Discrimination in Strategic Settings[*]

Dominik Duell[†]and Dimitri Landa[‡]

July 19, 2017

Abstract

In a laboratory investigation of a principal-agent relationship with moral hazard, we analyse strategically induced discrimination. We isolate the influence of the strategic environment from the effects of other sources of discrimination, including group statistics and taste for discrimination. We find that, in a strategic setting, principals who reward agents contingent on agent-generated outcomes also attribute good outcomes more readily to effort when they share a social identity with the agent. No such bias emerges either for principals whose reward decisions are not contingent on outcomes or for the principals in a non-strategic environment. In the strategic setting, agents tend to anticipate lower demands from same-identity principals and condition their effort choice on that expectation. We argue that these choices are not driven by reciprocity but related to attitudes toward risk.

Keywords: Strategic discrimination, principal-agent relationship, reciprocity, risk-preferences
JEL: J7, J15, J24, D83, D84

[*] The research presented in this paper was supported by NSF Grant $SES-1124265. Support through the ANR - Labex IAST is also gratefully acknowledged. Previous versions of this paper were presented at the Political Economy Seminar at Columbia University, the Center of Experimental Social Science Colloquium, Nuffield College, Oxford; the Human Nature Lab at Yale University, the North-American ESA Conference, the EPSA General Conference, and the Annual MPSA Conference. We are especially grateful to Alessandra Casella and Catherine Hafer for the extensive comments and conversations.

[†] Lecturer, University of Essex

[‡] Associate Professor, New York University

1 Introduction

Ethnic, racial, gender, and other forms of social identity influence how the people inhabiting them respond to others and how others respond to them. At its most benign, this influence brings to the table enriching diversity, 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); and, the under-representation of women and minorities in the legislative bodies of most Western democracies (Paxton, Kunovich and Hughes, 2007; Iversen and Rosenbluth, 2008; Griffin and Newman, 2007, 2008).

Consider the following example that encapsulates the insidious nature of discrimination in 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 from Bob, is considering whether it is wiser to re-allocate some of her time elsewhere, or to increase her effort in the hope of impressing Bob. Bob, who suspects that Alice may be under-investing, is less likely to attribute a good outcome to her effort, and more likely to attribute it to Alice's good luck. In effect, then, the quality of outcome Alice needs to generate to obtain a promotion is higher than the quality of outcome other employees need. If, realizing this, Alice is discouraged and chooses to invest less, Bob's suspicions are confirmed. In that case, Bob's interpretation of outcomes and Alice's expectation of a tougher standard would both be correct and consistent with each other and with the actions supporting them.

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 good outcomes come from *out-group* actors like Alice, these Bobs will be more inclined to associate success with favourable circumstances rather than with Alices' effort.

The employer-employee relationship, however, is fundamentally strategic in that outcomes will depend not only on the actions by an employee but also on the employee's expectations of the feedback from the employer. When we observe asymmetric attribution in these settings, it may be the consequence of a disposition toward prejudice, but it may also reflect correct, if clearly regrettable, beliefs about differences in performance arising from strategic responses to the asymmetric beliefs and choices of others. While the economic theory of principal-agent relationship and statistical discrimination has understood that it does not take a psychologically driven misattribution to create and sustain stereotypes (Phelps, 1972; Arrow, 1973; Spence, 1973; 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 literature 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 aim of this paper is to address this challenge by evaluating the extent of *strategically driven discrimination* as distinct from (pure) *statistical discrimination.* We do so with a laboratory experiment designed to isolate the distinctly strategic effects of individuals' responses to sharing social identity in a principal-agent environment – effects that are independent of asymmetric group-based generalisations, either rationally (statistically) or psychologically sustained. Our primary analysis focuses on the behavioural comparisons across identity-based matches of principals and agents (with shared vs. unshared social identity between them) across strategic vs. non-strategic interactive settings.

Our findings suggest that the patterns of beliefs associated with the ultimate attribution error may emerge without group-specific feedback, as a fundamentally strategic phenomenon, though not necessarily one consistent with equilibrium play. In strategic environments, principals tend to attribute good outcomes, on average, more readily to their agents' effort when they share a social identity; in turn, in expectation of their principals' reward decisions, in-group agents tend to invest more into effort, re-enforcing their principals' beliefs. When principals' and agents' choices are not strategically co-dependent, the relationship disappears.

Principals' beliefs are related to their reward strategies with respect to the agents: both the principals who demand more from their in-group agents and those who demand less tend to expect higher effort in in-group matches than the principals with symmetric expectations across identity matches. How far are those expectations off the mark? Overall, we find that principals tend to do better at anticipating the choices of in-group than out-group agents. While the agents tend to respond in a way that is consistent with expectations of mutual reciprocity with in-group principals, subject-level evidence suggests that (1) responses may have more to do with agents' risk preferences than with a norm of reciprocity; and (2) the principals underestimate the possibility that agents in out-group matches increase their effort in response to their expectations of higher demands from the principals.

The conjunction of the systematic differences in beliefs on the part of principals playing distinct types of reward strategies and the treatment effect of the strategic environment in creating these differences reinforces the "strategic" interpretation of discrimination in our data. While prejudice-based discrimination and statistical discrimination based on knowledge of group differences are important aspects of discrimination, our evidence suggests the value of accounting for the distinctly strategic elements of discriminatory settings. Apart from providing a more precise picture of the sources of discrimination, such an account speaks directly to the public debate on the most effective ways of addressing discrimination, as it is often easier to adjust incentives within institutions that regulate strategic interactions than changing people's prejudice or their past experiences with different groups.

2 Discrimination: Variety and Identification

We analyse discrimination and prejudice in the relationship of delegation found naturally in the contexts of the labor-management relations. Discrimination refers to a practice of treating persons who perform *equally* in a physical or material sense *unequally* in a way that is related to an observable characteristic such as race, ethnicity, or gender.[1] Note that, as defined, discrimination may or may

[1] See Altonji and Blank (1999); Holzer and Neumark (2000) for a more detailed elaboration of this definition in a labor market context. Discrimination can be "positive" in the sense of in-group favouritism or "negative" meaning unfavourable treatment of the out-group.

not be rationally sustained; when and whether it is is one of the central questions of this study. Prejudice – a key psychological determinant of discrimination – is a faulty or inflexible generalisation about members of a group (Allport, 1954). Unlike discrimination, which may be rationalisable with a set of potentially correct beliefs, prejudice, as defined above, necessarily entails a mistake.

Taste for discrimination An influential theoretical approach to analysing the determinants of discrimination views it as resulting from a *taste for discriminating* against out-group members (Becker, 1971; 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. A somewhat different version of this mechanism can be found in the social psychology work that ties prejudice, and the discrimination to which it may give rise, to a predisposition toward a particular kind of bias known as the *ultimate attribution error* (Pettigrew, 1979). This error manifests when individuals are biased in their attribution of positive outcomes to a disposition (an attitude or choice) when judging in-group members, but to a non-choice attribute (such as luck) when judging out-group members.[2]

Statistical discrimination In contrast to the taste for discrimination, *statistical discrimination* is expected 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) endogenises 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, both observational and experimental, report evidence consistent with statistical discrimination. Recent waves of audition studies (Bendick, 2007; Bertrand and Mullainathan, 2004; Goldin and Rouse, 2000) and "hit-rate" analyses (Knowles, Persico and Todd,

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

2001; Persico, 2002; Persico and Todd, 2006; Persico, 2009; Coviello and Persico, 2013) use sophisticated research designs to detect race and gender-based discrimination in the labor market, leaving aside the question of what drives discrimination in principal-agent settings. Falk and Zehnder (2007) show that players' behaviour in the trust game between city residents depended on the reputation of the city districts from which their partners were drawn, thus instantiating Phelpsian statistical discrimination, albeit in a strategic environment.

Strategic discrimination The idea of discrimination as a *specifically strategic* equilibrium phenomenon – the Arrovian version of statistical discrimination – has informed some of the important contributions to the debates regarding the desirability of policy interventions such as affirmative action programs. Such policy interventions can induce differences in employers' beliefs about how much effort members of different social identity groups exert, prompting the employers to discriminate. The resulting discrimination reduces incentives for members of the disadvantaged group to invest, creating a self-fulfilling prophecy (Loury, 1976; Coate and Loury, 1993).[3] Studies of the supply-side labor market document behaviour consistent with the strategic expectations at the core of the Arrovian approach. Pre-labor market discrimination has been shown to affect human capital investment of future generations and, in so doing, arguably to 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).[4]

Still, while these studies go some distance in distinguishing the taste for discrimination and the statistical discrimination mechanisms, they do not offer a clean test of the Arrovian strategic theory of statistical discrimination. The agents' rational (strategic) responses to discrimination are consistent with the possibility that what drives principals' discriminatory choices are psychological, taste-for-discrimination factors that may have little to do with statistical discrimination, let alone

[3] In this sense, employers allocating members of the social group with higher average qualifications to higher-skill jobs may be making ex ante optimal economic decisions (Lundberg and Startz, 1983; Lundberg, 1991). An opposite conclusion associated with strategically induced asymmetric beliefs about the behaviour 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' efforts on behalf of their constituents. Voters should, all else equal, prefer an out-group candidate because she will work to earn the electoral support that an in-group candidate will take for granted (Swain, 1993; Landa and Duell, 2015).

[4] 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).

an Arrovian, strategic version of it. A further challenge is that, when empirically evaluating the behaviour of principals (employers/managers), the predictions of the Phelpsian and the Arrovian versions of statistical discrimination are particularly hard to distinguish because the strategic response needs to be distinguished from the response to the differences in the group statistics that are typically part of the empirical background of the specific principal-agent interaction analysed.

To distinguish between these versions of statistical discrimination, a controlled laboratory environment may be particularly useful. Three previous laboratory studies seeking to capture different elements of the Arrovian strategic discrimination story are especially relevant to and provide an important comparative context for the present analysis. In the first of these studies, Fershtman and Gneezy (2001) provide evidence of differences in attribution in interactions with a strategic component (modeled as a trust game) and without it (modeled as a dictator game), but find that senders' stereotype-driven beliefs in the trust game are inconsistent with the return decisions, which do not vary with the group identity conditions. These results contrast with ours in the payoff-relevance of subjects' beliefs about the strategic play of their partners in the game. Because the payoffs on the receiver's choice in the trust game are independent of whatever beliefs she may have about the sender, the receiver has no affirmative reason in the lab to act on the stereotypes, whether senders' or her own. To capture the influence of a strategic environment on the possibility of a discriminatory action by the sender in response to the receiver's anticipation of discrimination against her, the experimental design needs to contain strategic feedback both before and after the receivers' choices (in our principal-agent setting, before and after the agents' effort choices).

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 (by facing higher costs of investment), and that discrimination persists, albeit non-linearly and non-monotonically, even after the differential in investment costs is removed. The design in Haan, Offerman and Sloof (2015) departs from that of Fryer et al. by endowing groups with the same investment costs but building into the treatment the exposure to the realised history of (apparently random) asymmetric play choices. Like Fryer et al., they find that the group that was less likely to be hired at the same quality signal then also invests less into the acquisition of quality. The feedback about group differences, like the public revelation of asymmetries in Fryer, Goeree and Holt (2005), creates and reinforces the stability of the pairing of low investment and low hiring rates.[5] In both studies, the publicity of the initial asymmetries is key, and that raises the question about the extent to which the strategic actions they report remain anchored in the seeded population statistics, or, to put it differently, in the distinctly Phelpsian framework of statistical discrimination. A further concern is that the initial asymmetries in the history of play that may encourage the manifestation of subjects' biases for consistency or against additional cognitive effort when exogeneously induced statistical reasons for such a history are no longer present.

The design of our study obviates both of these concerns by avoiding the seeding of discrimination with either the asymmetric group-level parameters or the asymmetries in the history of play. Such seeding naturally pushes in the direction of developing correct interpretations (attributions) of the determinants of payoff-relevant outcomes, whereas one of our key goals is to study how such interpretations develop endogenously. Further, whereas in Fryer et al. and Hahn et al. studies the principals are not endowed with distinct social identities, and so the observed discrimination in principals' choices is not driven by their attitudes to identity-driven relationships, both the principals and the agents in our study are assigned social identities that frame their relationship, allowing interpretations of outcomes to arise endogenously entirely in response to beliefs about the consequences of shared vs. unshared social identities – the mechanism at the core of Arrovian strategic statistical discrimination.[6]

Identifying Arrovian Statistical Discrimination in a Principal-Agent Environment We highlight three features of our experimental design that allow for the identification of strategic discrimination:

First, to get a handle on the size of the strategic effect, in contrast to the psychological one, we create counterfactual environments that make separation between these effects possible. We do so (1) by comparing the beliefs of principals whose punishment/reward strategies are constant in out-

[5] In a study currently in preparation, Hopfensitz, Reuben, and Rott (2016) show that any kind of exogenously set group stereotypes can easily be induced in a laboratory setting and generate vastly distinct behaviour across groups even if some of the stereotypes are in stark contrast to those individuals usually hold outside the laboratory (i.e., gender differences in skills).

[6] In earlier work, Jin, Yamagishi and Kiyonari (1996) demonstrate that in-group favouritism only arises when common-knowledge about group membership exist and not by mere assignment to groups and Yamagishi and Kiyonari (2000) show that expectations arising in a sequential prisoners dilemma game give room for identity-contingent behaviour, which is not present in a simultaneous version of the game. Both studies hint at a link between self-sustaining in-group bias in beliefs and behavior in distinct types of strategic environments.

come to those of *incentivising* principals whose punishment/reward decisions vary with the observed outcomes; and (2) by comparing the principals' beliefs in a treatment that implements a strategic environment to those in a corresponding non-strategic environment. The strategic environment has two-sided feedback, allowing the agents to condition their effort choices on their expectations of the principals' reward decisions and the principals to condition their reports of beliefs about agents' choices on their expectations of how agents likely evaluated their own expectations of being rewarded. This creates the possibility of mutually consistent identity-contingent incentivised beliefs. We also create payoff incentives for the agents that lead to pooling effort choices – low-quality types invest highly into effort, high-quality types invest little, which are constant across strategic and non-strategic treatment; in other words, differences between treatments must be driven by induced "strategicness" and not variation in the underlying payoff structure.

Second, to further separate strategically driven belief asymmetries from the non-strategic (Phelpsian) statistical belief asymmetries, we adopt a design that does not pre-treat subjects with reputations of social groups. In particular, we induce artificial group identities in a treatment related to the "minimal group paradigm" (Tajfel and Turner, 1986) – an approach to inducing a (weak) notion of identity that is seemingly unrelated to the behaviour of interest – and provide minimal feedback to subjects in the course of play. This approach also helps advance our overall goal of isolating the beliefs-driven determinants of strategic discrimination from the influence of other elements of the social environment that in real-world settings may also affect willingness to discriminate, e.g., reputation costs for discrimination.[7]

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

[7] We analyse the effects of some of these elements in a companion work, where we study the effects of policies aimed to limit discrimination.

3 A simple model of principal-agent relationships

3.1 Set-up

We capture the underlying strategic principal-agent relationship in a simple model of incomplete contracting. 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 to be uniform on that support. The agent chooses her effort level, $e \in \{1, 2, 3\}$, which is costly to herself. The outcome F is given by

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

where the noise, ω, is a random draw from a uniform distribution on $\{-1, 0, 1\}$. The payoffs of both the principal and the agent depend on F, though in different ways. The principal observes F and then decides whether to give a bonus, b, to the agent. The agent's payoff is given by $G(F, b, e)$, where

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

$G(\cdot)$ is, thus, increasing in F and b and decreasing in e.

The principal's payoff is F plus either t or e. The principal herself chooses whether she wants to double the t or e component in her payoff, but she must make her choice without directly observing t or e; she observes only F. 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. Note, D may be interpreted as principal's belief whether an outcome can be attributed more to the effort or type component. The game ends when these payoffs are realised.

3.2 Equilibria

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, which are the standard predictions in such games (Persson and

Tabellini, 2000; Bueno de Mesquita and Landa, 2015), the principal chooses to reward if and only if $F \geq z$, $z \in \{3, 4, 5\}$, and the agent chooses level of effort e^* such that $e^* + t = 4$. Thus, the agent of type 1 chooses effort 3, agent of type 2 chooses effort level 2, and agent of type 3 chooses effort level 1.[8] 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.

