Alleviating Discrimination*

Dominik Duell† and Dimitri Landa‡

December 14, 2018
Draft, please ask for the most recent version of this paper.

Abstract

In a laboratory investigation of a principal-agent relationship with moral hazard, we evaluate three interventions to alleviate discrimination based on group identity: (1) improving the principals' information regarding agents' effort; (2) creating uncertainty for the agents about the principals' identity; and (3) having principals announce a non-binding, identity-independent reward rule before agents' choices. All three interventions are, to varying degrees, effective in decreasing the principals' discriminatory actions and beliefs, but raise agents' suspicions of the principals and may lead them to expect even greater identity-contingent bias.

1 Introduction

Discrimination of specific ethnic, racial, or gender groups is widely implicated in the wage gap between men and women, and between whites and minorities (Altonji and Blank, 1999), the under-employment of blacks compared to whites (Western and Pettit, 2005) and the under-representation of women and minorities in legislative bodies of most Western democracies (Fox and Smith, 1998; Paxton, Kunovich and Hughes, 2007; Griffin and Newman, 2008). Although the evidence of the persistence of discriminatory patterns across the range of social, economic, and political areas is relatively robust and straightforward to document, those patterns often rest on a complex mix of individual and mutual beliefs, reinforced by statistical associations and focal strategic expectations. What interventions can be effective in dislodging these patterns remains little understood, despite the prominence of policy debates.

We report results from a series of experiments that model interventions seeking to reduce discrimination in *strategic* settings – settings, such as that of employers (principals) overseeing employees

*The research presented in this paper was supported by NSF Grant #SES-1124265, the NYU Research Challenge Fund Grant, and the ANR - Labex IAST.

†University of Essex

‡New York University

(agents), in which agents make choices in expectation of evaluation by principals, whose evaluations depend, in turn, on agents' responses to those expectations.

Discrimination – unequal treatment of persons who perform equally in a physical or material sense – has many sources, some directly, others indirectly connected to an observable characteristic such as race, ethnicity, or gender (Altonji and Blank, 1999; Holzer and Neumark, 2000). It may be driven by psychological factors such as prejudice (Allport, 1954) or, in economics parlance, a "taste for discrimination" against out-group members (Becker, 1971). However, it does not take a prejudice to create and sustain stereotypes generating discrimination (Phelps, 1972; Arrow, 1973). Situational factors, such as informational asymmetries that feed into statistical discrimination (Bertrand and Mullainathan, 2004; Goldin and Rouse, 2000; Knowles, Persico and Todd, 2001; Persico, 2002) and strategic expectations (Coate and Loury, 1993; Landa and Duell, 2015), often interacting with psychological factors, may also be influential in producing behavioral biases.

Perhaps the most frequently cited settings with discrimination are those best described as strategic in that individual actors' choices depend on their expectations of other actors' choices and vice versa. In those settings, discrimination may be a consequence of prejudice but it may also result from inferences about difference in performance arising from self-reinforcing beliefs about the choices by others that may be independent of the underlying distribution of group attributes (Arrow, 1973; Haan, Offerman and Sloof, 2015).

While the psychological causes of discrimination may be durable and difficult to shake, responses to situational factors are more calculated, and so, potentially more malleable. Policy interventions that focus on such factors may, thus, hold a particular promise. The interventions we analyze belong to that class. They target principals' beliefs about group-based differences in agents' choices, and, given the strategic feedback, the agents' beliefs about the principals' likely choices. They improve the quality of principals' information about the agents' choices, the coordination of mutual expectations, and the expected neutrality of oversight – specific measures that firms and organizations can take (and some have taken) to reduce the possibility of discrimination. These measures do not bind principals to particular non-discriminatory practices – and, in that sense, can be effective only insofar as the principals' underlying "taste for discrimination" is not too great – but they seek to close off channels that increase the likelihood of discrimination due to situational factors.

The overall pattern of our results suggests the following conclusions: First, the policy inter-

ventions we study are, in a laboratory setting, effective in checking the expression or formation of identity-based behavioral preference on the part of the principals, leading to a decrease in the identity-contingent biases in their beliefs and actions. But, second, and despite these successes on the principals' side, the effect on the agents is the opposite of what may be expected by the design of those interventions: they are not effective in decreasing, and may, in fact, increase, the agents' expectations of bias from the principals and the bias in their own choices. We provide further evidence suggesting that an increase in the agents' implicit mistrust of the principals' neutrality is accompanied by the agents' adopting a more transactional (less attachment-driven) disposition in their own choices.

