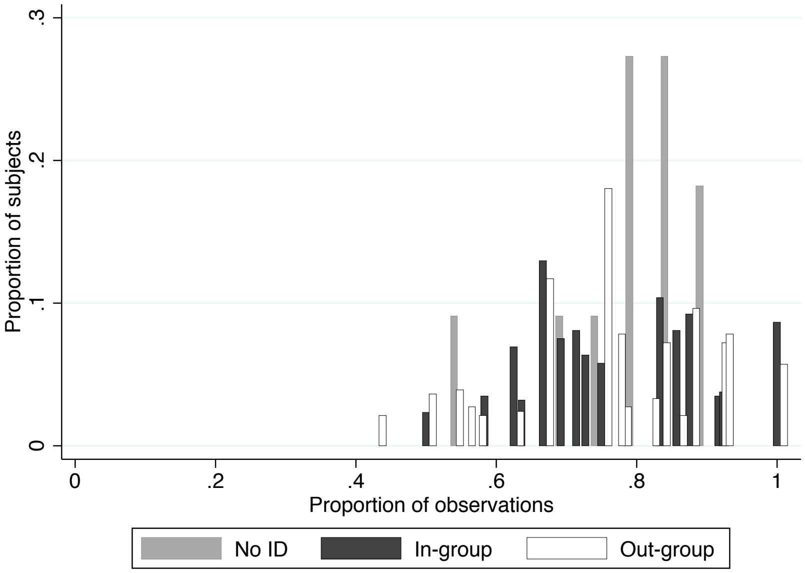
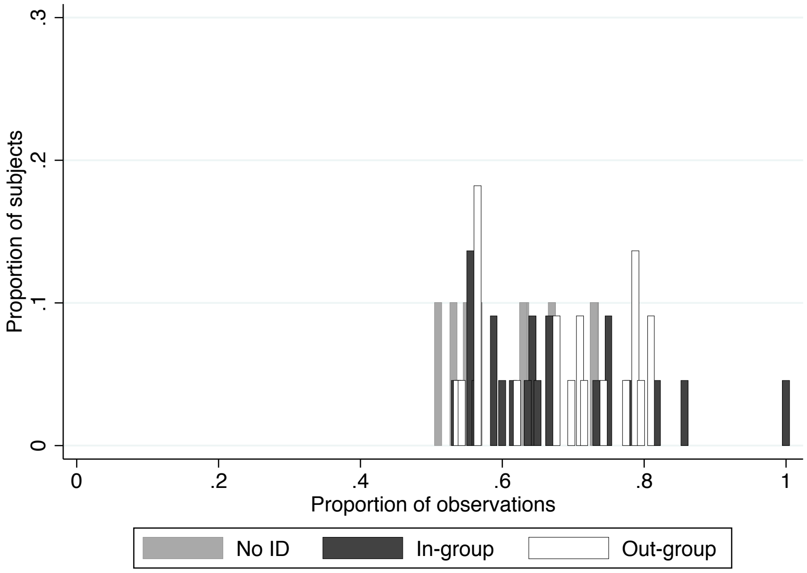


Figure B.2: Principals: Distribution of proportion of observations consistent with incentivizing strategy profile with a fixed threshold at medium outcome



Across subjects, there is variation in errors in classifying reward decisions associated with each potential outcome threshold. For some principals, the threshold that minimizes error, T^*, performs much better than any other threshold but for others either several threshold values are almost equally good in separating the decisions to award a bonus from those not to award, or many thresholds perform equally poorely and T^* is simply the least troublesome. To address concerns arising from computing T^*, we sum, fo each principal, all classification errors associated with outcomes below T^* and also sum those associated with the outcomes above it. Should both sums be large and similar to the error associated with T^*, such a principal would be choosing almost at random. In contrast, should both sums be large and, in addition, much larger than the error associated with T^*, such principals would be following T^* rather well. Two further sets of principals can be classified with these two measures. The first is the of principals with a small sum of errors above or below T^* that is associated with outcome values 3, 4, or 5. Given that 3, 4, and 5 are equilibrium thresholds that yield highest expected outcome values, the existence of such principals should not come as surprise. Finally, there is a set of principals with a small sum of errors above or below T^* that is associated with very low or very high outcomes. These principals are consistently staying away from a reward strategy that goes with the payoff-maximizing equilibrium profiles. Excluding principals that either choose almost at random and this last set of principals does not change the results presented in Section 5.2.2.