One can construct equilibria in which the threshold for receiving a bonus is $z \in \{1, 2, 6, 7\}$. Those equilibria are semi-separating, in that the principal's posterior beliefs about the agent's type are not uniform, and there is a critical value in the $\hat{F}$ space such that the principal will prefer to double type for $F > \hat{F}$ and effort for $F < \hat{F}$. Both in these semi-separating equilibria and in the pooling equilibria described above, the principals' choices are contingent on the outcomes they observe. We will, thus, refer to these equilibria 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.

We will call principals whose strategies call for rewarding outcomes meeting a threshold $F \geq z$, $z \in \{2, 3, 4, 5, 6, 7\}$ and not rewarding otherwise as *incentivising principals* and their strategies as *incentivising strategies.* Thus, principals in all OCP equilibria are incentivising principals and in all ONCP are not. The intuition is that incentivising strategies may create incentives for the forward-looking agents to invest into effort, whereas the strategies of always- and never-rewarding would clearly not.

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

[8] As is standard, these effort predictions are for agents endowed with the model payoffs in the lab. However, note that in the implemented game, subjects face two kinds of uncertainty: about the realised noise draw and the strategic uncertainty about principals' critical outcome thresholds for rewarding the agents. This means that the actual choices of our subjects in the role of agents may be contingent on their expectations of outcomes, and rewards, and reflect their underlying unmodeled risk preferences. We explore the relationship between the agents' choices and their risk attitudes in Section 5.2.4.

Note that the baseline game described above does not assign identities to the players. In the identity treatments of the experiment, we prime and reveal to group members their social identities by fixing labels to principals and agents and making them common knowledge within the pairs but do not alter the payoff structure described above. Because the payoff structure does not depend on these identities, one equilibrium behavioural expectation is that identity has no effect on behaviour. However, because players observe social identity matches, they may choose identity-contingent equilibria in which different equilibrium profiles are played in different identity matches (e.g., an OCP equilibrium profile with higher (lower) threshold for reward in in-group matches and an OCP equilibrium profile with lower (higher) threshold for reward in out-group matches). In this way, identity matches could matter as selectors of different equilibrium profiles.

The multiplicity of equilibria creates a strategic coordination problem for the players. The presence of this problem is an intentional feature of our design. The rationale is two-fold. First, contractual uncertainty of reward and promotion expectations is a wide-spread feature of empirical environments with incomplete contracts, and, in particular, of environments in which discrimination is typically reported. One of our primary goals is to understand how the players behave in environments of precisely this kind. Second, allowing the players to take auxiliary actions that can reduce uncertainty over mutual expectations (e.g., making cheap-talk announcements before or in the middle of play) can have a separate psychological self-committing effect that is distinct from the purely informational coordination effect, altering what we think is the standard baseline behaviour in such settings.

4 Experimental design

The structure of our laboratory experiment approximates the principal-agent relationship between an employer and an employee. We implement a principal-agent interaction in one main and two supporting treatments: the main – STRATEGIC – treatment features induced groups and the opportunity to reward the agent with a bonus (henceforth, referred to as the availability of the sanctioning device), following closely the model described above. The NON-IDENTITY treatment takes away the existence of induced group identities and the NON-STRATEGIC treatment removes the sanctioning device. Our experiment included 188 subjects, 94 in the role of a principal and 94 in the role of an agent, generating 3760 subject-round observations (see Table 1).

Table 1: Overview over experimental treatments, number of subjects (N), and number of subject-round observations (n)

	Identity	**No identity**
With sanctioning	STRATEGIC (N=110, n=2200)	NON-IDENTITY (N=38, n=760)
Without sanctioning	NON-STRATEGIC (N=40, n=800)	

In all treatments, 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).[9] Subjects are assigned to the role of either an *agent* (called "Player 1") or a *principal* ("Player 2") at the beginning of each session and remain in that role for the duration of the experiment. They are, then, randomly re-matched into pairs of one agent and one principal in each of 20 rounds of a session. The implemented random matching protocol is the perfect stranger matching for the first (number of subjects in the session)/2-rounds of each session, followed, in subsequent rounds, by subjects meeting previous matches again in random order once. The conjunction of the matching protocol and the anonymised interaction between subjects precludes direct exchanges between subjects. This design provides us approximately with as many independent observations as subjects in the experiment (94 agents and 94 principals distributed across 10 sessions) but learning within-subject may occur.

Group identity inducement At the beginning of each session of both the STRATEGIC and the NON-STRATEGIC treatments, subjects were assigned to groups according to their stated preferences for either *Klee* or *Kandinsky* paintings[10] and performed in a quiz collaboratively with their new fellow painter group members. Members of both groups, Klees and Kandinskys, in all

[9] Since we do this in each session and treatment condition in the experiment, treatment effects should not be affected.

[10] See Tajfel and Billig (1974), Chen and Li (2009), and Landa and Duell (2015) for the use of painter-preferences to induce identities. More detailed information on the experimental protocol can be found in Section B of the Supporting Information (SI).

treatments performed approximately equally well. In the subsequent principal-agent game part of the experimental session, the identities of both subjects within a matched pair were displayed for them on the screen along with icon-sized paintings by the corresponding artists. In this way, subjects learn whether they are in an *in-group* or *out-group* match.[11]

Principal-agent game with sanctioning device The game simulated in the lab, in both the STRATEGIC and NON-IDENTITY treatments, mirrors exactly the structure and payoffs laid out in Section 3. By monetarily incentivising subjects in the role of agents, we create concerns about outcomes because agents value receiving a bonus from the principals. Subjects in the role of principals benefit from high outcomes and thus may want to incentivise agents to exert high investment into effort. While the principal does not bear a direct cost of awarding the bonus, the principal's payoff, given that agent plays a best-response, varies with principal's bonus-awarding strategy. Given that the principal is facing no commitment problem in following through on her choice of that strategy, that choice effectively determines her payoff. In this way, the principal's bonus-awarding choice is incentivised across rationalisable strategy profiles, as is standard in moral hazard settings. The absence of a direct cost to the principal reduces exchange-based incentives central to the standard gift-exchange games,[12] allowing us to focus on how principals use the bonus to incentivise the agents or to express their preference for discrimination. As will become apparent in the analysis of our results, agents in our experiment clearly respond to their expectations of principals' demands.

Additionally, all subjects, both 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.[13] Before agents make their investment decision but after they observe their randomly assigned type, they are asked: "What minimal outcome do you think Player 2 will demand to give you a bonus?" They are shown payoffs, contingent on their answer and their type, as a function of the level of effort they may choose and the possible values of noise. Agents may

[11] Considerable experimental literature has shown the effectiveness of the minimal group paradigm in inducing the patterns of responses to identity that resemble those usually observed outside the laboratory with naturally occurring group identities (Eckel and Grossman, 2005; Chen and Li, 2009; Goette, Huffman and Meier, 2006; Bernhard, Fehr and Fischbacher, 2006; McLeish and Oxoby, 2007). 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.

[12] See Akerlof (1982); Fehr et al. (1998); Fehr, Kirchsteiger and Riedl (1998); Charness and Haruvy (2002); Charness (2004).

[13] Instructions are shown in Section B.3 of the SI.

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. The information about agents' expected minimal rewarded outcomes that they enter to obtain their contingent payoffs is understood to enable a more highly rewarded choice, and, as such, is indirectly incentivised monetarily. We collect this unobtrusively elicited information about the agents' beliefs and use in our analysis below.[14]

Principal-agent game without sanctioning device The NON-STRATEGIC treatment, similar to the STRATEGIC treatment, induces group identities but replaces the principals' sanctioning tool with exogeneously given incentives to the agents. In this treatment, agents' payoffs are given by $G(F, e) = \beta\sqrt{F} - e$, with $\beta = 4$. Note that, as in the STRATEGIC treatment, $G(\cdot)$ is increasing in outcome F and decreasing in effort e. The functional form of the payoffs and the parametrisation were chosen to be as close as possible to those in the STRATEGIC treatment and to induce optimal choices for agents, conditional on their type, that are identical to the optimal choices in the maximal principal welfare (3-4-5 threshold) outcome-contingent play equilibria in the STRATEGIC treatment game. In this game without sanctioning device, as in the game with sanctioning device, the optimal response for an agent is to choose $e^* = 4 - t$.

In the non-strategic environment, whatever asymmetry in beliefs is observed must be due to psychological, taste-for-discrimination factors like the ultimate attribution error. Using that behaviour as a baseline, we can interpret the behavioural 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.

A quick aside on an alternative treatment design: One could get at the difference between strategic and non-strategic settings with a design that randomly assigns a probability with which the sanctioning device is available rather than, as we do, taking it away altogether and exogeneously adjusting payoffs. The downside of that alternative in our setting is that a low probability of being rewarded/punished, which would be the case in the non-strategic setting, would imply that agents' effort would approach the minimal possible level, undermining the variation in the principal's beliefs

[14] More specifically, we capture agents' beliefs by recording the mean expected demands of all clicks they make in each round. Section A.4.2 in the Supporting information (SI) gives more data on frequency and extent of agents' use of this tool.

on whether an agent chose high or low effort and rendering largely meaningless our instrument eliciting them. Mindful of that problem, we opt instead, for the design that experimentally creates a strategic interaction, with its entailed necessity to form beliefs about others' beliefs and behaviours, and then juxtaposes that interaction with one that retains the relationship between agents' optimal effort and type but without incentives to engage in strategic calculation that frames the strategic interaction.

Summary of the experimental set-up At the beginning of each session in the STRATEGIC and NON-STRATEGIC treatments, subjects' identities are induced. In all treatments, then, subjects are randomly assigned to the role of either agent or principal and the sequence of moves in each round of the experiment is as follows:

1. Agents are assigned a *type* and privately informed about its realisation (1, 2, or 3).
2. Agents choose a level of *effort* (1, 2, or 3) and state their expectation about which minimal outcome principals demand to see to give a bonus (1-7, *expected demands* – agents' beliefs).
3. *Noise* and *outcome* are realised where the value of *outcome* is the sum of agent's *type* (1, 2, or 3), agent's chosen level of *effort* (1, 2, or 3), and a *noise* realization (-1, 0, or 1).
4. Principals learn the value of *outcome* (1-7).
5. Principals choose whether to attribute outcomes to *type* or *effort* (*attribution decision* – principals' beliefs) by doubling the payoff contribution of the *type* or *effort* component of *outcome*[15] and whether to give the agent a bonus *(reward decision)*.[16]
6. Round feedback: Principals observe whether type or effort was higher and agents learn principal's reward decision.

5 Results

We first present results on *aggregate behaviour* in general and in relation to the equilibrium predictions for the STRATEGIC treatment in particular (Section 5.1). These findings, however, are subject to interpretive challenges that require the assessment of subjects' choices in relation to their

[15] We elicit principals' beliefs about the relationship between effort and type rather than about the absolute values of effort and type; the former is less obtrusive and speaks directly to the idea of attribution bias that is central to our study. Also, we are interested in relative values in different treatments, not having an option to hedge does not bias our results.

[16] Principals' *reward decision* and the elicitation of agents' *expected demands* are omitted in the NON-STRATEGIC treatment. On the screen where principals make reward and attribution decision, we also asked principals whether they thought type or effort was the higher quantity. The correlation between *attribution decision* and the guess whether type or effort is higher is .74 ($p = .00$).

beliefs about their matched partners. To meet these challenges, which is one of the primary goals of our analysis, we investigate the relationships between principals' and agents' actions and beliefs at the *subject-level* in Section 5.2. There, we analyse behaviour of two sets of principals whose strategies suggest very different best responses from the agents: principals whose punishment/reward choices are constant in outcome and those whose choices vary with the observed outcomes. The comparison of respective attribution choices of these sets of principals gives us our first measure of the effect of strategic considerations on discrimination. We, then, look at the *average treatment effect* of the NON-IDENTITY treatment in Section 5.3.1, to explore what better accounts for the contrasts revealed in the subject-level comparisons, in-group favouritism or out-group discrimination. Finally, we compare the principals' beliefs in the STRATEGIC treatment to those in the NON-STRATEGIC treatment (Section 5.3.2) to get at our second measure of the effect of a strategic expectations on propensity to discriminate.[17]

5.1 Average treatment effects, take 1: aggregate behaviour and equilibrium predictions

In the aggregate, agents' level of effort decreases with type. The marginal effect of type on effort is $-.18\,(-.25, .10)$ in the STRATEGIC treatment, $-.28\,(-.44, -.13)$ in the NON-IDENTITY treatment, and $-.52\,(-.70, -.34)$ in the NON-STRATEGIC treatment (95% bootstrapped confidence bounds clustering on the subject-level are reported in parentheses throughout). Figure 1.A clearly shows this downward trend. Note that in our main, STRATEGIC, treatment this leads to a distribution of outcomes centered at 4, with 75% of observations falling within the range between 3 and 5. The figure also shows that principals in the NON-STRATEGIC treatment enjoy higher levels of effort from the agents than in the strategic settings, with or without identity (Section A.7 in the SI provides further details). The reason is, likely, broad strategic uncertainty in the strategic interactions, and we see comparable drop-offs in effort from in- and out-group agents consistent with that view.

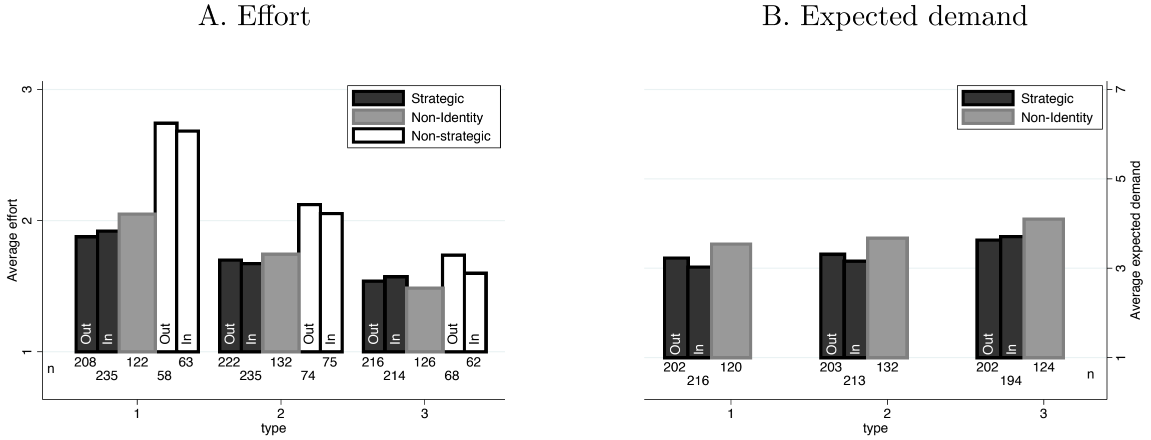
As the two panels of Figure 1 show, at this level of analysis, we find no difference in agents' effort or expected demands between in- and out-group matches.[18] The presence of the small but significant

[17] Summary statistics for the variables of interest in identity inducement stage and principal-agent game for all treatments are given in Section A of the SI.

[18] Marginal effects are estimated from the regression of effort shown in Table A.3 in the SI. The null effect of *in-*

increase in expected demand with assigned type (the marginal effect of type on expected demand is .23 (.05, .41) in the STRATEGIC treatment and .27 (.09, .47) in the NON-IDENTITY treatment) suggests the desirability of controlling for the interaction *type*×*expected demand* throughout our analysis.[19]

Figure 1: Agents' average effort (panel A) and average expected demand (panel B) over type for each treatment and by in-group status of the matched principal (for strategic and NON-STRATEGIC treatment). Averages displayed for in- and out-group matches separately in the STRATEGIC and NON-STRATEGIC treatments. The number of subject-round observations (n) is given below the bars.