2 Experiment

Sessions took place at the Center for Experimental Social Sciences/NYU and lasted 20 rounds with 14-22 participating subjects. Participants signed up via a web-based recruitment system that draws on a large pool of subjects from around the university and were not recruited from the authors' courses. The recruitment system blocked subjects from participating in more than one session of this experiment or similar experiments by the authors in the past. The experiments were programmed with the software z-Tree (Fischbacher, 2007) and subjects interacted anonymously via networked computers. After giving informed consent according to standard human subjects protocols, subjects received written instructions (as shown in the SI) 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. Before the principal-agent game stage commenced, subjects were asked questions concerning their understanding of instructions, 74% of participating subjects answered all questions correctly. 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 (Holt and Laury, 2002). Subjects received a show-up fee of \$7 and performance-based payments of on average \$23 for an experiment of 1 1/2 hours. Payments from the principal-agent game where taken from two randomly selected rounds. During the collaborative quiz in the identity inducement stage, a majority of members in both groups gave correct answers in four out of five painting quizzes. All session groups are balanced in risk-preferences as well as

in their (positive) experience in the collaborative quiz part of the group identity inducement (See Table S.3 in the SI).

The structure of our laboratory experiment approximates core aspects of the empirical principal-agent relationship between an employer and an employee or a voter and a representative. In a matched principal-agent pair, agents choose a costly effort and generate outcomes, which we model as the sum of chosen effort, agent's type (a randomly drawn integer), and a random noise draw. Principals observe the outcomes and decide whether they want to award agents a bonus – a special addition to agents' payoff. By monetarily incentivizing subjects in the role of agents, we create concerns about outcomes because agents value receiving a bonus from the principals. Subjects in the role of principals benefit from high outcomes and thus may want to incentivize agents to choose high effort.

We instantiate a principal-agent interaction in one *baseline* and three *intervention* treatments where the interventions aim to neutralize biases in beliefs and behaviors found in the baseline treatment. At the beginning of each round of the experiment, subjects are assigned to the role of either an *agent* or a *principal* and matched into pairs of one agent and one principal. They are randomly re-matched into pairs at the beginning of the next round but keep their role assignments throughout the course of the experiment. Baseline and intervention treatments consist of a group identity inducement stage and a principal-agent game.

2.1 Group identity inducement

At the beginning of each session, subjects were shown five pairs of paintings, one by Paul Klee and one by Vassily Kandinsky, and asked which painting they prefer in each pair. Based on their preference, subjects were assigned to be a *Klee* or a *Kandinsky* for the duration of the experiment.[1] Then, subjects participated in an activity within each identity group aimed at strengthening their attachment to these identities.[2] Unless noted otherwise, in the principal-agent game, the identities

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

[2] Considerable experimental literature has shown the effectiveness of the minimal group paradigm in inducing the patters of responses to identity, including in-group favoring discrimination, that resemble those usually observed outside the laboratory with naturally occurring group identities. (Chen and Li, 2009) and (Landa and Duell, 2015) provide evidence that "weak" induced identities significantly affect subject behavior with respect to their willingness to reward or punish in-group members across the range of strategically distinct settings. (Eckel and Grossman, 2005) show that the weakness of identity inducement does not bias results in the wrong direction.

of both subjects within a matched pair were displayed for them on the screen. Subjects, thus, learn whether they are in an *in-group* or *out-group* match. In our principal-agent interaction, social identities are exogenously tied to subjects' payoffs, which allows to elicit effects of identity, including subjects' responses to identity, without "feeding" them to the subjects.

2.2 Principal-agent game

In the baseline game, the sequence of moves in each round of the experiment is as follows:

1. Agents are assigned a *type* and privately informed about its realization (1, 2, or 3) drawn from a uniform distribution.

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 demanded outcome*).

3. *Noise* and *outcome* are realized where the value of *outcome* is the sum of agent's *type* (1, 2, or 3), agent's chosen *effort* (1, 2, or 3), and a *noise* realization (-1, 0, or 1) drawn from a uniform distribution.

4. Principals learn the value of *outcome* (1-7).

5. Principals choose whether to attribute outcomes to *type* or *effort* (*attribution decision*) and whether to give the agent a bonus *(reward decision)*.

When the principal thinks that effort was higher than type in generating the observed outcome we call this behavior *attribution to effort* and we label as *in-group bias in attribution* when the principal is more likely to attribute to effort in in-group than in out-group matches.

2.3 Baseline and policy interventions

The ***Baseline*** treatment implements the two-stage setting described above without amendment and we consider the following interventions:

The ***Observable Effort*** treatment provides principals with better (complete) information about agent's effort. In this intervention, principals cannot hold wrong beliefs about agent's effort and, in turn, agent's effort cannot condition on principal's (wrong) attribution of outcomes. If discrimination by the principals is a consequence of mistrust that is fed by the uncertainty over agents' choices, then this intervention should neutralize that effect. While in practice, this condition is not always possible to implement, often there are measures that principals can adopt to improve their

information about agents' choices.