C Agents' beliefs

Figure C.1: Agents' inquiries of payoff consequences of expected demanded minimal outcomes

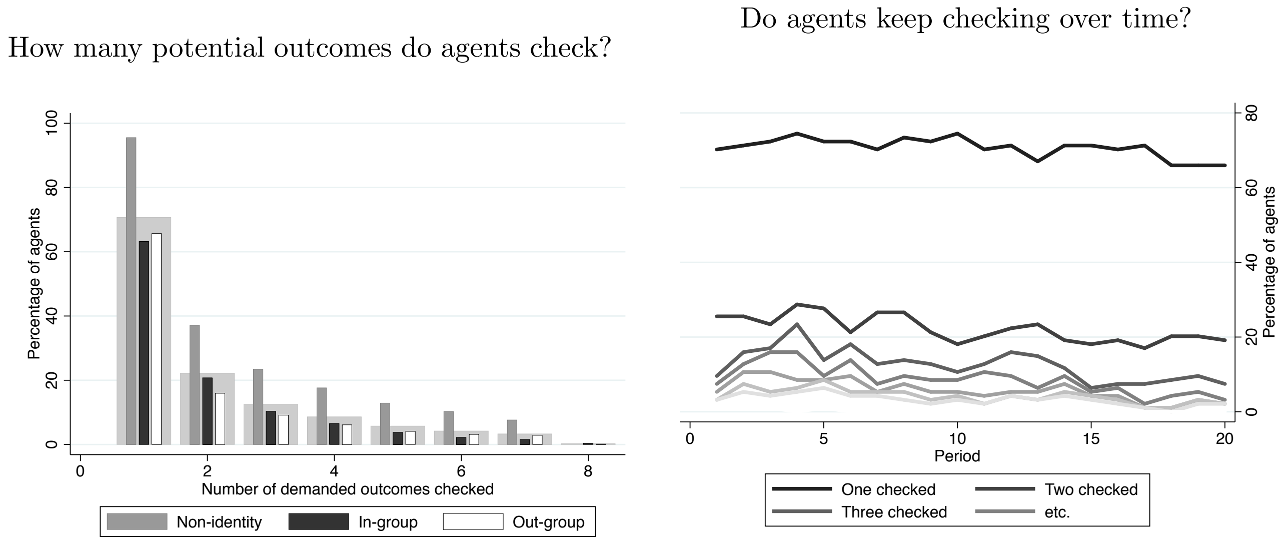


Table C.1: Least square regression of agents' effort on covariates for main treatment; standard errors are computed based on clustering by session.

Variable	**Coefficient**	**(Std. Err.)**
Expected Outcome Demand	0.249	(0.118)
Difference in expected Demand	-0.149	(0.328)
Demand × Difference in Demand	0.091	(0.069)
Ingroup	-0.025	(0.375)
Ingroup × Demand	0.043	(0.131)
Ingroup × Difference in Demand	0.455	(0.636)
Ingroup × Demand × Difference in Demand	-0.272	(0.146)
Type	-0.152	(0.136)
Demand × Type	-0.020	(0.053)
Difference in Demand × Type	-0.101	(0.161)
Demand × Difference in Demand × Type	0.017	(0.039)
Ingroup × Type	0.050	(0.186)
Ingroup × Demand × Type	-0.029	(0.060)
Ingroup × Difference in Demand × Type	0.004	(0.284)
Ingroup × Demand × Difference in Demand × Type	0.036	(0.061)
Period	-0.008	(0.007)
Intercept	1.455	(0.276)
N	966	
Session	6	
Number of agents	55	

D Instructions

Introduction

During the following experiment, we require your complete undivided attention and ask that you follow instructions carefully. Please turn off your cell phones and, for the duration of the experiment, do not take actions that could distract you or other participants, including opening other applications on your computer, reading books, newspapers, and doing homework.