A. Effort

B. Expected demand

Turning to principals' behaviour, we observe an overall increase in the rate with which agents are rewarded with increasing outcomes (Figure 2). The marginal effect of outcome on rate rewarded is .07 (.03, .11) in in-group matches and .10 (.06, .13) in out-group matches in the STRATEGIC treatment, as well as .06 (.01, .11) in the NON-IDENTITY treatment.[20] The figure also shows the principals' greater willingness to reward in-group rather than out-group agents. The marginal effect of in-group on probability of reward, holding outcome at its mean, is .11 (.02, .23) in the STRATEGIC treatment. In fact, absolute differences in rates of reward by principals in in- and out-

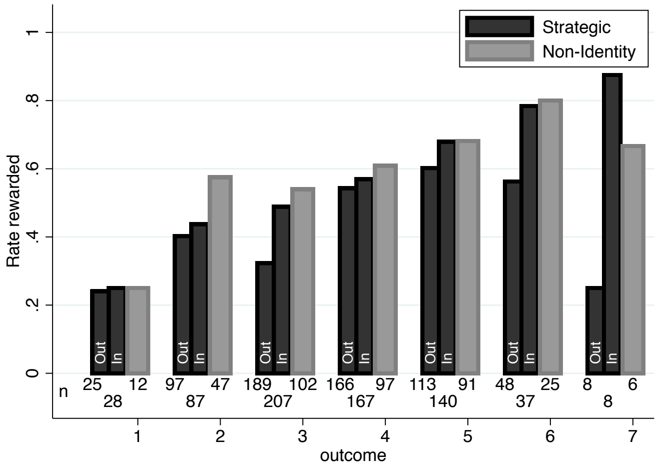
group on effort is indicated by the statistical insignificance associated with the coefficient on type and the interaction *type*×*treatment* in the same regression and the null result on expected demand by the insignificance of *in-group* in the regression of expected demand shown in Table A.4 in the SI.

[19] Behavior in the NON-STRATEGIC treatment provides a robustness check on whether agents are correctly incentivised by the differences in monetary payoffs offered. Subjects in the NON-STRATEGIC treatment show almost perfect pooling behaviour, 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. There is not difference in agents' average effort choices in in- and out-group matches (2.11 vs 2.17, difference = .06 (−.09, .21))

[20] Marginal effects are estimated from the regression of reward decision shown in Table A.5 in the SI.

group matches are substantial at both outcomes lower than 4, .48 vs .36, difference $= .13\,(.03, .23)$, and outcomes higher than 4, .75 vs. .60, difference $= .15\,(.04, .26)$, but become smaller at an outcome of 4, difference $= .03, (-.09, .15)$.

Figure 2: Principals' rate rewarded over outcome in STRATEGIC and NON-IDENTITY treatments. Rate displayed for in- and out-group matches separately in the STRATEGIC treatment. The number of subject-round observations (n) is given below the bars.



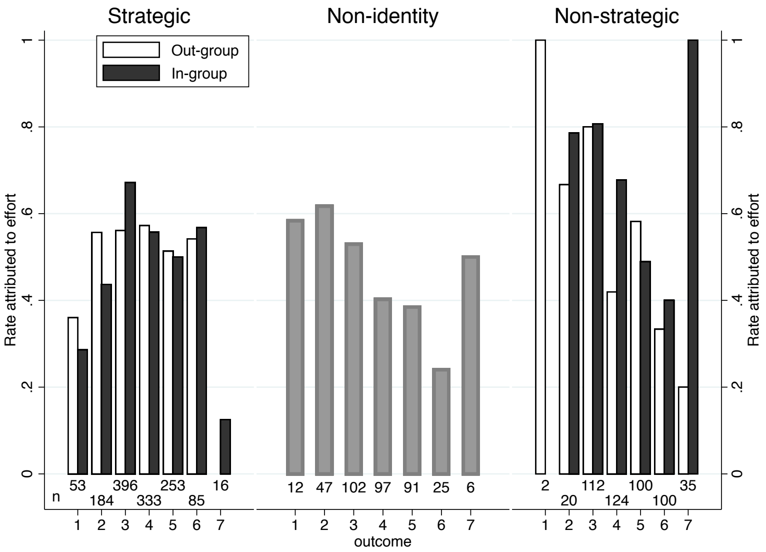
We summarise these aggregate-level findings for the applicable treatments in the following two results:

Result 1 *Subjects' behaviour is consistent with key features of outcome-contingent play: (a) principals' reward choices are systematically increasing in observed outcome, and (b) agents' effort is decreasing with type.*

Result 2 *(a) The average rate of bonus awarding by the principals is significantly higher in in-group matches than in out-group matches. (b) On average, agents' and principals' beliefs and agents' effort levels are not systematically different across in-group and out-group matches.*[21]

[21] Note that, by construction of our treatments, parts (a) and (b) are relevant to the STRATEGIC treatment, and parts (b) other than agents' beliefs also to the NON-STRATEGIC treatment.

Figure 3: Rates of outcome attributed by principals to effort over outcome for each treatment. Rate displayed for in- and out-group matches separately in the STRATEGIC and NON-STRATEGIC treatments. The number of subject-round observations (n) is given below the bars.



Turning next to the evidence in attribution choices by the principals, we find a significant difference in the average rates by which outcomes are attributed to effort and not type between STRATEGIC and NON-STRATEGIC treatments for low outcomes (3 and below), $-.24\,(-.36, .12)$, but not for high outcomes (4 and above), $.01\,(-.12, .15)$. The effect of displaying identity markers (STRATEGIC vs. NON-IDENTITY treatments) is positive and significant for high, $.15\,(.00, .30)$, but not significant for low outcomes, $.00\,(-.15, .15)$ (see Figure 3).[22] Comparing averages correctness of attribution across treatments, we find no significant differences (the rates are .34, .36, and .36 in the STRATEGIC, NON-STRATEGIC, and the NON-IDENTITY treatments, respectively (the fuller analysis of attribution decisions is below).[23] These results suggest that the strategic considerations in the presence of social identity may, indeed, affect principals' attribution choices, but the evidence is mixed, and interpreting these *average* attribution values is difficult because the

[22] Results are similar when we look at outcomes below 3 vs. 3 and above, or below 5 vs. 5 and above. Table A.6 in the SI provides regression-based evidence that shows a potential (conditional on outcome) treatment effect but no average difference in attribution decisions between in- and out-group matches. The coefficient on treatment and the marginal effect of treatment are statistically significant in a regression on observations in STRATEGIC vs. NON-STRATEGIC treatments and controls; marginal effects of treatment condition NON-STRATEGIC (vs. STRATEGIC) is $.10(.00, .20)$. There is no significant NON-IDENTITY average treatment effect.

[23] In conjunction with the comparison of agents' effort choices, the correctness of attribution underlies the comparison of realized principals' payoffs between treatments, suggesting that they do best in the NON-STRATEGIC treatment, with the other two treatments being close to one another. The contrast between variations in agents' effort and in principals' attribution across treatments suggests that the effects of the differences in strategic uncertainty appear greater for the agents than principals.

beliefs about effort they represent must depend on the incentives that the corresponding principals expect to be inducing with their bonus-awarding strategies. Because those strategies differ across principals, the aggregate results may conceal the true relationships between award decisions and attributions. A similar challenge arises with respect to the interpretation of the aggregate results on agents' actions and beliefs.

A further reason for interpretive caution is suggested by the comparison of attribution data panels in Figure 3. A casual glance at the figure suggests that while the patterns of attribution decisions in the NON-IDENTITY and in NON-STRATEGIC treatments are quite similar, with attributions to effort decreasing in outcome, the pattern in the STRATEGIC treatment is clearly different, suggesting different underlying mechanisms and the value of a disaggregated analysis that could reveal them. To take proper account of subjects' decisions, we next turn to the subject-level analysis that anchors principals' attribution and agents' effort choices in their respective expectations about each other's (likely) choices. As we will see from that analysis, the aggregate values described above conceal systematic variations in subjects' beliefs and in their belief-contingent strategies, and with them, considerable evidence of strategically driven discrimination.

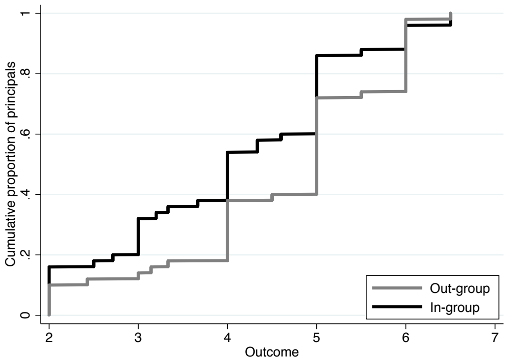
5.2 Subject-level Analysis

5.2.1 Principals

A key element of our approach in the subject-level analysis is connecting subjects' observable action choices to their elicited beliefs. We start by distinguishing between two straightforwardly distinct behavioural groups of principals: those whose bonus-awarding strategies are contingent on the received outcomes and those whose strategies are not. Recall that these two types of strategies correspond to two types of equilibria in the game, OCP equilibria and ONCP equilibria. In the context of the best-responses to these strategies by the agents, we refer to the principals playing the former (outcome-contingent) strategies as incentivising, and to those playing the latter (outcome-independent) strategies as non-incentivising.

Identifying incentivising principals and their in-group biases in reward decisions In the STRATEGIC treatment sample, incentivising principals constitute 76% of the principals, their behaviour as such is likely to generalize to other draws from the laboratory subject pool, and accounting for it is of first order of importance.[24] What constitutes *good* performance in the eyes of incentivising principals, however, is not well defined ex ante. Different outcome thresholds in the 3-5 range are consistent with OCP equilibria that, in our model, maximize the agents' effort. This means that even restricting attention to these equilibria, principals' attribution decisions may be driven by attribution biases that would be "canceling" each other at any exogenously fixed level of performance in that range. To get a valid measure of attribution bias, we need to evaluate attributions at the thresholds of good/bad performance that are subject-specific. To get a measure of such thresholds for each of the incentivising principals, we compute the individual-specific threshold values of outcome that minimize errors in categorizing their respective reward decisions.[25] These threshold values provide natural individual-specific definitions of what outcomes a given principal perceives as good performance (at and above the threshold) as opposed to bad performance (below the threshold).

Figure 4: Cumulative distribution of incentivising principals' outcome thresholds above which they are willing to reward the agent in the STRATEGIC treatment.



The inferred principal-specific reward thresholds vary from 2 to 7. Figure 4 gives the cumulative distribution of those thresholds. The average threshold is lower (3.96) in in-group matches than in out-group matches (4.53), implying that incentivising principals are less demanding in the in-group matches than in out-group matches; the difference in means is $-.57\,(-1.11, .03)$ and the p-value in the associated equality of distribution test (Wilcoxon) is smaller than .05. To reward, 50% of

[24] 11 out of 55 principals give a bonus in every round and 2 principals never give a bonus in any round; this leaves 42 principals whose reward choices vary with outcome.

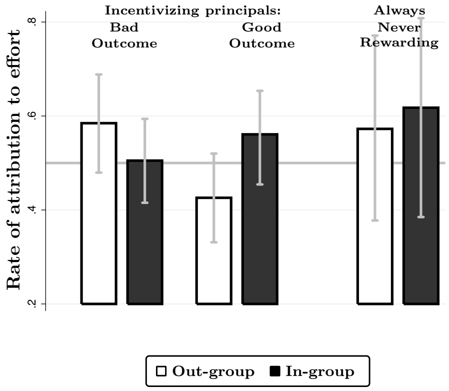
[25] The average reward decisions incorrectly classified with the error-minimizing threshold is .23 and .22 in- and out-group matches, respectively. This suggests that principals mostly behave consistently with their inferred individual thresholds in their reward decisions.

incentivising principals demand to see lower outcomes from in-group agents than from the out-group agents, while a significantly smaller share, 20%, demand the reverse. To summarise:

Result 3 *The bulk of principals in the STRATEGIC treatment play incentivising reward strategies. Among these incentivising principals, significantly more demand higher outcomes for rewarding out-group than in-group agents.*

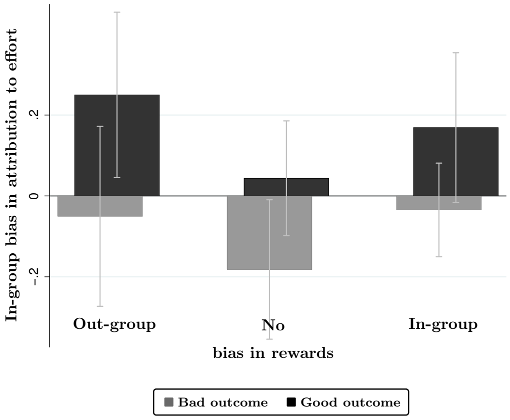
Principals' attribution choices vs. reward biases We next turn to describing the incentivising principals' attribution choices relative to their reward decisions. Figure 5 displays the average attribution of outcomes to agents' effort in in-group and out-group matches separately for the incentivising and for the non-incentivising principals and for the endogenously derived good and bad outcomes. The figure shows that principals playing the incentivising bonus-rewarding strategies attribute outcomes to effort in out-group than in-group matches when the observed outcome is bad, at rates of .58 vs .51, respectively, with a difference of $.08\,(-.01, .17)$ but more often to effort in in-group than out-group matches when the outcome is good, .56 vs .43, respectively, with a difference of $.13\,(.03, .24)$. Non-incentivising principals (who always or never reward) attribute outcomes to effort in both in-group and out-group matches at similar rates: .57 and .62 (the difference of $.05\,(-.10, .19)$ is not systematically different from zero). In short, while there exists a systematic attribution asymmetry between in-group and out-group matches for the incentivising principals, there is no such asymmetry for non-incentivising principals.

Figure 5: Principals' rates of attribution to effort; 95% confidence bounds based on a subject-level clustered bootstrap. Good (bad) outcomes defined as those outcomes above (below) the relevant principal's individual outcome-threshold for rewarding.



Are the attribution asymmetries for the incentivising principals evidence of group-specific biases? To fix concepts, let the *in-group bias in attribution* at O, $b(O)$ be the attribution to effort in in-group matches at O minus the attribution to effort in out-group matches at O. But how should we think about a principal's bias more generally *across outcomes*? We cannot measure it for the incentivising principals by simply comparing average attribution choices above and below the good outcome thresholds in in-group and out-group matches because the sets of good outcomes tend to be different in those matches.[26] An incentivising principal who is willing to reward in-group agents for lower outcomes than out-group agents may appear to be more likely to attribute good outcomes to effort in the in-group than in the out-group matches but may, in fact, be group-neutral in attribution at a fixed level of outcome. Avoiding the confounding effects of differences in good outcome thresholds by measuring in-group bias in attribution at the level of the individual principal would be problematic because, for a particular subject, a given good outcome in the in-group matches may not have an equivalent outcome in the out-group matches, and certainly does not have an equivalent bad outcome. We get around this problem by making inferences based on the behaviour in in-group and out-group matches of comparable principals, pooling together principals who show similar biases in reward decisions. Figure 6 gives the in-group bias in attribution for different principals: out-group biased, no bias, and in-group biased in rewards.

Figure 6: Average difference in the rates of incentivising principals' attribution to effort between in-group and out-group matches (= in-group bias in attribution) over in-group bias in rewards of incentivising principals; 95% confidence bounds based on a subject-level clustered bootstrap.



[26] In contrast, because those sets are the same for the non-incentivising principals, that comparison is the right measure of their group-specific bias.

This average difference in attribution indicates a U-shaped relationship between in-group bias in attribution and in-group bias in rewards. The observed difference in rate of attribution of good outcomes to effort in in-group and out-group matches is $.25\,(.05, .45)$ aggregated for principals who are more demanding of the in-group than of the out-group agents (in-group bias in rewards < 0) and $.17\,(-.02, .35)$ for principals who are less demanding of the in-group agents (in-group bias in rewards > 0). The in-group bias in attribution for principals who do not differentiate between in- and out-group agents in terms of demanded outcomes in their reward decisions is indistinguishable from zero: $.04\,(-.10, .19)$.[27]

To summarise:

Result 4 *(a) Incentivizing principals tend to be in-group biased in attribution of good outcomes to effort, while non-incentivising principals show no attribution bias. (b) For incentivising principals with in-group and out-group reward biases, in-group bias in attribution of good outcomes to effort is higher than for incentivising principals with no reward bias.*

5.2.2 Agents

Agents' beliefs and biases in effort Recall from our Section 5.1 that, on average, agents in the STRATEGIC treatment invest slightly more into effort in in-group matches than in out-group matches but that this average difference is relatively small in size. The subject-level analysis of agents' choices, however, reveals systematic group identity biases that respond to agents' anticipation of biases in principals' reward decisions. Our findings on agents' recorded beliefs are summarised as follows:

Result 5 *While there is 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.*

Agents' beliefs about principals' biases are asymmetric, consistent with the overall direction of bias in principals' actual reward choices. In particular, while the agents' expectations of bias in

[27] Section A.4.1 in the SI gives more information on the robustness of this results, in particular, we show in a regression framework that in-group bias in attribution is systematically different from zero for many values of in-group bias in rewards as well as that holding fixed a particular outcome value and pooling the attribution decisions of principals with a similar level of bias in rewards (on a more fine grained scale than the one with 3 categories shown above) does not change our interpretation.