The ***Announce Rule*** treatment departs from the *Baseline* in adding the announcement by the principals of an identity-independent reward rule before the agent makes her choice of effort. The intuition behind this treatment is that one of the factors contributing to discrimination may be mistrust due to the uncertainty about mutual expectation. Assuming that the principal does not deviate from his announced rule, the effect of the intervention would be to create a focal set of joint expectations. If the agent expects that the principal will not deviate, either because there is no upside to the principal from doing that or because deviating from the announcement creates a psychic cost for the principal, the agent's effort is less likely to be based on the expectation of bias in principal's choices. In turn, this would have an effect of weakening the principals' expectation that the agent chooses effort contingent on identity and eventually lower the bias in principals' own choices. Insofar as the empirically verifiable performance targets are meaningful and can be ex ante anticipated, the practical implementation in the workplace of the measures that are modeled by the *Announce Rule* intervention is straightforward: the companies would ask their supervisors to disseminate broadly the information about the relevant performance targets and make the promotion rules maximally transparent.

Finally, the ***Don't See ID*** treatment gives the agent no information about principal's group identity. The principal knows this, and so knows that the agent cannot condition her effort choices on whether her group identity matches the principal's identity. The expectation is that this will have the effect of weakening the principal's expectation of behavioral group differences resulting from agents' anticipation of bias in principals' reward decisions. This treatment follows the idea that if discrimination is a consequence of a strategically induced behavioral equilibrium, then weakening the discriminatory feedback from agents' choices by removing the possibility of conditioning the effort choice on the principal's identity should remove the asymmetry in principals' beliefs about agents' choices, and so remove that strategically induced reason for discrimination. One possible practical implementation of this intervention in the workplace would be as mixed-identity panels of supervisors with random post-performance assignments to evaluate an employee. Of course, as indicated in the Introduction, neither of the anticipated effects of the interventions may weaken psychological reasons for discrimination; insofar as the interventions prove successful, it is because they are targeting the values of the situational factors.

2.4 Equilibrium predictions of a simple principal-agent game

The game implemented in the laboratory approximates a simple model of incomplete contracting where a principal faces an agent with privately known competence (agent's type $t \in \{1, 2, 3\}$ with a commonly known uniform prior on that support). The agent chooses a costly effort level, $e \in \{1, 2, 3\}$ and the principal then observes the outcome $F = t + e + \omega$, where noise ω is a random draw from a uniform distribution on $\{-1, 0, 1\}$. After observing the outcome the principal decides whether to award bonus b, to the agent with the principal's payoff $u_r = G(F, b, e)$, which is $\beta\sqrt{F+1} - \alpha e$ if the bonus is awarded and $\beta\sqrt{F} - \alpha e$ if the bonus is not awarded. $G(\cdot)$ is, thus, increasing in F and b and decreasing in e. Simultaneously with choosing whether to award the bonus, the principal also chooses whether she wants to double the t or the e component (both unobserved) in the outcome function F. The principal's payoff is $F + De + (1 - D)t$, where $D = 1$ is the principal's decision to double e. The game ends when these payoffs are realized. While the Don't see ID- and the Announce rule-treatment do not differ in the interaction implemented in the laboratory from the structure of the simple model, in the Observable Effort treatment principals will not only learn the outcome before making their reward decision but also the by the agent chosen level of effort.

In the laboratory we set $\alpha = 1.95$ and $\beta = 6$ generating various sets of equilibria. In the baseline model, equilibria with the highest expected welfare for the principal, the principal awards a bonus if and only if $F \geq z$, $z \in \{3, 4, 5\}$, and the agent chooses effort e^* such that $e^* + t = 4$. These are pooling equilibria, and the principal's beliefs in these equilibria are such that she is indifferent between doubling e or t. Another set of equilibria with a threshold for receiving a bonus of $z \in \{1, 2, 6, 7\}$ are semi-separating, in that the principal's posterior beliefs about the agent's type is 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 $F < \hat{F}$. In both sets of equilibria, the principals' choices are contingent on outcomes they observe (= outcome-contingent-play, OCP, equilibria); principals whose strategies consist of rewarding choices that are outcome-contingent are called incentivizing principals and their strategies as incentivizing strategies. In a different kind of equilibrium, the principal awards the bonus independently of outcome (= outcome-noncontingent-play, ONCP) and the agents choose minimal effort, inducing partial separation through outcomes. Here, the principal will always prefer to double type. Given the payoff function, the principal will always prefer the pooling OCP equilibria

to the equilibria with semi-separation. That is, given the payoffs, the principal always prefers to obtain highest possible expected outcome F and live with more uncertain attribution to playing an equilibrium in which it is easier to make a correct attribution at the cost of a low expected outcome F.