This is an experiment on group decision-making. In this experiment you will make a series of choices. At the end of the experiment, you will be paid depending on the specific choices that you made during the experiment and the choices made by other participants. If you follow the instructions and make appropriate decisions, you may make an appreciable amount of money.

This experiment has 3 parts. Your total earnings will be the sum of your payoffs in each part plus the show-up fee. We will start with a brief instruction period, followed by Part 1 of the experiment. After Part 1 is completed, we will pause to receive instructions for Part 2 and complete the session accordingly.

If you have questions during the instruction period, please raise your hand after I have completed reading the instructions, and your questions will be answered out loud so everyone can hear. Please restrict these questions to clarifications about the instructions only. If you have any questions after the paid session of the experiment has begun, raise your hand, and an experimenter will come and assist you. Apart from the questions directed to the experimenter, you are expressly asked to refrain from communicating with other participants in the experiment, including making public remarks or exclamations. Failure to comply with these instructions will result in the termination of your participation and the forfeiture of any compensation.

Part 1

In Part 1 of the experiment, everyone will be shown 5 pairs of paintings by two artists, Paul Klee and Wassily Kandinsky. You will be asked to choose which painting in each pair you prefer. You will then be classified as member of the “KLEEs” (or “a KLEE” as a shorthand) or member of the “KANDINSKYs” (or “a KANDINSKY” as a shorthand) based on which artist you prefer most and informed privately about your classification. Everyone’s identity as a KLEE or as a KANDINSKY will stay fixed for the rest of the experiment (that is, in both Part 1 and Part 2 of the experiment).

You will then be asked to identify the painter (Klee or Kandinsky) of five other paintings. For each of those paintings, you will be asked to submit two answers: your initial guess and your final answer. After submitting your initial guess, you will have an opportunity to see the initial guesses of your fellow KLEEs if you are a KLEE, or of fellow KANDINSKYs if you are a KANDINSKY, and then also an opportunity to change your answer when you are submitting your final answer.

If you are a KLEE and a half or more of KLEEs give a correct final answer then, regardless of whether your own final answer was correct or incorrect, you and each of your fellow KLEEs will receive $1. Similarly, if you are a member of the KANDINSKYs and a half or more of KANDINSKYs give a correct final answer then, regardless of your own final answer, each of the KANDINSKYs, including you, will receive $1. However, if you are a KLEE and more than a half of KLEEs give an incorrect final answer, then, regardless of whether your own final answer was correct or incorrect, you and each of the KLEEs will receive $0. And similarly, if you are a KANDINSKY and the final answers from more than a half of KANDINSKYs were incorrect, then you and each of your fellow KANDINSKYs will receive $0 regardless of what answer he or a she gave personally.

In addition, if you and your fellow group members answer at least as many quiz questions correctly than members of the other group, you will receive an additional payoff of $1. That is, if you are a KLEE and you and your fellow KLEEs give more correct answers than the KANDINSKYs, you receive the additional payoff. If you are a KANDINSKY and you and your fellow KANDINSKYs give more correct answers than the KLEEs, you receive the additional payoff.

We will now run Part 1 of the experiment. After Part 2 has finished, we will give you instructions for Part 2.

Part 2
We will now move on to Part 2 of the experiment. Part 2 will consist of 20 different rounds. At the beginning of the first round, you will be randomly assigned a role of either Player 1 or Player 2. You will keep that role for the rest of Part 3 of the experiment. Throughout this part of the experiment, you will also retain your identity as a member of the KLEEs or a member of the KANDINSKYs, as assigned in Part 2 of the experiment.