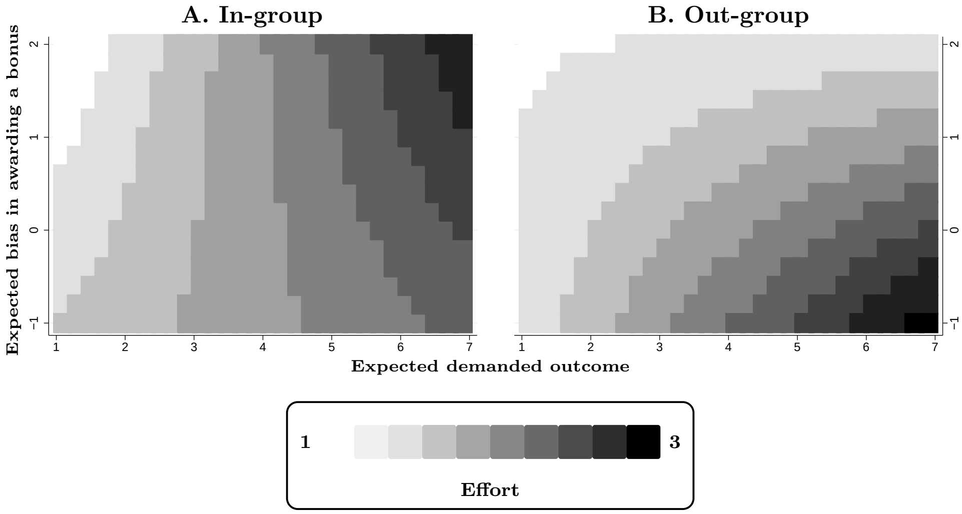
principals' reward decisions range from 1 to 7 with the mean at 3.5 $(3.21, 3.75)$, the average difference in expected demands between the in- and out-group matches is $-.12\,(-.30, .05)$ indicating that the distribution of expected in-group bias in principals' reward choices is skewed.

Turning now to the relationship between agents' recorded beliefs and effort choices, we find that in both, in-group and out-group matches, *effort is increasing with expected demands* (henceforth referred to as *the expected demand effect*). A one-unit increase in expected demands leads to an increase in effort of .20 $(.10, .29)$ in in-group matches and .19 $(.06, .31)$ in out-group matches. Further, we find that *in-group bias in agents' effort is increasing with their expectation of the principals' in-group bias in rewards* (henceforth referred to as *the expected bias effect*). When agents believe that to be rewarded, they are expected to deliver lower outcomes in in-group than in out-group matches, their effort is predicted to be .14 $(.02, .26)$ higher in in-group than in out-group matches; when they believe higher outcome is required for reward in in-group than out-group matches, the difference estimate is $-.08\,(.04, -.20)$. Differences in effort choices are smallest for agents who do not expect identity-contingent differences in principals' demands.[28] In sum, properly accounting for the level of expected demands corrects the impression of no difference between agents' behaviour in in-group and out-group matches.

Figure 7 presents predicted values of effort and shows both the expected demand effect and the expected bias effect in operation.

[28] Estimates in this paragraph are taken from regression Model 4 in Table A.10 in the SI. Note, we are excluding 4 of the 55 agents in the STRATEGIC treatment because they failed to make consistent choices in the risk elicitation task.

Figure 7: Predicted levels of effort over expected bias plotted over expected demand



The following result summarises our key conclusions:

Result 6 *Agents' choices display an expected demand effect in in-group and out-group matches as well as an expected bias effect.*

Note that the expected demand effect and the expected bias effect work, on average, in opposite directions. While according to the expected demand effect, agents' expectations of lower demands from in-group than out-group principals should induce higher effort in out-group than in in-group matches, such expectations lead, according to the expected bias effect, to higher effort in in-group than in out-group matches.

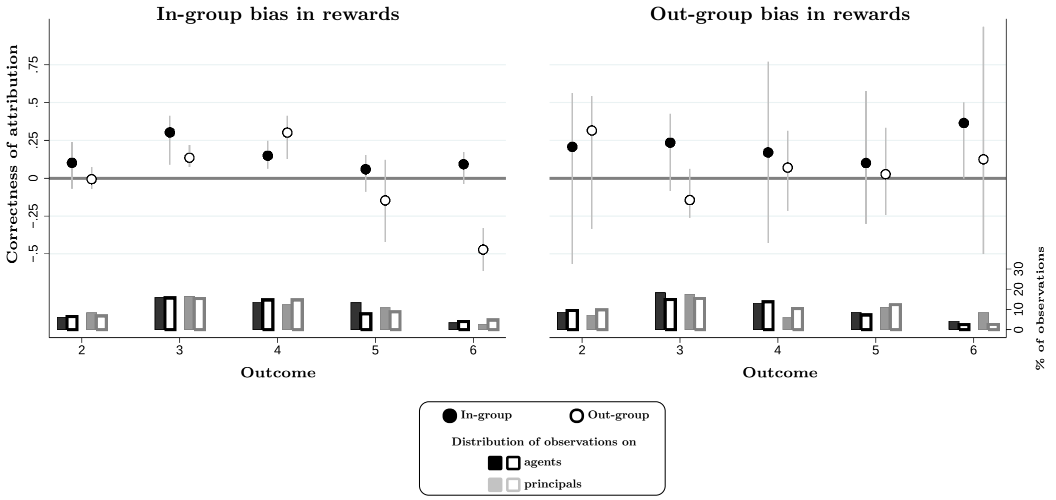
5.2.3 Are the attributions correct?

Our key results show the existence of systematically asymmetric group-specific choices and judgments by the principals and agents' responsiveness to the expectation of such asymmetries. But are principals' attributions ultimately correct in their assessments of the agents' decisions?

Principals' attribution choices could entail two distinct kinds of mistakes: (1) holding fixed a given kind of identity match, principals may incorrectly attribute outcomes to higher effort vs. higher type; or, (2) they may erroneously think that agents who share an identity with them tend to choose higher effort than the agents who do not. Our design directly incentivises the correctness of beliefs to avoid (1) and indirectly, with the same elicitation mechanism, to avoid (2) as well. Because, for reasons indicated above, the aggregate-level assessments of these mistakes are problematic, in what follows we focus on two conditional assessments, each providing a natural way of proceeding in relation to a given type of attribution mistake. In both exercises, we ask: are principals' attributions correct when agents' expectations are correct?

Are principals right *within* identity matches? We consider behaviour within counterfactual principal-agent pairs that match on the actual (for the principals) and the expected (by the agents) reward threshold outcomes. At each level of outcome in a distinct identity match condition, we record the principal's *correctness in attribution within identity match.* Holding fixed the outcome, this quantity measures the difference in the proportion of observations for which the agents' effort levels were larger than their type and the proportion of observations where principals' correctly attribute those outcomes to effort. Figure 8 plots the correctness measure where the value of 0 on the y-axes corresponds to the principals' always correctly guessing the ordering of type and effort for the given outcome levels. We show negative deviations (underestimation of agents' effort relative to type) and positive deviations (overestimation) from a correct guess.

Figure 8: *Correctness in attribution within identity match* at each level of outcome for the counterfactually matched incentivising principals and agents with the corresponding expectations about in-group bias in rewards. Results are shown for two groups of pairs distinguished by the sign of expected/actual in-group bias in rewards; the unit of this analysis is groups of agents/principals who show similar levels of (expected) in-group bias (values within the interval of 1 on the (expected) in-group bias in rewards-scale); 95% confidence bounds based on a subject-level clustered bootstrap.



Note, first, that the most on-the-mark attributions are by the in-group biased principals in in-group matches and by the out-group biased principals in the out-group matches – for higher than average outcomes (4 and above). The attributions are not systematically different from what would be a perfect guess, but tend to be too high, especially in in-group matches, for lower than average outcomes.

Second, for pairs with (expected) in-group bias > 0, principal's attribution choice is closer to correct in in-group rather than out-group matches; the most systematic attribution mistakes are due to the under-attribution to effort in the higher than average range. Relating this back to our motivating example, Bob's interpretation of Alice's performance tends to under-appreciate Alice's effort. Relating to the discussion of the two effects on agents' choices we saw in the previous section, we may say that principals focus on the expected bias effect, and under-appreciate the implications of expected in-group bias in rewards on the manifestation of the expected demand effect.

Are principals right *across* identity matches? In Section 5.2.1, we provided evidence that principals who are in-group biased in rewards are also in-group biased in attribution. However, as the evidence of a robust expected demand effect in agents' choices suggests, agents respond to higher expectation (in this case, in out-group matches) by increasing their effort to meet the demand. Even if agents' choices are subject to the expected bias effect, if they expect the demands in the out-group matches to be sufficiently high relative to the in-group demands, the expected demand effect may override the expected bias effect, producing a *higher*, not lower, effort in the out-group matches. Our regression-based estimate of agents' effort reinforces this conclusion. When the agents expect to be facing symmetric demands from in-group and out-group principals, they choose higher effort in the former (the difference is $.14\,(-.09, .37)$ in favour of the in-group). But the sign of the difference flips if the agents now expect to meet higher demands in the out-group match: expecting that the principal's demand is two outcome points higher in out-group than in in-group matches (i.e., expecting a strong in-group bias), lifts the average effort in out-group matches to $.27\,(-.70, .17)$ above the effort in in-group matches.[29]

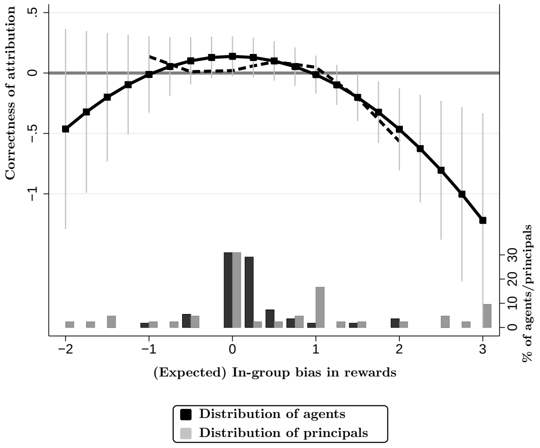
The strategic consistency of the in-group bias in rewards and the in-group bias in attribution is, thus, a function of the size of the reward bias. When the latter is relatively small, the two biases

[29] Estimates are based on Model 4 in Table A.10 in the SI.

are mutually consistent. When the in-group bias in rewards is large enough, the principals should be expecting the size of the expected demand effect to counter the size of the expected bias effect; in this case, the persistent in-group bias in attribution is evidence of the principals' under-appreciating the symmetric force of the expected demand effect and over-weighting the expected bias effect.

Figure 9 may be thought of as a visualisation of the assessment of whether the principals get the net balance of these effects right. The values in the figure correspond to pairs of incentivising principals and agents, matching principals' in-group bias in rewards and agents' expectations of in-group bias. The distance from zero on the vertical axis gives a measure of *correctness of attribution across identity matches.* It is computed as the difference between (1) the average difference between attribution to effort in in-group and out-group matches at a given outcome and (2) the difference between the proportions of observations with effort greater than type in in-group and out-group matches.[30]

Figure 9: *Correctness of attribution across identity matches* over in-group bias in rewards. Markers give the predicted correctness of attribution estimated from a regression of correctness of attribution on (expected) in-group bias in rewards and its squared value. The dashed line gives the lowess estimate from the raw value of correctness of attribution. The unit of analysis pairs matched groups of agents/principals who show similar levels of (expected) in-group bias (values within the interval of 1 on the (expected) in-group bias in reward scale); 95% confidence bounds based on a subject-level clustered bootstrap.



The evidence in the figure reinforces our interpretation. It is the principals who are in-group biased in rewards who display the largest deviation from a correct guess about agents' in-group bias

[30] This difference is a perfect proxy for the difference between average effort in in- and out-group matches because type values are realised from a uniform distribution. We use the difference in proportions in computing correctness of attribution across identity matches because it is a direct measure of the attribution decisions principals have to make (answering the question "what do you think is higher, effort or type?")

in effort, consistent with our conjecture of their failing to correctly anticipate the strength of the expected demand effect in out-group agents.

5.2.4 Agents' effort and risk attitude

Agent's choice are a comparison of a sure value (cost of investment) to an expected value of a lottery (outcome and reward, contingent on realizations of random variables). It would be reasonable to suppose that in making this choice, subjects will respond to the explicitly given payoffs in ways that track their personal unmodeled risk preferences. This would induce a variation in effort choice where our prediction of agent choices for a given type t — in particular, for the outcome-contingent-play (OCP) equilibrium — expects no variation relative to the differences in the expected retention threshold $z \in \{3, 4, 5\}$.

Indeed, the expected demand effect and the expected bias effect described above are related to the agents' risk preferences. Our first conclusion regarding the consequences of subjects' risk preferences is one of no effect. Using the measure of risk-attitude constructed from the elicited attitudes, we find no significant relationship between the agents' risk preference and their expectation of demand from the principals in either in-group or out-group matches. The marginal effect of risk aversion on expected demand is not systematically different from zero in either in-group or out-group matches ($-.01\,(-.13, .11)$ and $.07\,(-.06, 20)$, respectively).[31] This suggests that if risk preferences have an effect, it is on agents' effort choices, not their beliefs.

The second conclusion regarding risk preferences reveals an effect. We find that risk preferences are key to the relationship between expected demands and effort and to the relationship between expected in-group bias and effort. Figure 10.A shows the marginal effect of expected demands on effort plotted over risk-aversion – the number of safe choices as elicited by the Holt and Laury (2002)-list at the beginning of each session. As this plot makes clear, the expected demand effect systematically increases with agents' risk-aversion.

[31] Marginal effects are estimated from a regression of expected demand on risk-aversion, in-group status, their interaction, and round of play; errors are clustered by subject.

Figure 10: Marginal effect of expected demands on effort (= expected demand effect) and difference in marginal effect of expected in-group bias on effort (= difference in expected bias effect) over risk-aversion (number of safe choices in the Holt and Laury (2002)-list). Effect is estimated from Model 4 in Table A.10; 95% confidence bounds based on a subject-level clustered bootstrap.

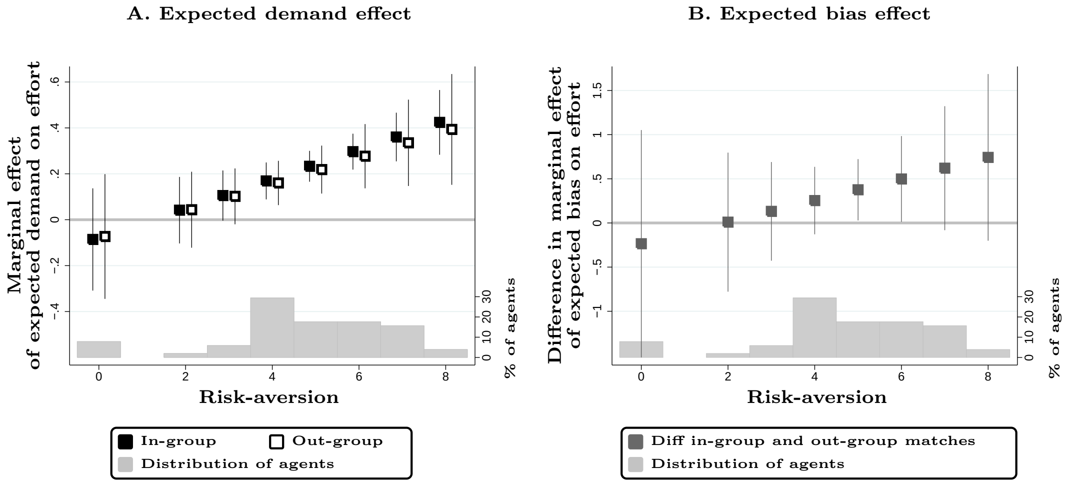


Figure 10.B shows that the expected bias effect also grows stronger with risk-aversion and tilts the difference in effort in in-group and out-group matches towards higher investment into effort when an agent shares an identity with the matched principal. The difference in marginal effect of expected bias on effort is systematically larger in in-group than in out-group matches and increasingly so for more risk-averse agents, while there is no such difference for more risk-acceptant agents.