The environment described in baseline game implemented in the laboratory is "identity-free;" inducing group identities does not alter the payoff structure describe above. One equilibrium behavioral expectation is, thus, that shared identity has no effect on behavior. However, because players observe social identity matches and there are multiple identity-free equilibria, the game with the identity treatment also admits identity-contingent "composite" equilibria in which different equilibrium profiles are played in different identity matches. In this way, identity matches could matter as "selectors" of different equilibrium profiles. The equilibrium predictions for the intervention treatments remain the same, except that equilibrium attribution behavior in the Observable Effort-treatment shifts slightly. In the Baseline, Don't See ID-, and Announce Rule-treatments, only observing the outcome, principals are equally likely to attribute outcomes to effort than to type assuming they follow the expected welfare maximizing incentivizing strategy of the OCP equilibria. In the Observable Effort-treatment, principals separately observe effort and type + noise; assuming that principals expect the welfare maximizing OCP equilibria being played, they should infer from seeing a given effort the type associated with it in these equilibria: low effort and high type, medium effort and medium type, and high effort and low type. Note, in equilibrium, attribution of outcomes to effort or type should not be identity-contingent as long as the randomly drawn types are balanced across identity match; Table S.2 in the SI shows that balance is given.

3 Results

To ascertain the effectiveness of policy interventions, we compare (i) principals' group-contingent biases in reward decisions and their attributions of outcomes to agents' effort in the *Baseline* and in the intervention treatments; and (ii) group-contingent biases in agents' effort choices and differences in their expectations of group-contingent biases in principals' rewards, in the *Baseline* and in the intervention treatments.

The systematically group-contingent attribution by a principal of the observed outcomes to her

matched agent's effort (a choice variable) rather than to her type (a feature of the agent fixed by the experimenter) would suggest the presence of a bias. When that attribution excuses lower outcomes in in-group matches by reference to the unchosen aspects and attribute the higher outcomes to chosen ones, while doing the reverse in out-group matches, the bias would be consistent with a phenomenon known as "the ultimate attribution error" and familiar from previous studies (Pettigrew, 1979) as a standard feature of discriminatory behavior.

Given that our primary interest is in how many individuals within an institution our interventions are able to reach, we quantify the change in bias as change in the share of subjects who show bias.[3]

3.1 Baseline

We are interested in the behavior of those principals who reward agents contingent on observed outcomes (in contrast to principals who always or never reward their matched agents). These are the principals whose choices, if anticipated by agents, could improve principals' outcomes, by creating for the agents the expectation that their effort may matter for obtaining a reward. We will refer to these principals as *incentivizing* principals.[4] In the *Baseline*, incentivizing principals are more likely to reward in-group agents than out-group agents, and they are more likely to show in-group bias in their attribution of outcomes to effort. The outcome principals demand to see to award a bonus is, on average, 3.96 for in-group agents but 4.53 for out-group agents, with the statistically significant difference of .57 (.17, .97) between the two. Also, incentivizing principals attribute good outcomes to the effort of in-group agents at a rate of .56 (.45,.66), while they do so of out-group agents only at a rate of .43 (.31,.54; difference = .13 (.03, .24)). Agents' effort is .14 (.02, .26) higher in in-group than in out-group matches when they expect the principal to be in-group biased in their reward decisions, i.e., to demand to see lower outcome to award a bonus in in-group than in out-group matches. If they do not expect to be favored by their in-group principal, their effort is not significantly different between in- and out-group matches (difference = -.08 (.04, -.20)).

The effect of introducing commonly known group identities into the principal-agent interac-

[3] In the SI, individual-level analysis is performed to verify that results are robust to outliers and provide additional insights about the functioning of the interventions.

[4] This subset of principals accounts for 76% of participating principals in the baseline, 78% in the *Don't See ID*-, 66% in the *Observable Effort*-, and 88% in the *Announce Rule*-treatments.