Matched group
In each round, all participants in the experiment will be randomly matched into pairs, each consisting of one Player 1 and one Player 2. Because every participant will be randomly re-matched with other participants into a different group in each round of the experiment, the composition of matched pairs will vary from one round to the next. All of participants' interactions will take place anonymously through a computer terminal, so your true personal identity will never be revealed to others, and you will not know who precisely is in your pair in any round of the experiment. However, every time you are matched with another participant (Player 1 or Player 2), you will be told whether that participant is a member of the KLEEs or a member of the KANDINSKYs.

In each round, a member of the group who takes on the role of Player 1 in that round will be randomly assigned a number, which we will refer to as Player 1's *special number*. That number will be shown only to that participant and never to other participants in the experiment. You should know, however, that Player 1's *special number* is one of three possible numbers: 1, 2 or, 3, and is chosen by the computer for assigning to Player 1 so that each of these numbers is equally likely to be picked. In each round, Player 1 is assigned a new *special number*, which stays fixed until the round ends, at which point a new *special number* is assigned. As with all other players, her identity as a member of the KLEEs or a member of the KANDINSKYs does not change from one round to the next.

Choices within each round of the experiment
At the beginning of each round, in each group, the member who is designated as Player 1 will choose a number: 1, 2, or 3, which you can think of as Player 1's level of *effort*. Please note that, while Player 1's *effort* is her choice, Player 1's *special number* is not her choice, but is assigned to Player 1 by the computer. Player 1's choice of *effort* will help determine *the choice outcome* in that round. In particular, *the choice outcome* will be computed as follows:

the choice outcome = Player 1's effort + Player 1's special number + random bump,

where the possible values of the *random bump* are -1, 0, or 1, and any one of these three values will be possible and equally likely to occur.

For example, suppose that a given Player 1's *special number* is 2, he or she chooses a level of *effort* equal to 1, and the realized value of the *random bump* is -1. Then *the choice outcome* is 2 + 1 - 1 = 2.

After *the choice outcome* is computed, it will be shown to Player 2. However, Player 2 will not see Player 1's *special number* nor her choice of *effort* nor the realized value of the *random bump.*
After seeing *the choice outcome,* Player 2 will be given an opportunity to *increase* the outcome by doubling the contribution to outcome of either Player 1's *effort* or of her *special number* – whichever of those two Player 2 decides to increase. A new outcome will, then, be computed, based on the corresponding *choice outcome*, but now increased because of the doubled contribution of *effort* or

special number, as indicated by Player 2. We will refer to this new resulting outcome as *the increased outcome.*

For example, suppose that a given Player 1's *special number* is 2, he or she chooses a level of *effort* equal to 1, and the realized *random bump* is -1. Suppose, further, that Player 2 decides to increase the outcome by raising the contribution of *effort.* Then *the increased outcome* is 2 + [2(1)] - 1 = 3. (Note that the product in the square brackets [] is the newly increased value of *effort.*) If, in contrast, Player 2 decides to raise the contribution of Player 1's *special number,* then *the increased outcome* is [2(2)] + 1 - 1 = 4. (Note that the product in the square brackets [] is now the newly increased contribution of Player 1's *special number.*)

Of course, if Player 1 had chosen a level of *effort* equal to 3, instead, then, with her *special number* (2) and the realized *random bump* (-1), *the choice outcome* would be 1 + 3 - 1 = 3. If Player 2 had further chosen to increase the outcome by increasing the contribution of Player 1's *special number,* then *the increased outcome* would be 2(1) + 3 - 1 = 4. But if Player 2 had chosen to increase the contribution of Player 1's *effort,* then *the increased outcome* would be 1 + 2(3) - 1 = 6.