Result 7 *Both the outcome demand effect and the expected bias effect increase with agents' risk-aversion.*

A plausible way of understanding the behavioural motivations behind the patterns in Figure 10 is by conceiving of the agents as viewing the bonus as a reference payoff and seeking to insure themselves against losing it with investment into effort. Consistent with this interpretation, when the agents anticipate higher outcome demands, the more risk-averse among them react more strongly by investing more on the margin to meet those demands. In this way, risk-aversion drives the demand effect as a behavioural instantiation of purchasing insurance against losing a reference payoff. However, if that payoff is too distant – too risky – the insurance premium may become too expensive to be worth purchasing, and so we should see the more risk-averse agents losing interest in it faster. Perhaps, the status of the bonus as a reference payoff itself becomes for the risk-averse agents less plausible when the risks associated with it become too great. Put somewhat differently, those agents for whom the gap between in-group and out-group expectations is high tend to regard the bonus in the out-group matches as a particularly distant prospect. That would account for the relationship we see in Figure 10.B.[32] If this interpretation is right, the patterns depicted in the two panels of Figure 10 should be most pronounced when agents have lower types. Indeed, that is the case: the effect of risk preference on the expected demand effect and on the expected bias effect is highest for type 1 agents (See Figure A.3 in the SI).

5.3 Average treatment effects, take 2

In this section, we re-visit the analysis of treatment effects, but now consider them with subject-level filters that address the ecological inference concerns we raised for interpreting aggregate-level data.

5.3.1 NON-IDENTITY treatment

We begin by re-considering how subject behaviour in the STRATEGIC treatment compares to behaviour in the absence of activated social identities (the NON-IDENTITY treatment) – a comparison that will now help pinpoint the aspects of the identity relationship responsible for the reported results.

First, we find that principals' expectations in the NON-IDENTITY treatment lie between those of principals in in-group and out-group matches in the STRATEGIC treatment, but closer to those in the out-group. The average principal-specific outcome threshold in the NON-IDENTITY treatment is 4.45, in contrast to a higher average threshold of 4.53 in out-group matches and a lower average threshold of 3.96 in in-group matches in the STRATEGIC treatment, with the difference in thresholds in in-group and out-group matches systematically different from zero. This result suggests that the effect of inducing identities on principals' reward biases manifests primarily in a change in the relative status of the in-group agents; principals tend to treat out-group agents essentially the same way they treat agents in the identity-free environment, but treat the in-group

[32] According to classic prospect theory (Kahneman and Tversky, 1979), individuals' decisions under uncertainty – where risk attitudes matter – vary with the reference point to which potential outcomes are compared. In particular, when expecting a loss, individuals show a distaste for insuring loss. Kőszegi and Rabin (2007), however, argue that when risk and the possibility of insuring it are anticipated – rather than coming as a surprise – behaviour is driven by risk-aversion. Our interpretation is in line with this argument. Risk-averse agents are insuring themselves against the expected risk of not receiving the bonus. If the bonus is believed unobtainable, the (risky) outcome of receiving it ceases to be the reference point and agents drop their investment into effort.

agents significantly more favourably.

Second, we find that in contrast to the STRATEGIC treatment, in which the attributions by incentivising principals above the good outcome threshold are significantly different between in-group and out-group matches (higher in the former than in the latter), the incentivising principals' attributions to effort in the NON-IDENTITY treatment are not distinguishable from random for either good or bad outcomes. This contrast suggests, consistent with the idea of strategic discrimination, that the effect of the identity treatment is to create asymmetric behavioural expectations associated with the additional information entailed in the identity markers – an effect that is smoothed out in the less informative identity-free environment. Of course, as the counterfactual matching analysis we saw above suggests, principals are not using this additional information equally well in the two identity group match conditions: on average, they anticipate the agents' choices in the in-group matches better than in the out-group matches.

5.3.2 NON-STRATEGIC treatment

As we emphasized above, the only substantial difference between the STRATEGIC and the NON-STRATEGIC treatments is that in the NON-STRATEGIC treatment, principals cannot choose whether to give a bonus. In the NON-STRATEGIC environment, whatever asymmetry in beliefs is observed must be due to psychological, taste-for-discrimination factors like the ultimate attribution error. Because of the nature of the NON-STRATEGIC treatment, and in order to avoid contaminating the results with pre-treatment effects, we do not incorporate into this treatment an element that would identify subjects who would, if this were a STRATEGIC treatment, behave as incentivising principals. Similarly, while the notion of a good vs. bad outcome is well-defined in the STRATEGIC treatments because it is identified endogenously relative to the reward decision, no such natural endogenous identifier exists in the NON-STRATEGIC treatment.

Without identifying incentivising type in the set of principals and the subject-specific values of good vs. bad outcomes in the NON-STRATEGIC treatment, we cannot make what would be the perfect comparison, at the subject level, to the STRATEGIC treatment. What we offer is two imperfect comparisons that, nonetheless, go some distance toward clarifying the behavioral comparison of the effects of these treatment.

The first comparison is at the endogenously specified subject-specific good vs. bad outcome thresholds in the STRATEGIC treatment and the exogenously specified common thresholds in the NON-STRATEGIC treatment. While the latter thresholds may be smoothing the variations across principals, it is important to see that a key reason for evaluating treatment effects in the STRATEGIC treatment at the subject-specific thresholds – different reward strategies – is moot in the NON-STRATEGIC treatment, suggesting that the reasons for variation there are attenuated.

Recall that incentivising principals in the STRATEGIC treatment display a substantial in-group bias in their attribution of good outcomes to effort – $.13\,(-.03, .24)$ – but no bias for bad outcomes ($-.08\,(-.17, .01)$). We estimate attribution bias for the set of all principals in the NON-STRATEGIC treatment, drawing the line of "good" outcomes with respect to the NON-STRATEGIC treatment at 5 or above, and "bad outcomes" at 3 or below. In contrast to the STRATEGIC treatment, we observe in this treatment no significant bias in attribution choices between in- and out-group matches, neither when the principals observed bad outcomes, $.01\,(-.16, .18)$, nor when they observed good outcomes $-.03\,(-.22, .16)$.[33]

Given that we cannot separate incentivising from non-incentivising principals in the NON-STRATEGIC treatment, the estimate of in-group bias in attribution choices in the latter will be averaging across those two types of principals and, intuitively, will be lower than the bias observed among incentivising principals in the STRATEGIC treatment. Our second comparison addresses this concern. As a more (perhaps, most) conservative estimate of the average treatment effect, we compare attribution decisions of the 75% most in-group biased principals in the NON-STRATEGIC treatment – the share of incentivising principals among all principals in the STRATEGIC treatment (principals whose choices, as a group, unlike those of always/never rewarding principals, reveal identity preference biases we describe above). Strikingly, we find that the attribution bias among these most biased principals in the NON-STRATEGIC treatments at good outcomes is still smaller than that of the incentivising principals in the STRATEGIC treatment, .09, $(-.13, .30)$.[34] Both of these comparisons point in the same direction suggesting that the strategic nature of the principal-agent

[33] Since at median outcomes of 4 we do not have the theoretically or empirically grounded predictions for how ultimate attribution error manifests itself that we have at low or high outcomes and in the STRATEGIC treatment, the good/bad outcome thresholds for the incentivising principals are mainly distributed between 3 and 5, we restrict our attention in the comparison of behaviour in strategic and non-strategic environments to what would be (relatively) clear cases of low and high outcomes. Applying a threshold separating good and bad outcomes in the NON-STRATEGIC treatment at 3 or 5 does not change the interpretation of the results.

[34] Note that this comparison required reducing our sample of principals in the NON-STRATEGIC treatment, driving down statistical power.

relationship in the STRATEGIC treatment is responsible for the bias in attribution we report.[35]

6 Interpreting the evidence

We argue that the totality of evidence provided is best understood as a manifestation of strategic discrimination, whereby the principals' attribution choices are identity-contingent because they reflect their beliefs about agents' responses to principals' reward choices and the agents' effort choices are investment decisions responding to those expectations. In this account, incentivising principals in the STRATEGIC treatment expect that good outcomes are more likely to be a product of effort in in-group than in out-group matches, and that in out-group matches they are, on the margin, more due to lucky draws of noise or type. Because principals are less likely to reward out-group than in-group agents at high outcomes, out-group agents, and especially the more risk-averse among those, will choose marginally lower effort than they would as in-group agents, which reinforces the principals' bias.

There is another prominent possibility. The conjunction of the principals' reward bias in favour of the in-group agents and the agents' effort increasing in the expectation of that bias may implicate a norm of mutual reciprocity rather than a strategic discrimination. Such a norm may correspond to an equilibrium of a different game – played outside the lab – in which identity-indexed interactions are repeated and the mutual in-group favouritism (reciprocity) is the focal equilibrium. Such an equilibrium may motivate subjects' interpretations of the proper behaviour in social identity contexts, including the one implemented in our experiment. The principals' attribution choices, then, would be understood as simply describing the expectation that comes with that norm.

We cannot, of course, rule out the possibility that the reciprocity interpretation is correct for at least some subset of subjects in the sample, but the totality of evidence we present above casts doubt on the power of the reciprocity account as a general explanation.

First, our results on the effect of risk attitude suggest both that the investment interpretation has independent support and that the power of reciprocal favouritism, if it is the interpretation of

[35] As an aside, recall that in our report on average treatment effects with respect to agents' behavior, we noticed lower levels of effort exerted by low-type agents in the STRATEGIC treatment when compared to the NON-STRATEGIC treatment. We also showed that risk preferences systematically influence how agents' expectations inform their effort choices. Given that a strategic environment is characterized by higher levels of uncertainty than a non-strategic one, the existence of more risk-averse than risk-seeking subjects depresses effort for low types.

the expected bias effect, is highly contingent on agents' underlying risk attitude.

Second, the reciprocity interpretation does not sit well with what we saw from the most natural subset of principals who may be expected to be anticipating reciprocal favouritism from the agents: the principals who always reward agents. For those principals, operating with a norm of reciprocity should entail the expectation of higher effort from all their matched partners; however, we do not find that those principals are more likely to attribute good outcomes to effort (difference $= .12\,(-.14, .40)$).

Third, evidence from the NON-STRATEGIC treatment shows that the attribution asymmetries that we saw in the STRATEGIC treatment are a function of the strategic environment. Even if the reward choices were somehow based on expectations of reciprocity, it is clear that attribution choices are responding to features of the environment created in the lab rather than induced by considerations from outside the lab. The NON-STRATEGIC treatment also provides evidence against another interpretation of principals' behaviour: that the principals who are discriminating in beliefs are just the kind of people who are particularly more likely to discriminate in rewards – that is, they may just be the people who are simply more prone to discriminate. If that were the case, then 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 possibility does not 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.

Fourth, while the strategic discrimination account requires self-awareness, the mutual reciprocity account less so, and arguably the strongest version of the reciprocity-driven identity effect may be instinctive rather than intentional. In the exit survey following our STRATEGIC treatment, we asked questions that allow us to evaluate the relationship between subjects' self-awareness and their choices in the experiment. We find that a large proportion of subjects discriminated intentionally – that is, with self-awareness – responding to the expectations of each other's play. In the survey, 44% of incentivising principals indicate that they were influenced in their reward decision by the group membership of their matched agent, in contrast to no principals who always or never awarded a bonus to agents saying that group identities mattered. The contrast with respect to the attribution decision is less stark but still significant: 35% of incentivising principals claimed to be influenced by group membership in their attribution choices but only 23% of principals who always or never rewarded. Further, within the set of incentivising principals, awareness of one's own bias in reward decisions increases attribution of good outcomes to effort in in-group matches in contrast to out-group matches. For those who are aware of their reward biases, the in-group bias in the attribution of good outcomes to effort is $.27\,(.06, .49)$ in contrast to $.14\,(-.03, .30)$ for those who admit no such self-awareness. In sum, principals whose reward and attribution choices are asymmetric tend to be aware of it, and, further, principals who are more likely to attribute good outcomes to effort in the in-group than in the out-group matches tend to be more aware of their asymmetric treatment of agents in their bonus award decisions.

As a final thought about the robustness of our interpretation of findings, our research design seeks to separate non-strategic statistical responses from strategically driven ones, so it is important to rule out the possibility that the behaviour 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 in a given round. However, it may be possible, in principle, that subject behaviour tracks different individual experiences in in-group vs. out-group matches. If that is the case, our background argument that the artificial identity does away with the possibility of a non-strategic statistical discrimination would be weakened. If however, principals' attributions are not related to their experiences in previous rounds, we can dismiss this concern. In the STRATEGIC treatment (and in the relevant comparisons to the NON-STRATEGIC treatment), we find no relationship between the difference in level of outcomes observed in previous rounds between in-group and out-group matches and principals' reward and attribution decisions in the current round. Also, agents' beliefs are not affected by their experiences of reward decision of the matched principal while agents' effort choices are positively related to such an experience in the STRATEGIC treatment (see Section A.6 in the SI).

We still argue that the absence of history-of-play effects on principals' beliefs, their reward decisions, and agents' beliefs means that the asymmetric responses to identity matches that we observe cannot be based on experienced statistical differences between groups. The asymmetric treatment of groups must arise from un-reinforced asymmetric beliefs about what to expect from the behaviour of the members of the two groups.

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 analyses those expectations in a strategic principal-agent setting that models relationships between employers and employees. Our goal is to provide a behavioural evaluation of the correlates of a discrimination that is fundamentally strategic, induced 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 generalisations, whether rationally or psychologically sustained.

One of our key findings is that principals' and agents' choices may, through potentially correct mutual expectations, sustain a pattern of beliefs that is observationally equivalent to those conforming to the ultimate attribution error 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' biased award choices.

In strategic environments discriminating behaviour does not necessarily go together with prejudicial stereotyping. Biases in principals' reward choices correlate with principals' beliefs about agents' effort choices, and agents' effort choices are responding to agents' identity-contingent expectations. In such settings, asymmetric identity-contingent interpretations of agent performance that are observationally equivalent to prejudice may not be based on incorrect beliefs about differences in agents' performance but on an anticipation – a plausibly correct one – of the greater ability to incentivise those agents with whom principals share a group identity. Out-group agents, like Alice from our motivating example, may be justified in expecting that they are treated more harshly by their principals and in reducing effort in (correct) anticipation of lower likelihood of receiving a recognition for their effort. While their rational response may provide a rationale for Bobs' discriminatory actions, we show that the strategically acting principals tend to under-appreciate the effort from the out-group agents or, equivalently, out-group agents may be justified in reducing their effort still farther than they in fact do.

As a positive matter, the evidence of strategic discrimination suggests that the existing measures of prejudice in the observational studies may be partial-equilibrium: they may be identifying a joint measure of prejudice and rational expectations associated with an equilibrium performance rather than prejudice alone. More broadly, the analysis of subject behaviour we present suggests that the strategically induced attribution asymmetries may be *a*, if not *the*, first-order phenomenon when it comes to accounting for discriminatory choices by principals. As a matter of normative policy design, the analysis suggests that discrimination may not be easily combatted by providing stakeholders with information that seeks to improve mutual empathy. If discrimination arises endogenously and does not require exogenously set beliefs about group differences, general information about the out-group partners may not alter the endogenously determined asymmetries. A more promising approach to alleviating strategic discrimination may be reward schemes that cannot differentiate between in- and out-group agents, and reward agents on observable measures of performance without conditioning on principals' second-guessing of their causes.

References

Akerlof, George A. 1982. "Labor contracts as partial gift exchange." *The Quarterly Journal of Economics* pp. 543–569.

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.

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

Becker, Gary S. 1971. *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.

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

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

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

Charness, Gary. 2004. "Attribution and reciprocity in an experimental labor market." *Journal of Labor Economics* 22(3):665–688.

Charness, Gary and Ernan Haruvy. 2002. "Altruism, equity, and reciprocity in a gift-exchange experiment: an encompassing approach." *Games and Economic Behavior* 40(2):203–231.

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.