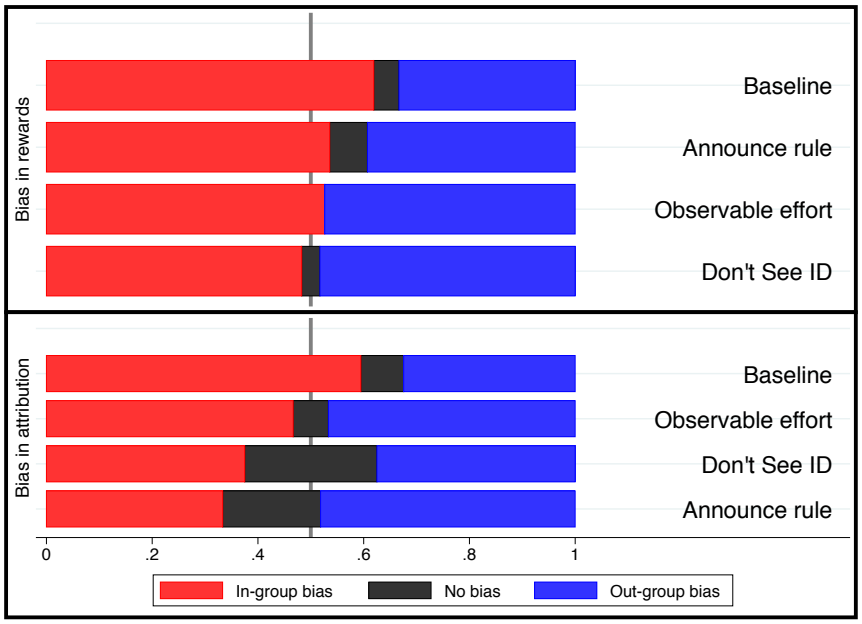

Figure 1: **Principals' biases:** Proportions of incentivizing principals by group bias in the reward decisions (*Bias in rewards*) and in attribution of good outcomes to effort (*Bias in attribution*), N = 118

tion is to increase the favorable treatment of in-group members and the unfavorable treatment of out-group members,[5] reinforcing the interpretation of these results as instances of identity-based discrimination.

3.2 Interventions – principals

The bottom panel in Figure 1 demonstrates the decrease in the prevalence of in-group bias in principals' attribution choices from *Baseline* to the interventions treatments. All treatment effects are significant at standard levels. The share of principals who are biased in their attribution of good outcomes to effort toward the in-group decreases from .60 (.47, .72) in the *Baseline* to .47 (.25, .69) in the *Observable Effort*-, to .38 (.23, .52) in the *Don't See ID*-, and to .33 (.19, .48) in the *Announce Rule*-treatment.

The effect of interventions on principals' bias in their reward decisions, represented in the top panel in Figure 1, follows in the direction of the changes in the principals' attribution choices, but is somewhat less pronounced. In all the treatment conditions, fewer principals show in-group bias

[5] In a version of the baseline treatment without group identities we find that principals demand for a reward bonus a value of outcome that is between the respective values for in- and out-group matches in the baseline reported here and show no bias in their attribute of good outcomes to effort.

than in the *Baseline*. However, the effect is significant at standard levels only in the *Don't See ID* treatment, in which the share of principals who are in-group biased in rewards is .48 (.37, .61) – a drop from .62 (.50, .73) in the *Baseline*. The corresponding shares of in-group biased principals in the *Observable Effort* treatment and in the *Announce Rule* treatment are .53 (.33, .72) and .54 (.40, .67), respectively.[6] We provide appropriate statistical tests as well as summary statistics and the full distribution of biases of principals (and agents) in the SI.

The reduction in reward bias is driven, in all interventions, by the more favorable treatment of out-group agents than in the *Baseline*, while changes in attribution to effort occur in both, in-group and out-group matches (See Figures S.5 and S.7 in the SI, respectively).

In summary, interventions reduce principals' bias in rewards and even more so in their attribution of good outcomes to effort. The least effective intervention has been to make the agent's effort perfectly observable, while obscuring the principals' identity or allowing principals to announce a non-binding reward rule before agents choose their effort are somewhat more promising.

3.3 Interventions – agents

While our intervention treatments prove to be promising in reducing discrimination in principals' choices that may be attributable to situational factors, agents' responses are of a different nature. Agents show slightly higher levels of bias in effort and expect higher bias in principals reward decision in the intervention treatments than the *Baseline*. The bottom panel in Figure 2 illustrates the steady increase in bias in effort when moving from *Baseline* to *Observable Effort*-, *Announce Rule*-, and *Don't See ID* treatments.

The share of agents who choose higher effort in the in-group matches (in-group biased in effort) is .40 (.31, .49) in the *Baseline*. It remains essentially the same in the *Observable Effort* treatment, but increases to .47 (.35, .59) in the *Announce Rule*- and .60 (.48, .71) in the *Don't See ID*-treatments (with the latter difference of .19 (.04, .35) being significantly different from zero).

Similarly, agents are more likely to expect in-group bias in the intervention treatments than in the *Baseline* (top panel of Figure 2). The share of agents who expect in-group biased principals

[6] In the *Announce Rule* treatment, we observe some principals deviating from their reward decisions in the announced rule. The deviations systematically favor the in-group agents (the difference in deviation from the average announced rule to the actual threshold a principal sets is higher by .28 (.03,.53) in in-group than in out-group matches). Despite this bias in deviations, the overall bias in rewards in this treatment is still lower than in the *Baseline*.