In addition to deciding how to increase the *choice outcome*, Player 2 also decides if she wants to give Player 1 a *bonus* - a special addition to Player 1's payoff in that round.

After *the increased outcome* is shown to Player 2 and Player 2's bonus decision is shown to Player 1, the round ends and the players proceed to the next round.

This completes the description of a single round of play. I will now describe how your payoff for the experiment will be calculated.

Payoffs

If you are participating in the role of Player 1, your payoff in a given round will depend on *the choice outcome* in that round (and so indirectly, on your *special number*, your *effort* level, and the realized *random bump*) but also directly on the chosen level of *effort* and on the decision of Player 2 you are matched with whether to give you a *bonus.*

Please look now at Table 1 on page 9 of these instructions. This table gives you the values of Player 1's payoffs for all possible values of your *special number*, your *effort* level, and the realized *random bump.* For your convenience we are reproducing a piece of this table in the text of these instructions. Please, turn back to page 6 of the instructions.

Special Number	Effort	Random Bump	Outcome	Bonus	No Bonus
		-1	1	**6.54**	**4.05**
	1	0	2	**8.44**	**6.54**
		1	3	**10.05**	**8.44**
		-1	2	**6.49**	**4.59**
1	2	0	3	**8.10**	**6.49**
		1	4	**9.52**	**8.10**
		-1	3	**6.15**	**4.54**
	3	0	4	**7.57**	**6.15**
		1	5	**8.85**	**7.57**

Suppose, for example, that in a given round, your *special number* was 1, your *effort* was 2, and the *random bump* was -1. You can see in the table above that the resulting choice outcome is 2. Suppose that Player 2 decided not to give you a *bonus* this round. You will find your payoff for this example by finding *special number* equal to 1 in the left-most column, *effort* equal to 2 in the column second from the left, and *random bump* equal to -1 in the third column from the left. Then, you will see in the right-most column of this row of Table 1 that your payoff for that round will be $4.59.

Suppose, however, that you are considering a higher level of *effort*, say 3. If the random bump happens to be same, -1, then the outcome will be 3. If the Player 2 decides to give you a *bonus* in this case, then your payoff in this round can be found by locating *special number* equal to 1 in the left-most column, *effort* equal to 3 in the second column from the left, *random bump* equal to -1, and then looking at the second to last column of this row, which shows a payoff of $6.15.

To give you further assistance in visualizing your choices as Player 1, we will also provide you the relevant payoff information on the screen as you are making your *effort* choices. This information will be equivalent to what you see in Table 1. Please look now at page 8 of this handout, which reproduces a screenshot similar to what you will see each round. The screenshot shows a question that we will ask Player 1 as a part of his *effort* choice: "What minimal outcome do you think Player 2 will demand to give you a bonus?" Then, for a given such outcome that you are specifying, the screen will show you what payoffs you may get with what probabilities (corresponding to different random bumps) given different available choices of *effort*.

If you are participating in the role of Player 2, your payoff in a given round will be equal to *the increased outcome* you obtained in that round – that is, it will depend on *the choice outcome* produced by Player 1 you are matched with (and so on Player 1's *special number*, her choice of *effort*, and the realized *random bump*), as well as on your decision on how to increase it.

Please look now at Table 2 on page 10 of the instructions where you can see how Player 2's payoffs are computed from *the choice outcome* and Player 2's decision how to increase it. Now, for example, suppose that in a given round, Player 1's *special number* was 2, she chose a level of *effort* equal to 1, and the value of the *random bump* was -1. If you chose to increase the outcome by increasing *effort*, then your payoff in that round is

$$2 + [2 \times 1] - 1 = \$3$$

In contrast, if you chose to increase the outcome by increasing Player 1's *special number*, then your payoff in that round is