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.

Fehr, Ernst, Erich Kirchler, Andreas Weichbold and Simon Gächter. 1998. "When social norms overpower competition: Gift exchange in experimental labor markets." *Journal of Labor economics* 16(2):324–351.

Fehr, Ernst, Georg Kirchsteiger and Arno Riedl. 1998. "Gift exchange and reciprocity in competitive experimental markets." *European Economic Review* 42(1):1–34.

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 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.

Haan, Thomas, Theo Offerman and Randolph Sloof. 2015. "Discrimination in the Labour Market: The Curse of Competition between Workers." *The Economic Journal* .

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.

Jin, Nobuhito, Toshio Yamagishi and Tohko Kiyonari. 1996. "Bilateral dependency and the minimal group paradigm." *Japanese Journal of Psychology* .

Kahneman, Daniel and Amos Tversky. 1979. "Prospect Theory: An Analysis of Choices under Risk." *Econometrica* 47(2):263–91.

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).

Kőszegi, Botond and Matthew Rabin. 2007. "Reference-dependent risk attitudes." *The American Economic Review* pp. 1047–1073.

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. 2015. "Social Identity and Electoral Accountability." *American Journal of Political Science* 59(3):671–89.

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.

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.

Yamagishi, Toshio and Toko Kiyonari. 2000. “The group as the container of generalized reciprocity.” *Social Psychology Quarterly* pp. 116–132.

Supporting information

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	55	1100
STRATEGIC	Kandinskys	55	1100
	Total	110	2200
NON-IDENTITY	Total	38	760
	Klees	21	420
NON-STRATEGIC	Kandinskys	19	380
	Total	40	800
		188	3760

The STRATEGIC treatment condition generated 55 Klees, subjects who preferred paintings by Paul Klee most of the time, and 55 Kandinskys, subjects who preferred those by Vassily Kandinsky most of the time. In the NON-STRATEGIC 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

Table A.2: Means (standard deviation), minimum, and maximum values of type, effort, outcome, attribution decision (0 = attributed to type, 1 = attributed to effort), and reward decision (0 = not rewarded, 1 = rewarded) by treatment.

	STRATEGIC		**NON-IDENTITY**	**NON-STRATEGIC**			
Variable	In-group	Out-group		In-group	Out-group	Min	Max
Type	1.97 (.82)	2.01 (.80)	2.01 (.81)	2.00 (.79)	2.05 (.79)	1	3
Effort	1.79 (.78)	1.73 (.79)	1.76 (.84)	2.11 (.77)	2.17 (.76)	1	3
Outcome	3.70 (1.3)	3.68 (1.3)	3.81 (1.3)	4.03 (1.1)	4.14 (1.2)	1	7
Attribution	.555 (.50)	.534 (.50)	.455 (.50)	.66 (.48)	.57 (.50)	0	1
Reward	.594 (.49)	.483 (.50)	.605 (.49)	-	-	0	1

A.3 Aggregate behaviour

Table A.3: Agents' effort regressed on a treatment-dummy (STRATEGIC serves as base category), agent's type, in-group status of the matched principal (STRATEGIC vs NON-STRATEGIC, the interactions of those variables, and round of play.

VARIABLES	**STRATEGIC** and **NON-STRATEGIC**	**STRATEGIC** and **NON-IDENTITY**
treatment	1.15***	0.27
	(0.198)	(0.240)
type	-0.16***	-0.17***
	(0.047)	(0.037)
treatment × *type*	-0.33***	-0.11
	(0.098)	(0.088)
in-group	0.05	
	(0.133)	
treatment × *in-group*	-0.04	
	(0.189)	
in-group × *type*	-0.01	
	(0.056)	
treatment × *in-group* × *type*	-0.03	
	(0.086)	
round	-0.01**	-0.01**
	(0.004)	(0.004)
Constant	2.12***	2.16***
	(0.129)	(0.112)
Observations	1,720	1,700
R-squared	0.141	0.049

Standard errors clustered by subject in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table A.4: Agents' expected demand regressed on a treatment-dummy (STRATEGIC serves as base category), agent's type, in-group status of the matched principal (STRATEGIC vs NON-STRATEGIC, the interactions of those variables, and round of play.

VARIABLES	**STRATEGIC**	**STRATEGIC** and **NON-IDENTITY**
NON-IDENTITY		0.41
		(0.356)
type	0.21***	0.27***
	(0.073)	(0.071)
NON-IDENTITY × *type*		0.01
		(0.120)
in-group	-0.34	
	(0.249)	
in-group × *type*	0.12	
	(0.111)	
round	-0.01	-0.01*
	(0.009)	(0.008)
Constant	3.09***	2.94***
	(0.220)	(0.215)
Observations	1,230	1,606
R-squared	0.031	0.049

Standard errors clustered by subject in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table A.5: Principals' reward decisions regressed on outcome, in-group status of the matched agent, the interactions of those variables, and round of play in the STRATEGIC treatment and reward decision regressed on outcome, a treatment-dummy (STRATEGIC serves as base category), the interaction of those variables, and round of play in STRATEGIC and NON-IDENTITY treatment.

VARIABLES	**STRATEGIC**	**STRATEGIC** and **NON-IDENTITY**
NON-IDENTITY		0.59
		(0.654)
outcome	0.27***	0.33***
	(0.081)	(0.072)
NON-IDENTITY × *outcome*		-0.06
		(0.137)
in-group	-0.02	
	(0.441)	
in-group × *outcome*	0.12	
	(0.097)	
round	-0.05***	-0.04***
	(0.013)	(0.011)
Constant	-0.69*	-0.78**
	(0.378)	(0.336)
Observations	1,320	1,700

Standard errors clustered by subject in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table A.6: Principals' attribution decision regressed on a treatment-dummy (STRATEGIC serves as base category), outcome, in-group status of the matched agent (STRATEGIC vs NON-STRATEGIC treatment), the interactions of those variables, and round of play.

VARIABLES	**STRATEGIC** and **NON-STRATEGIC**	**STRATEGIC** and **NON-IDENTITY**	**NON-STRATEGIC**
treatment	2.01**	0.67	
	(0.838)	(0.563)	
outcome	-0.04	-0.02	-0.49***
	(0.085)	(0.072)	(0.164)
treatment × *outcome*	-0.45**	-0.27*	
	(0.188)	(0.143)	
in-group	-0.08		0.37
	(0.447)		(1.096)
treatment × *in-group*	0.53		
	(1.153)		
in-group × *outcome*	0.03		-0.01
	(0.109)		(0.216)
treatment × *in-group* × *outcome*	-0.06		
	(0.236)		
round	-0.02*	-0.03**	0.02
	(0.011)	(0.011)	(0.015)
Constant	0.51	0.56*	2.11***
	(0.340)	(0.317)	(0.756)
Observations	1,720	1,700	400

Standard errors clustered by subject in parentheses
*** p<0.01, ** p<0.05, * p<0.1

A.4 Subject-level analysis

A.4.1 Principal's choices

Table A.7: Logistic regression of principals' reward decision on covariates in NON-IDENTITY and STRATEGIC treatment. Model (4)-(6) are run on incentivising principals only.

		STRATEGIC					
VARIABLES	NON-IDENTITY	(1)	(2)	(3)	(4)	(5)	(6)
outcome	0.27**	0.33***	0.33***	0.27***			
	(0.115)	(0.073)	(0.073)	(0.081)			
good outcome					2.46***	2.42***	2.37***
					(0.198)	(0.197)	(0.301)
in-group			0.42*	-0.02		0.41**	0.37
			(0.227)	(0.441)		(0.179)	(0.225)
in-group×outcome				0.12			
				(0.097)			
in-group×good outcome							0.09
							(0.325)
round	-0.01	-0.05***	-0.05***	-0.05***	-0.06***	-0.07***	-0.07***
	(0.021)	(0.013)	(0.013)	(0.013)	(0.023)	(0.024)	(0.024)
Constant	-0.51	-0.69**	-0.90**	-0.69*	-0.80***	-0.99***	-0.97***
	(0.615)	(0.341)	(0.354)	(0.378)	(0.264)	(0.263)	(0.258)
Observations	380	1,320	1,320	1,320	1,000	1,000	1,000

Standard errors clustered by subject in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table A.8: Logistic regression of principals' attribution of effort decision on covariates in NON-IDENTITY and NON-STRATEGIC treatment.

	NON-IDENTITY				NON-STRATEGIC		
VARIABLES	(1)	(2)	(3)	(4)	(1)	(2)	(3)
outcome	-0.29**	-0.33**	-0.48***		-0.50***	-0.49***	-0.49***
	(0.124)	(0.130)	(0.171)		(0.176)	(0.177)	(0.165)
rewarded		0.53	-0.34				
		(0.479)	(0.919)				
rewarded × outcome			0.24				
			(0.220)				
good outcome				0.16			
				(0.372)			
in-group						0.34	0.37
						(0.364)	(1.096)
in-group × outcome							-0.01
							(0.216)
round	-0.02	-0.01	-0.02	-0.01	0.02	0.02	0.02
	(0.023)	(0.023)	(0.024)	(0.023)	(0.015)	(0.015)	(0.015)
Constant	1.08**	0.89	1.42**	-0.07	2.31***	2.12***	2.11***
	(0.525)	(0.612)	(0.690)	(0.346)	(0.759)	(0.775)	(0.756)
Observations	380	380	380	380	400	400	400

Standard errors clustered by subject in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table A.9: Logistic regression of principals' attribution of effort decision on covariates in the STRATEGIC treatment.

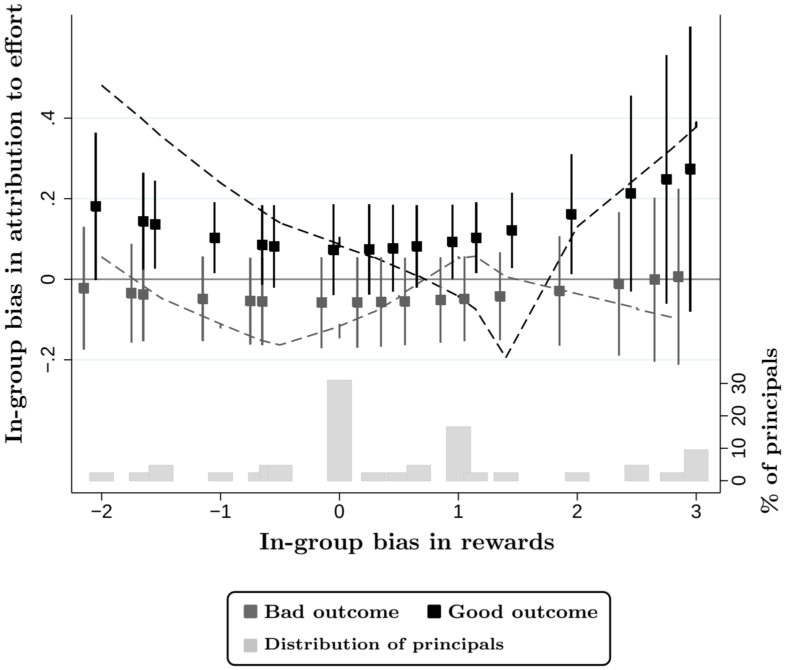
	Incentivizing principals						
VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)
in-group		0.08	-0.31	-0.25	-0.23	-0.49**	-0.50**
		(0.154)	(0.197)	(0.194)	(0.230)	(0.217)	(0.216)
good outcome	-0.21	-0.22	-0.67**	-0.59**	-0.58*	-0.67**	-0.62*
	(0.236)	(0.235)	(0.300)	(0.295)	(0.319)	(0.300)	(0.356)
in-group bias in rewards				0.09	0.10		
				(0.078)	(0.096)		
good outcome×*in-group*			0.85***	0.71**	0.68**	0.82***	0.82***
			(0.297)	(0.290)	(0.305)	(0.296)	(0.298)
in-group×*in-group bias in rewards*					-0.06		
					(0.160)		
good outcome×*in-group*×*in-group bias in rewards*					0.06		
					(0.199)		
round	-0.04**	-0.04**	-0.03**	-0.03**	-0.03**	-0.03**	-0.03**
	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)
in-group bias in rewards^2						-0.11	-0.12
						(0.115)	(0.117)
in-group×*in-group bias in rewards*^{2}						0.15***	0.15***
						(0.056)	(0.055)
good outcome×*in-group*×*in-group bias in rewards*^{2}						-0.03	-0.03
						(0.072)	(0.072)
outcome							-0.03
							(0.100)
Constant	0.59***	0.55**	0.72***	0.64**	0.63**	0.72***	0.81**
	(0.224)	(0.239)	(0.259)	(0.267)	(0.285)	(0.259)	(0.372)
Observations	840	840	840	840	840	840	840
Log-likelihood	-575.8	-575.6	-571.1	-569.8	-569.7	-566.8	-566.7

Standard errors clustered by subject in parentheses

*** p<0.01, ** p<0.05, * p<0.1

We also look at principals' attribution choices in a regression framework to assess the robustness of results on Principals' in-group bias in attribution in relation to their in-group bias in rewards as shown in Figure 6 where we pooled observations into three categories: in-group bias in rewards, no bias in rewards, and out-group bias in rewards; this is necessary, in particular, to explore our results across the full domain of values of in-group bias in rewards. To this end, Figure A.1 is based on the results from a regression of attribution to effort on in-group status, whether an outcome is below or above the threshold (whether an outcome is a bad or good outcome) for each level of in-group bias in rewards, the particular outcome observed, as well as covariates. Based on the regression estimates we generate the marginal effect of in-group vs. out-group status of the agent on attribution (= in-group bias in attribution) of good and bad outcomes over principals in-group bias in rewards (markers). We also, again, superimpose a curve of lowess estimates of the directly observed average of in-group bias in attribution for each level of principals' in-group bias in rewards for good and bad outcomes (dashed lines). Estimates are taken from Model 7 for incentivising principals in Table A.9. Informed by U-shaped curve drawn by the lowess estimator of average in-group bias in attribution, we fit a model that includes the square of in-group bias in rewards.

Figure A.1: Average difference in the rates of attribution to effort between in-group and out-group matches (= in-group bias in attribution) over in-group bias in rewards of incentivising principals in the STRATEGIC treatment. 95% confidence bounds are shown based on a principal-level clustered bootstrap and a curve fitted by the lowess estimator of in-group bias in attribution at a given pair of above/below the threshold and in-group bias in rewards.



Lowess estimates and regression estimates, similar to the average differences in attribution between in- and out-group shown in Figure 6, indicate a U-shaped relationship between in-group bias in rewards and in-group bias in attribution.

A.4.2 Agents' choices

90% 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. 28% of agents also investigate the payoff consequences of a second minimal outcome demanded and 16% 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, centered around 4.

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 behaviour 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 reward.

Defining this measure as only the first click by an agent does not change the results of our analysis.