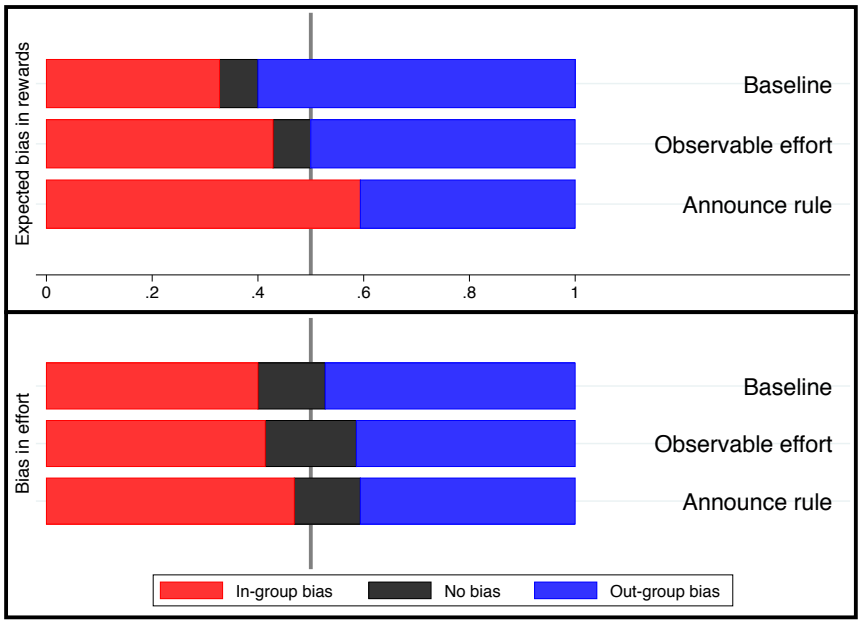

Figure 2: **Agents' biases:** Proportions of agents by group bias in effort choices (*Bias in effort*) and by expectation of principals to be in-group biased, out-group biased, or not biased in their reward decisions (*Expected bias in rewards*), N=153

increases from .33 (.23, .34) in the *Baseline*, to .43 (.30, .56) when principals perfectly observe agent's effort, to .52 (.40, .63) when agents do not know their matched principals' identity, and to .59 (.45, .74) when principals are able to announce a non-binding rule. The effects in the latter two intervention treatments are significantly different from zero at conventional levels, and in the *Observable Effort* treatment nearly so. Even though agents in the *Don't See ID* treatment do not observe principals' group identity, they are still situated in an in-group or out-group match from the principal's perspective and agents' choices reflect their beliefs about the expected response from a principal.

3.4 Interpreting agents' behavior

How can the interventions lead to an increase in agents' expectation of in-group bias in principal's reward decisions, given that they reduce that bias and the bias in principals' beliefs? To shed light on the mechanisms triggered by the interventions, we consider how agents' effort choices vary with their expectations: of the outcome their principal demands for reward and of the group bias in the principal's reward decisions.

Figure 3 shows agents' conditional effort levels in in- and out-group matches in the *Baseline*

and in the intervention treatments. The left half of the figure (in-group matches) shows that in the *Baseline*, effort is increasing with expected demands but also with the expectation of being favored by the principal (higher levels of effort (darker colors) can be found in the top left corner). The *Baseline* panel in the right half shows a complementary pattern in the out-group matches: effort is increasing when the agents are expecting the bias against them to be weaker. The comparison to the panels for the intervention treatments suggests a rather different underlying process there than in the *Baseline*. In in-group matches of the *Observable Effort*- and *Don't See ID* treatments, agents *decrease* their effort with the expectation of being favored by their principal in in-group matches (the change in color from light to darker color moving from top to bottom of the figure). In the *Announce Rule* treatment, the agents' responses to the expectation of favorable bias are non-monotone. In the out-group matches, it's the *Announce Rule* treatment that shows the strongest pattern opposite of what we see in the *Baseline*: agents increase, rather than decrease, their effort in expectation of being less favored by the identity-differing principals.

We interpret agents' increased expectation of in-group bias in principals' reward decisions in the intervention treatments as evidence of agents' *increased suspicion of the principals' disposition toward discriminatory actions*. The interventions focalize the agents' sense of the principals' susceptibility to identity-based behavioral preference, while, at the same time, casting that preference in a negative light.

As a response to the expectation of in-group favoritism, agents in in-group matches come to view group bias in the intervention treatments in more transactional terms: while in the *Baseline*, the identity affinity is a source of positive affect, encouraging a corresponding costly effort choice, in the interventions, it is (more) instrumental – a permission to choose lower effort when the bias is expected to be in one's favor because it is sufficient for expecting a reward, and an inducement to choose a higher effort when the expected bias is unfavorable to improve the prospect of getting a reward.