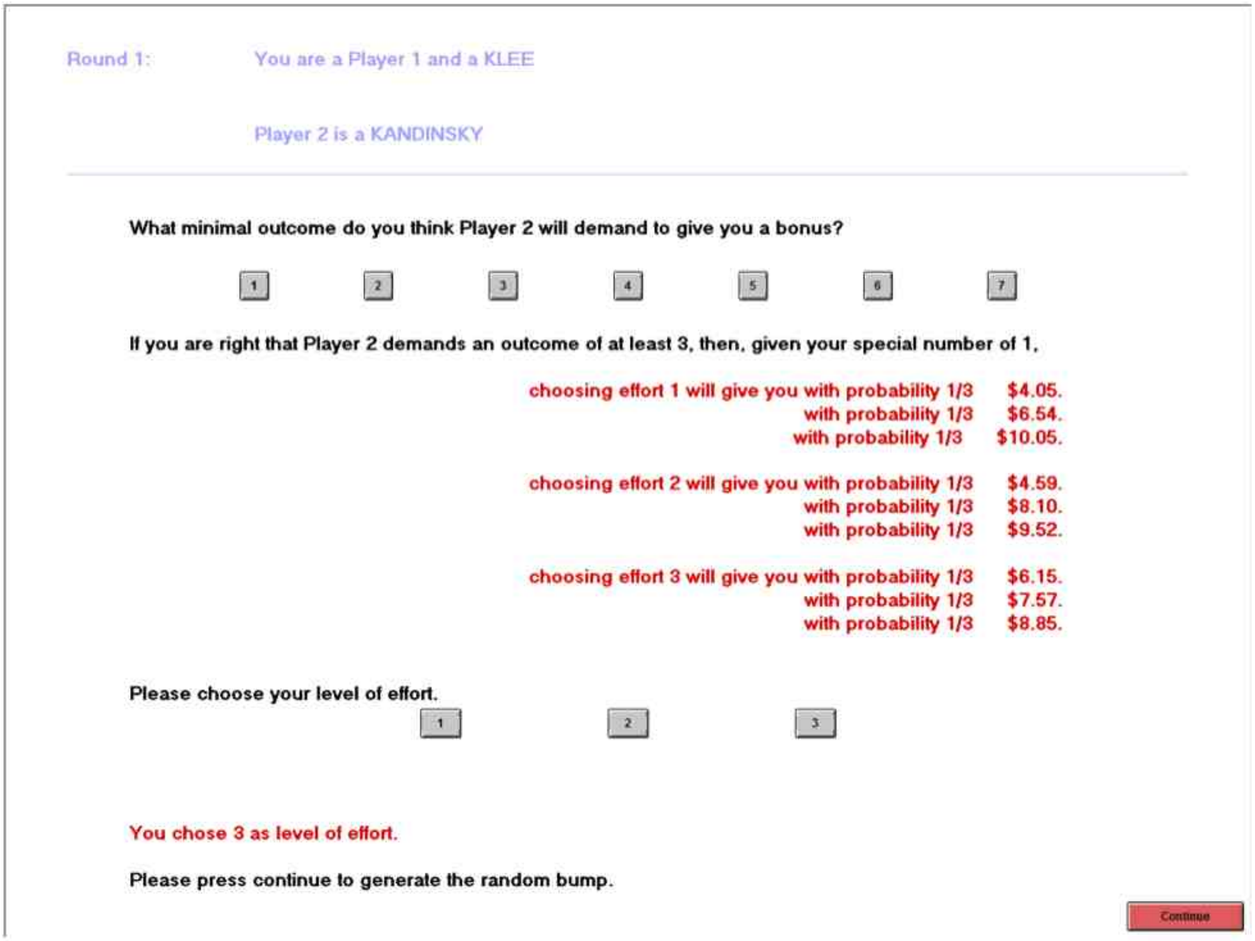
$$[2 \times 2] + 1 - 1 = \$4$$

You will see this by finding *special number* equal to 2 in the left-most column, *effort* equal to 1 in the second column from the left, and *random bump* equal to -1 in the third column from the left. The value in the same wow of the next column shows that the *the choice outcome* associated with this example is 2. The values in this row in the two columns on the right, then, tell you what *the increased outcome* and thus your payoff from this round as Player 2 will be. In case you decide to double *special number*, your payoff will be 4. In case you decide to increase *effort*, your payoff will be 3.