Figure A.2: Agents' inquiries of payoff consequences of expected demanded minimal outcomes

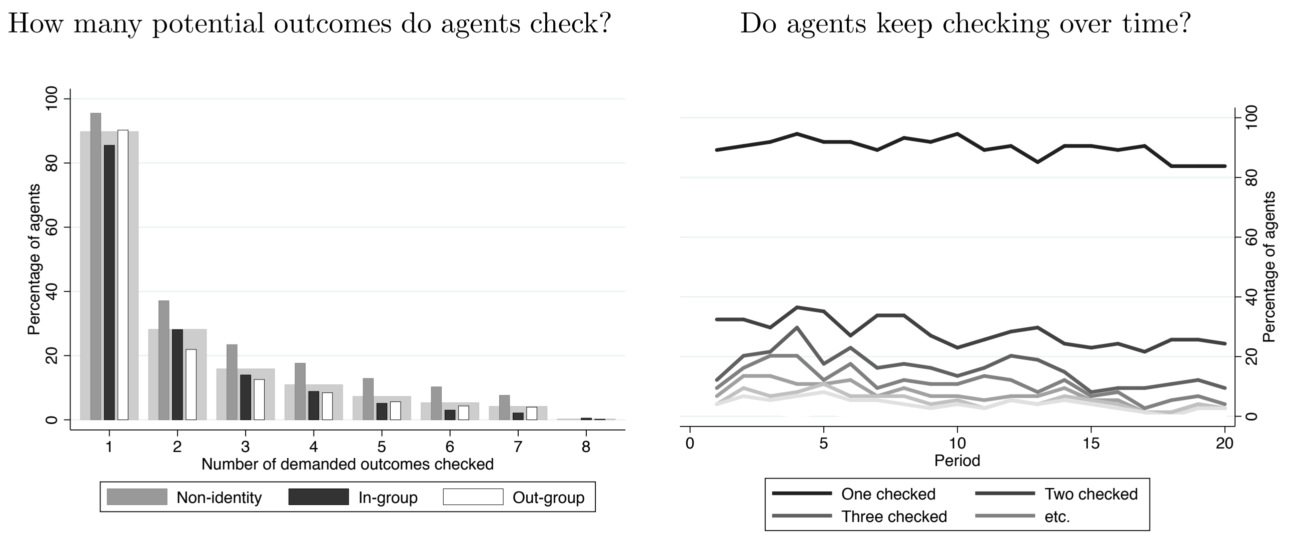


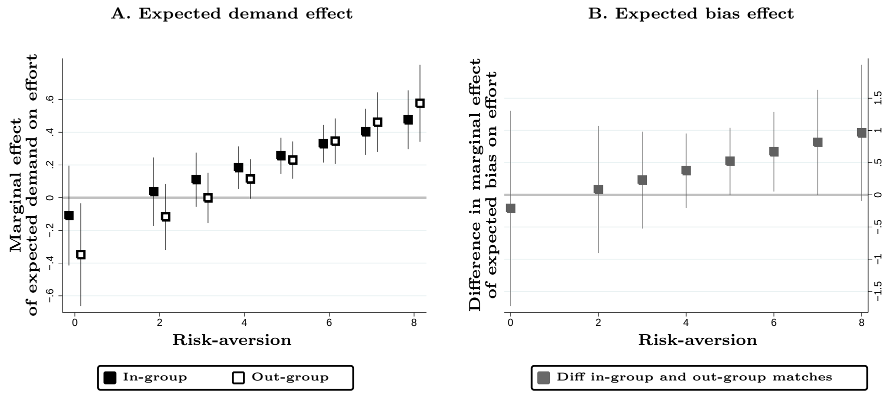
Table A.10: Regression of agents' effort on covariates for NON-IDENTITY and NON-STRATEGIC treatment. *Risk-aversion* is measured by the number of *safe choices* made in a (Holt and Laury, 2002)-list; four subjects with inconsistent choices moving through the list – switching back and forth between safe and risky option are excluded.

VARIABLES	(1)	(2)	(3)	(4)
in-group	0.07	-0.33	-0.09	-1.10
	(0.138)	(0.308)	(0.290)	(0.977)
type	-0.16***	-0.08	-0.07	-0.32
	(0.046)	(0.137)	(0.144)	(0.455)
expected demand		0.20**	0.29***	-0.27
		(0.075)	(0.083)	(0.199)
expected in-group bias			0.10	-0.57
			(0.215)	(0.952)
type × *expected demand*		-0.03	-0.04	0.06
		(0.035)	(0.036)	(0.105)
expected demand × *expected in-group bias*			-0.11**	0.16
			(0.048)	(0.262)
in-group × *expected demand*		0.17*	0.07	0.34
		(0.096)	(0.085)	(0.264)
in-group × *type*	-0.02	0.03	0.02	0.28
	(0.059)	(0.145)	(0.146)	(0.432)
in-group × *expected in-group bias*			-0.17	0.46
			(0.175)	(0.846)
in-group × *expected demand* × *type*		-0.03	-0.03	-0.10
		(0.042)	(0.043)	(0.115)
in-group × *expected demand* × *expected in-group bias*			0.12*	-0.22
			(0.067)	(0.273)
risk aversion				-0.31*
				(0.179)
in-group × *risk aversion*				0.25
				(0.179)
type × *risk aversion*				0.06
				(0.098)
expected demand × *risk aversion*				0.12***
				(0.041)
expected in-group bias × *risk aversion*				0.13
				(0.215)
type × *expected demand* × *risk aversion*				-0.02
				(0.023)
expected demand × *expected in-group bias* × *risk aversion*				-0.06
				(0.061)
in-group × *expected demand* × *risk aversion*				-0.07
				(0.048)
in-group × *type* × *risk aversion*				-0.07
				(0.078)
in-group × *expected in-group bias* × *risk aversion*				-0.15
				(0.169)
in-group × *expected demand* × *type* × *risk aversion*				0.02
				(0.021)
in-group × *expected demand* × *expected in-group bias* × *risk aversion*				0.09
				(0.059)
round	-0.00	-0.00	-0.00	-0.00
	(0.004)	(0.005)	(0.005)	(0.005)
Constant	2.11***	1.45***	1.21***	2.73***
	(0.131)	(0.258)	(0.263)	(0.877)
Observations	1,020	949	949	949
R-squared	0.033	0.152	0.180	0.240

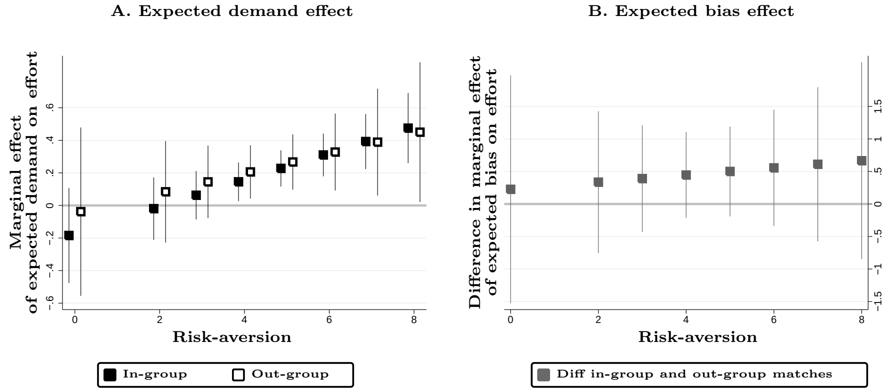
Standard errors clustered by subject in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Figure A.3: Marginal effect of expected demands on effort (= expected demand effect, Panel A) and difference in marginal effect of expected in-group bias on effort (= difference in expected bias effect, Panel B) over risk-aversion (number of safe choices in the Holt and Laury (2002)-list) by type. Marginal effects are estimated from Model 4 in A.10; standard errors bootstrapped with subject-level clustered errors.

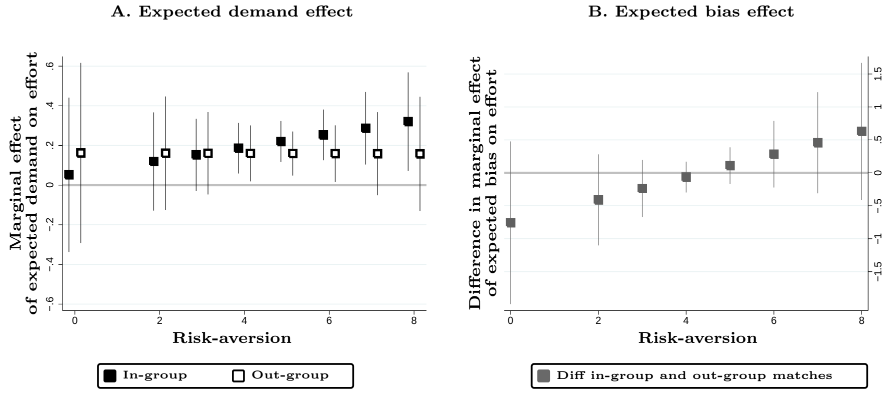
Low type



Medium type



High type



A.5 Average treatment effects: NON-STRATEGIC treatment

Table A.11: Logistic regression of attribution decision on indicators of treatment status, being classified as non-incentivising principals, in-group status, and high (good) outcome as well as the interactions of those variables and round of play. For non-incentivising principals and principals in the NON-STRATEGIC treatment, high outcomes are defined as those that are above > 4, in contrast to low outcomes (< 4. For incentivising Principals, good outcomes are defined as those that are above the principals individual reward threshold as defined in Section 5.2.1.

VARIABLES	
non-incentivising	-0.43
	(0.398)
NON-STRATEGIC	0.87*
	(0.524)
in-group	-0.32
	(0.196)
non-incentivising × *in-group*	0.89*
	(0.516)
NON-STRATEGIC × *in-group*	1.40*
	(0.740)
high (good) outcome	-0.66**
	(0.298)
non-incentivising × *high (good) outcome*	1.69***
	(0.629)
NON-STRATEGIC × *high (good) outcome*	-1.16*
	(0.659)
in-group × *high (good) outcome*	0.85***
	(0.294)
non-incentivising × *in-group* × *high (good) outcome*	-1.95**
	(0.819)
NON-STRATEGIC × *in-group* × *high (good) outcome*	-1.57*
	(0.833)
round	-0.02**
	(0.012)
Constant	0.60**
	(0.243)
Observations	1,229

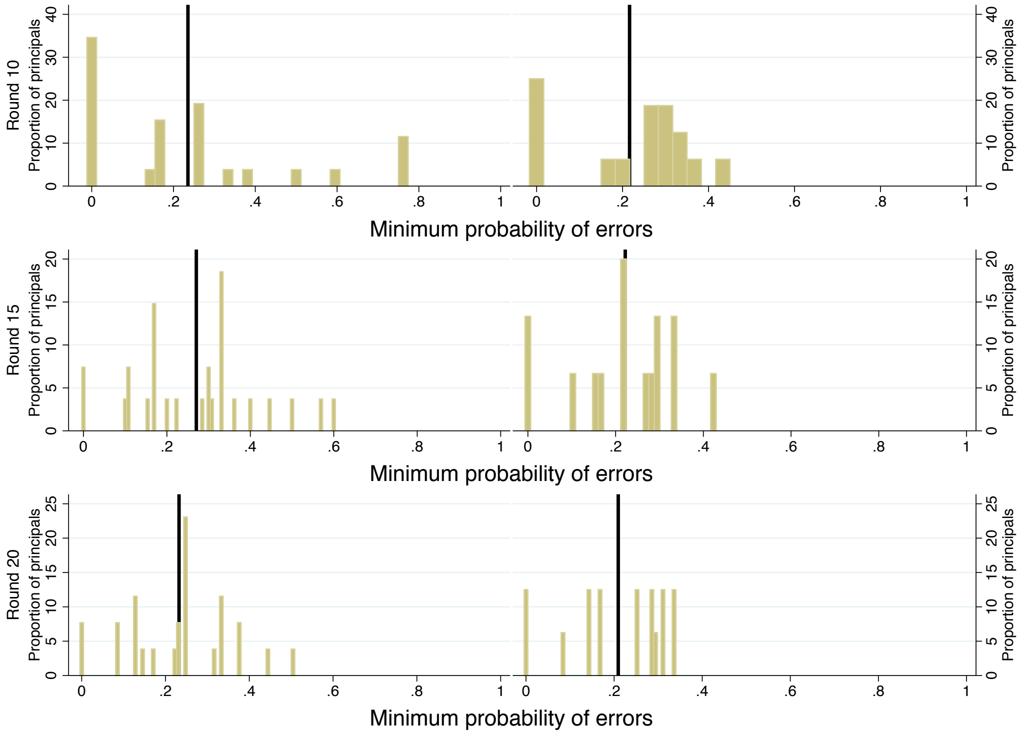
Standard errors clustered by subject in parentheses
*** p<0.01, ** p<0.05, * p<0.1

A.6 History of play

A.6.1 Principal's reward and attribution decisions

As expected, with increasing number of rounds played, the threshold in the outcome space above which incentivising principals are willing to reward the agent improves with respect to minimizing committed categorization errors. Figure A.4 shows a decrease in the spread of the probability of errors associated with the error minimizing threshold computed for each principal (while the mean remains constant); in other words, the computation of principals' thresholds becomes more precise with round of play. There is an element of noise we seem unable to pick up with our definition of each individual principals' threshold; the categorization error associated with the threshold that minimizes errors lingers around a probability of .2 of committing a categorization error.

Figure A.4: Distribution of the probability of an error in categorizing reward decisions associated with principal's reward threshold (= error minimizing threshold above which incentivising principals are willing to reward the agent).



Looking at principals reward decisions in the aggregate, we do not find a relationship of experience of favourable treatment in general and in in-group and out-group in particular; we express favourable past experience in current round t as the average outcome in round 1 to t-1. Table A.12 shows no significant effect of experience on current reward choices; here we model reward decisions as a function of outcome, favourable past experience (overall and separated by in- and out-group), the in-group status of the matched agent (applicable in the comparison STRATEGIC and NON-STRATEGIC treatment), the interaction of those variables, and round of play.

Table A.12: Logistic regression of principals' reward decisions on outcome, average of past outcomes, and a NON-IDENTITY treatment dummy for STRATEGIC and NON-IDENTITY treatment and, separately, for the STRATEGIC treatment but now reward decisions regressed on average of past outcomes in the in- and out-group.

VARIABLES	STRATEGIC and NON-IDENTITY	STRATEGIC
NON-IDENTITY	1.78	
	(2.517)	
outcome	0.34***	0.31***
	(0.073)	(0.085)
outcomes in the past	0.03	
	(0.248)	
in-group		-0.38
		(1.702)
outcomes in the past in the in-group		0.19
		(0.308)
outcomes in the past in the out-group		-0.11
		(0.221)
NON-IDENTITY × *outcome*	-0.05	
	(0.139)	
NON-IDENTITY × *outcomes in the past*	-0.31	
	(0.601)	
in-group × *outcome*		0.10
		(0.100)
in-group × *outcomes in the past in the in-group*		0.06
		(0.275)
in-group × *outcomes in the past in the out-group*		0.07
		(0.280)
Constant	-1.39	-1.72
	(1.082)	(1.497)
Observations	1,615	1,083

Standard errors clustered by subject in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Modeling the attribution decisions of incentivising principals as a function of outcome, the in-group status of the matched agent, and, similar to above, past outcome experience, shows that there is also no effect of a history of favourable experience with any agent, in-group agents, or out-group agents on the decision whether to attribute outcomes to effort. For our argument of the existence of strategic discrimination, behaviour among incentivising principals in the STRATEGIC treatment, because they accept to act in a strategic environment, and comparing those to principals in the NON-STRATEGIC treatment is the relevant counterfactual; Model (2) and (4) in Table A.13 gives the regression results for this comparison.

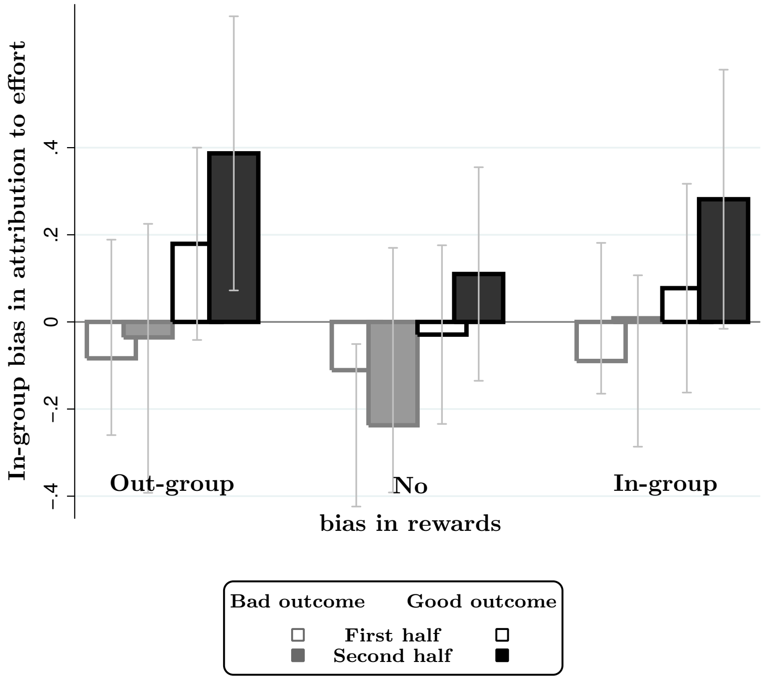
Table A.13: Logistic regression of principals' attribution decisions on outcome, in-group status of the matched principal, average of past outcomes in STRATEGIC, NON-STRATEGIC treatment, where the treatment-variable takes the STRATEGIC treatment as its base category, and in the STRATEGIC treatment on average of past outcomes in the in- and out-group separately; standard errors are computed based on clustering by subject. Model (2) and (4) exclude non-incentivising principals in the STRATEGIC treatment from the analysis.