4 Discussion and conclusion

This study analyzed effectiveness of policy interventions seeking to reduce discrimination due to situational — informational and/or strategic — factors. While the psychological effects of exposing

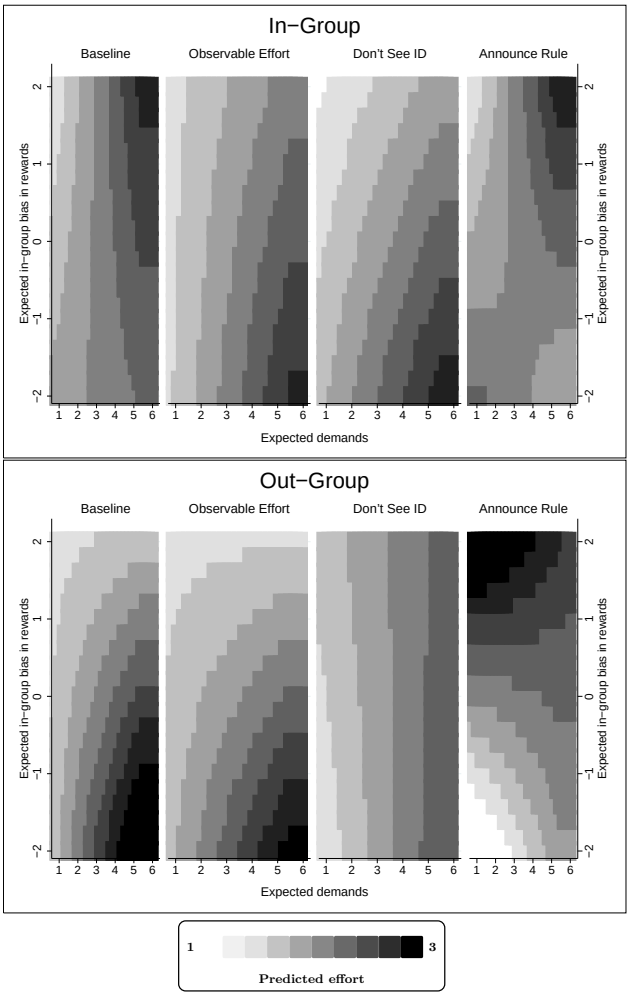

Figure 3: Predicted levels of effort over expected in-group bias in demands and expected demands for in- and out-group matches; estimates are taken from a regression of effort on agent's type, whether they share an identity with the matched principal, expected demands from the principal, expected in-group bias in principal's demands, and the interaction of these variables (Regression results are displayed in Table S.11 in the SI), N=153

subjects to elements of our treatments may point to other explanations of the treatment effects we describe, it is unlikely that those explanations are applicable in the setting considered here. First, while it has been shown that exposing individuals to information about out-groups reduces their biases (Fiske, 1998), providing additional knowledge about other group's behavior in our setting with minimal groups where subjects know equally little about in- and out-group, cannot be an instance of this effect. Second, taking the perspective of the member of a stigmatized out-group may reduce implicit bias (Galinsky and Moskowitz, 2000), the kind of perspective-taking we ask subjects to engage in – forming beliefs about others' behavior – targets explicit expectations in a strategic environment and not psychologically sustained implicit bias. Third, we cannot rule out that some of the observed reduction of in-group bias among principals is related to fewer opportunities to engage in reciprocity (Falk and Fischbacher, 2006; Rabin, 1993). However, while the *Don't See ID* treatment certainly lowers the ability to reciprocate given that agents do not learn principals' group identity, the *Observable Effort* treatment makes rewarding good behavior and punishing bad ones – the definition of following a norm of reciprocity – easier. Both treatments deliver similar changes in biases, suggesting that lower ability to reciprocate is unlikely as an explanation in our setting.

The results we present are a mix of good and bad news. The good news is that the interventions we evaluate appear to succeed in reducing principals' group-contingent attribution bias and differential reward decisions. In interpreting this result, it is important to consider its relationship to the possible presence of an experimenter effect meaning some or all behavioral effects reported here are not due to the content of the interventions but rather to subjects' providing answers in the presence of a figure of authority – the researcher running the experiment. There are two reasons why this concern about experimental validity is unfounded. First, the intervention effects we report are differences from the *Baseline*, which primes the nature of the experiment in similar ways as the *Don't See ID*- or the *Observable Effort*-treatments. The treatment that carries the highest potential for alerting subjects in the experiment to the fact that the study is about group biases is, arguably, the *Announce Rule* treatment. The results in that treatment, however, are very close to the results in the other treatments. To the extent that the presence of the observer mattered, we may consider the effect of that treatment to be an overestimate of what the effects would be without the observer, suggesting that the other two treatments may be more effective at decreasing the principals' discriminatory behavior. Second, the real-world phenomena the experiment approximates more often

than not occur in the presence of individuals, some in positions of moral, economic, or political authority, who are commonly understood to monitor interactions such as those we model in the lab. From this perspective, an experimenter effect describes simply the sense of being observed which one should want to model in the first place. To the extent that the presence of the observer had any effect, the conclusion that that effect weakened rather than strengthened the external validity of the experiment is not applicable generally.