Again, your total payoff for the experiment will be the two highest round payoff from three randomly chosen rounds plus your payoffs from Part 1 of the experiment plus the show-up fee of $7.

If you have any questions, please ask them now.

Figure 1: Screen shot



Round 1: You are a Player 1 and a KLEE

Player 2 is a KANDINSKY

What minimal outcome do you think Player 2 will demand to give you a bonus?

If you are right that Player 2 demands an outcome of at least 3, then, given your special number of 1,

choosing effort 1 will give you with probability 1/3 $4.05.
with probability 1/3 $6.54.
with probability 1/3 $10.05.

choosing effort 2 will give you with probability 1/3 $4.59.
with probability 1/3 $8.10.
with probability 1/3 $9.52.

choosing effort 3 will give you with probability 1/3 $6.15.
with probability 1/3 $7.57.
with probability 1/3 $8.85.

Please choose your level of effort.

You chose 3 as level of effort.

Please press continue to generate the random bump.

Table 1: Player 1's round payoff

Special Number	Effort	Random Bump	Outcome	Bonus	No Bonus
1	1	-1	1	**6.54**	**4.05**
		0	2	**8.44**	**6.54**
		1	3	**10.05**	**8.44**
	2	-1	2	**6.49**	**4.59**
		0	3	**8.10**	**6.49**
		1	4	**9.52**	**8.10**
	3	-1	3	**6.15**	**4.54**
		0	4	**7.57**	**6.15**
		1	5	**8.85**	**7.57**
2	1	-1	2	**8.44**	**6.54**
		0	3	**10.05**	**8.44**
		1	4	**11.47**	**10.05**
	2	-1	3	**8.10**	**6.49**
		0	4	**9.52**	**8.10**
		1	5	**10.80**	**9.52**
	3	-1	4	**7.57**	**6.15**
		0	5	**8.85**	**7.57**
		1	6	**10.02**	**8.85**
3	1	-1	3	**10.05**	**8.44**
		0	4	**11.47**	**10.05**
		1	5	**12.57**	**11.47**
	2	-1	4	**9.52**	**8.10**
		0	5	**10.80**	**9.52**
		1	6	**11.97**	**10.80**
	3	-1	5	**8.85**	**7.57**
		0	6	**10.02**	**8.85**
		1	7	**11.12**	**10.02**

Table 2: Player 2's round payoff

				Increased Outcome when	
Special Number	**Effort**	**Random Bump**	**Outcome**	**Special Number Doubled**	**Effort Doubled**
1	1	-1	**1**	**2**	**2**
		0	**2**	**3**	**3**
		1	**3**	**4**	**4**
	2	-1	**2**	**3**	**4**
		0	**3**	**4**	**5**
		1	**4**	**5**	**6**
	3	-1	**3**	**4**	**6**
		0	**4**	**5**	**7**
		1	**5**	**6**	**8**
2	1	-1	**2**	**4**	**3**
		0	**3**	**5**	**4**
		1	**4**	**6**	**5**
	2	-1	**3**	**5**	**5**
		0	**4**	**6**	**6**
		1	**5**	**7**	**7**
	3	-1	**4**	**6**	**7**
		0	**5**	**7**	**8**
		1	**6**	**8**	**9**
3	1	-1	**3**	**6**	**4**
		0	**4**	**7**	**5**
		1	**5**	**8**	**6**
	2	-1	**4**	**7**	**6**
		0	**5**	**8**	**7**
		1	**6**	**9**	**8**
	3	-1	**5**	**8**	**8**
		0	**6**	**9**	**9**
		1	**7**	**10**	**10**