		STRATEGIC and			
	All	NON-STRATEGIC			
VARIABLES	treatments	(1)	(2)	(3)	(4)
NON-IDENTITY	-0.59				
	(1.609)				
NON-STRATEGIC	-0.61	1.37	1.38	0.83	0.69
	(1.612)	(2.923)	(2.978)	(2.455)	(2.503)
outcome	-0.02	-0.02	-0.10	-0.02	-0.10
	(0.071)	(0.082)	(0.091)	(0.085)	(0.104)
outcomes in the past	0.01	-0.01	0.06		
	(0.200)	(0.231)	(0.232)		
in-group		-0.16	-1.41	0.87	-0.36
		(1.218)	(1.084)	(1.208)	(1.319)
outcomes in the past in the in-group				-0.02	-0.12
				(0.251)	(0.278)
outcomes in the past in the out-group				0.11	0.23
				(0.156)	(0.137)
in-group ×*outcome*		0.01	0.08	0.04	0.13
		(0.112)	(0.124)	(0.110)	(0.128)
in-group ×*outcomes in the past*		0.04	0.27		
		(0.297)	(0.267)		
in-group ×*outcomes in the past in the in-group*				-0.06	0.02
				(0.223)	(0.253)
in-group ×*outcomes in the past in the out-group*				-0.21	-0.08
				(0.187)	(0.225)
NON-IDENTITY ×*outcome*	-0.28**				
	(0.140)				
NON-IDENTITY ×*outcomes in the past*	0.35				
	(0.353)				
NON-STRATEGIC ×*outcome*	-0.46**	-0.44**	-0.36*	-0.54**	-0.46**
	(0.193)	(0.197)	(0.201)	(0.224)	(0.232)
NON-STRATEGIC ×*outcomes in the past*	0.70*	0.16	0.08		
	(0.365)	(0.599)	(0.605)		
NON-STRATEGIC ×*in-group*		-4.25	-3.04	-3.12	-1.93
		(4.072)	(4.060)	(4.126)	(4.192)
NON-STRATEGIC ×*outcomes in the past in the in-group*				0.42	0.52
				(0.485)	(0.501)
NON-STRATEGIC ×*outcomes in the past in the out-group*				-0.03	-0.15
				(0.317)	(0.309)
NON-STRATEGIC ×*in-group* ×*outcome*		-0.07	-0.15	-0.09	-0.18
		(0.251)	(0.257)	(0.230)	(0.239)
NON-STRATEGIC ×*in-group* ×*outcomes in the past*		1.18	0.97		
		(0.899)	(0.896)		
NON-STRATEGIC ×*in-group* ×*outcomes in the past in the in-group*				-0.25	-0.32
				(0.695)	(0.709)
NON-STRATEGIC ×*in-group* ×*outcomes in the past in the out-group*				1.18*	1.05
				(0.640)	(0.658)
round	-0.02*	-0.02*	-0.02*	-0.02	-0.02
	(0.010)	(0.011)	(0.012)	(0.014)	(0.016)
Constant	0.38	0.47	0.51	0.06	0.25
	(0.881)	(0.953)	(1.044)	(1.198)	(1.299)
Observations	1,995	1,634	1,330	1,412	1,161

Standard errors clustered by subject in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Even though learning, in the shape of forming beliefs about agents' behaviour given past experience with agents' performance (outcomes), seems not to exist, attribution to effort changes over round of play. The U-shaped relationship between in-group bias in rewards and in-group bias in attribution to effort for good outcomes becomes more pronounced in the second half of the experiment. Figure A.5 produces Figure 6 above, which is one of the crucial elements in our argument for the existence of strategic discrimination.

Figure A.5: Average difference in the rates of incentivising principals' attribution to effort between in-group and out-group matches (= in-group bias in attribution) over in-group bias in rewards of incentivising principals in first (round 1 to 10) and second half (round 11 to 20) of the experiment; 95% confidence bounds are shown based on a principal-level clustered bootstrap.



A.6.2 Agents's effort and expectations about principal's demands

Elaborating on the effect of history of play on agents, we see that agents' beliefs do not respond to individual agents' experience with reward decisions of their matched principals. Table A.15 shows no significant effect of the past rate of being rewarded overall, in in-group, or in out-group matches on agents' current expectations of principals' demands. There is, however, an effect of favourable treatment as out-group agent in the past in terms of principals' reward decisions on agent's current effort choice in the Strategic treatment (Table A.14). In particular, in the STRATEGIC treatment, the marginal effect of an increase in the rate of reward in the in-group in past rounds raises effort of agents in in-group matches by .46 (.12, .81). A rise in receiving a reward in the out-group increases effort in in-group matches (.41 (.02, .80)) and out-group matches (.44 (.03, .86)); marginal effects are estimated from Model (2) in Table A.14). Given that this relationship seems not to be related to a positively updated belief about the likelihood of receiving a reward from principals in the current round, we do not think that this finding takes away from our interpretation of strategic discrimination.

Table A.14: Regression of agents' effort on type, rate rewarded in the past in the STRATEGIC vs NON-IDENTITY treatment, where the treatment-variable takes the STRATEGIC treatment as its base category. And, regression of agents' effort on type, rate rewarded in the past in the in- and out-group separately, and in-group status of the matched principal in the STRATEGIC treatment.

	STRATEGIC and	STRATEGIC	
VARIABLES	NON-IDENTITY	(1)	(2)
NON-IDENTITY	0.12		
	(0.465)		
type	-0.22***	-0.19***	-0.18***
	(0.042)	(0.053)	(0.056)
rewarded in the past	0.40*	0.32	
	(0.208)	(0.257)	
expected demand	0.17***	0.14**	0.16***
	(0.038)	(0.054)	(0.051)
in-group		-0.20	-0.26
		(0.211)	(0.220)
rewarded in the past in the in-group			0.02
			(0.215)
rewarded in the past in the out-group			0.41**
			(0.199)
NON-IDENTITY × *type*	-0.12		
	(0.103)		
NON-IDENTITY × *rewarded in the past*	-0.21		
	(0.384)		
NON-IDENTITY × *expected demand*	0.06		
	(0.088)		
in-group × *type*		-0.06	-0.08
		(0.057)	(0.060)
in-group × *rewarded in the past*		0.20	
		(0.233)	
in-group × *expected demand*		0.07	0.05
		(0.058)	(0.052)
in-group × *rewarded in the past in the in-group*			0.45**
			(0.204)
in-group × *rewarded in the past in the out-group*			0.03
			(0.206)
round	-0.00	0.00	0.00
	(0.005)	(0.006)	(0.006)
Constant	1.34***	1.41***	1.24***
	(0.180)	(0.233)	(0.218)
Observations	1,521	1,164	1,032
R-squared	0.153	0.147	0.184

Standard errors clustered by subject in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table A.15: Regression of agents' expected demands on type, rate rewarded in the past in the STRATEGIC vs NON-IDENTITY treatment, where the treatment-variable takes the STRATEGIC treatment as its base category. And, regression of agents' effort on type, rate rewarded in the past in the in- and out-group separately, and in-group status of the matched principal in the STRATEGIC treatment; standard errors are computed based on clustering by subject.

	STRATEGIC and		
VARIABLES	NON-IDENTITY	STRATEGIC	
NON-IDENTITY	0.58		
	(0.567)		
type	0.27***	0.23***	0.22***
	(0.077)	(0.077)	(0.083)
rewarded in the past	0.44	0.48	
	(0.433)	(0.446)	
in-group		-0.27	-0.30
		(0.266)	(0.278)
rewarded in the past in the in-group			0.74
			(0.481)
rewarded in the past in the out-group			0.01
			(0.411)
in-group × *type*		0.08	0.13
		(0.118)	(0.133)
NON-IDENTITY × *type*	0.00		
	(0.127)		
NON-IDENTITY × *rewarded in the past*	-0.27		
	(0.730)		
round	-0.01	-0.00	0.00
	(0.009)	(0.011)	(0.014)
Constant	2.59***	2.65***	2.44***
	(0.398)	(0.430)	(0.548)
Observations	1,521	1,164	1,032
R-squared	0.051	0.034	0.049

Standard errors clustered by subject in parentheses
*** p<0.01, ** p<0.05, * p<0.1

A.7 Welfare effects

We find lower round payoff in the STRATEGIC than the NON-STRATEGIC and NON-IDENTITY treatments (See Table A.16). The treatment effect of a non-strategic environment is a decrease in round payoff by .70 (.45, .96) tokens from principals' out-group matches but only a drop of .55 (.24, .86) tokens from in-group matches. This effect is driven by the difference in round payoffs of principals in out-group matches. The treatment effect of a non-identity environment is .16, $(-.02, .34)$; these marginal effects are estimated from the regression in Table A.16 below. From Section 5.1 we already knew that average effort is much higher in the NON-STRATEGIC treatment and slightly higher still in the NON-IDENTITY treatment than in the STRATEGIC treatment (See Table 1.

Table A.16: Regression of subjects' round payoff on treatment, round of play, role of subject (dummy for principal), and the interaction of those variables for all treatments. And, the same regression for only the STRATEGIC and NON-STRATEGIC treatment also including in-group status of the matched partner on the right hand side.

VARIABLES	All treatments	STRATEGIC and NON-STRATEGIC
NON-IDENTITY	0.05	
	(0.156)	
NON-STRATEGIC	0.03	0.00
	(0.140)	(0.189)
round	0.00	0.01
	(0.007)	(0.009)
principal	-0.20	-0.13
	(0.130)	(0.204)
in-group		0.20
		(0.135)
principal × *round*	-0.00	-0.01
	(0.011)	(0.016)
in-group × *round*		-0.02*
		(0.012)
principal × *in-group*		-0.13
		(0.281)
principal × *in-group* × *round*		0.02
		(0.022)
NON-IDENTITY × *round*	0.01	
	(0.014)	
NON-IDENTITY × *principal*	0.17	
	(0.272)	
NON-IDENTITY × *principal* × *round*	-0.01	
	(0.023)	
NON-STRATEGIC × *round*	0.00	0.01
	(0.014)	(0.017)
NON-STRATEGIC × *principal*	0.66***	0.72**
	(0.220)	(0.332)
NON-STRATEGIC × *in-group*		0.05
		(0.214)
NON-STRATEGIC × *principal* × *round*	-0.01	-0.01
	(0.020)	(0.028)
NON-STRATEGIC × *in-group* × *round*		-0.01
		(0.019)
NON-STRATEGIC × *principal* × *in-group*		-0.13
		(0.451)
NON-STRATEGIC × *principal* × *in-group* × *round*		-0.00
		(0.034)
Constant	5.79***	5.68***
	(0.080)	(0.112)
Observations	3,760	3,000
R-squared	0.016	0.020

Standard errors clustered by subject in parentheses
*** p<0.01, ** p<0.05, * p<0.1

B Experimental design

B.1 Set up

Sessions were carried out at the Center for Experimental Social Sciences/NYU. Each experimental session lasted 20 rounds with 14-22 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 rounds.

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 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 realised *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*."

B.2 Group identity inducement

At the beginning of each session of both the STRATEGIC and the NON-STRATEGIC treatments, subjects were shown 5 pairs of paintings, with one painting by Paul Klee paired with one by Vassily Kandinsky, and were asked which painting they prefer in each pair. Based on which painter a subject preferred in a majority of pairs, he/she was assigned to be a *Klee* or a *Kandinsky*.[36]

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 the same identity group before everyone was asked for their final answer. Subjects within a group

[36] See Tajfel and Billig (1974), Chen and Li (2009), and Landa and Duell (2015) for the use of painter-preferences to induce identities in Social Psychology, Economics, and Political Science.

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.[37] (Members of both groups, Klees and Kandinskys, in all treatments performed approximately equally well.)

B.3 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.

[37] The 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 described here.

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 realised 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 realised 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 realised *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 realised *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 realised *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 realised *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	-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**

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 realised *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

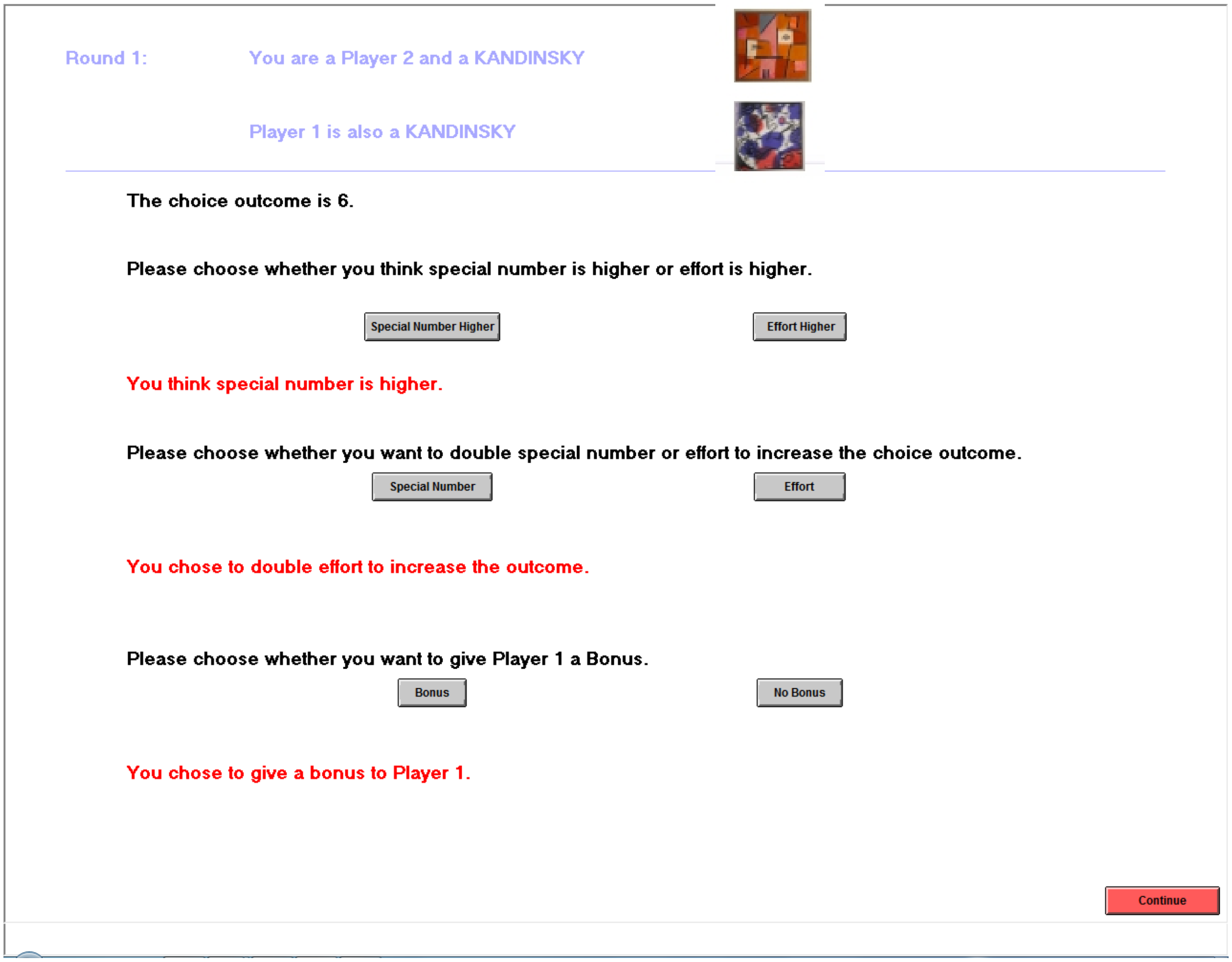
$$[2 \times 2] + 1 - 1 = \$4$$

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 2 and a KANDINSKY

Player 1 is also a KANDINSKY

The choice outcome is 6.

Please choose whether you think special number is higher or effort is higher.

You think special number is higher.

Please choose whether you want to double special number or effort to increase the choice outcome.

You chose to double effort to increase the outcome.

Please choose whether you want to give Player 1 a Bonus.

You chose to give a bonus to Player 1.

Table 1: Player 1's round payoff

Special		**Random**			
Number	**Effort**	**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**