The bad news from our results is that interventions drive a wedge in the expectations of principals and agents. The priming of the possibility of identity-based discrimination in the interventions appears to have raised agents' suspicions of and led them to expect even greater bias from the principals. In the substantive context of discrimination, this result is significant for at least two different reasons. First, to the extent that the expectation of discrimination has detrimental psychological effects on the concerned individuals, it suggests that measures that reduce discrimination should not be automatically assumed to reduce all of its negative consequences. And second, to the extent that agents' expectation of principals' bias feeds into agents' biased choices, those choices may have an eventual effect of undermining the salutary consequences of policy interventions on the principals, bringing back the vicious circle of discrimination re-enforced by the expectation of discrimination, itself induced by discrimination.

Turning, finally, to policy implications, the first broader lesson from our analysis is that successfully reducing discrimination may benefit from focusing on both psychological and situational factors that may contribute to it. The second lesson is policy interventions targeting principals' choices should include reaching out to agents. Anti-discrimination policies need to bring agents' expectations of principals' choices in line with those choices themselves, rather than focusing exclusively on measures to improve those choices.

References

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.

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.

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.

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

Falk, Armin and Urs Fischbacher. 2006. "A theory of reciprocity." *Games and economic behavior* 54(2):293–315.

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

Fiske, Susan T. 1998. "Stereotyping, Prejudice, and Discrimination. S. 357–411 in: Daniel T. Gilbert, Susan T. Fiske und Gardner Lindzey (Hg.): Handbook of Social Psychology.".

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

Galinsky, Adam D and Gordon B Moskowitz. 2000. "Perspective-taking: decreasing stereotype expression, stereotype accessibility, and in-group favoritism." *Journal of personality and social psychology* 78(4):708.

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. 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.

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

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

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.

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.

Rabin, Matthew. 1993. "Incorporating Fairness into Game Theory and Economics." *American Economic Review* 83(5):1281–1302.

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

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

Supporting information

1 Tables

Table S.1: Number of subjects, distribution of group identities, and number of observations by treatment (20 rounds for each subject).

Treatment		× of subjects	× of observations
	Klees	55	1100
Baseline	Kandinskys	55	1100
	Total	110	2200
Don't see ID	Klees	44	740
	Kandinskys	30	740
	Total	74	1480
Observable effort	Klees	33	580
	Kandinskys	25	580
	Total	58	1160
Announce rule	Klees	31	640
	Kandinskys	33	640
	Total	64	1280
Non-identity	Total	38	760
		344	6880

Table S.2: Difference in means and distribution of type in in-group and out-group match as faced by the agent across treatments; p-Value taken from Wilcoxon test.

Treatment	**Difference**	p-Value
Baseline	.04 (-.06,.14)	.42
Don't See ID	.04 (-.07,.16)	.45
Observable Effort	.02 (-.11,.15)	.77
Announce Rule	.01 (-.12,.14)	.83

Table S.3: Elicit risk aversion, positive group experience in collaborative painter quiz, and demographics across treatments. Risk aversion is measured as the number of safe choices in a standard, low stakes (Holt and Laury, 2002)-list. Positive group experience for a subject means to face a majority of in-group members (excluding the subject) who give the right answer in the collaborative painter quiz.

Variable	Baseline	Don't see ID	Observable effort	Announce rule	Difference Baseline vs treatments?
Risk aversion (scale 0 - 10)	4.83	4.95	5.19	5.14	No
% positive group experience	97.3	96.8	97.9	97.8	No
% Female	53.7	41.1	36.8	38.3	Yes (All)
% White	19.8	14.3	18.4	16.7	No
% Asian	57.3	80.3	63.2	58.3	Yes (Don't see ID)
Age	20.1	23.1	22.2	23.2	Yes (All)
% econ majors	25.0	10.7	15.8	18.8	Yes (Don't see ID)

Table S.4: Summary statistics: Means (standard deviation), minimum, and maximum values of type, effort, outcome, doubling decision (0 = type doubled, 1 = effort doubled), and bonus decision (0 = no bonus awarded, 1 = bonus awarded) by treatment

Variable	Baseline	Don't See ID	Observable effort	Announce Rule	Non-identity
Type	1.99 (.81)	2.00 (.79)	2.01 (.81)	1.97 (.83)	2.01 (.81)
Effort	1.76 (.79)	1.65 (.74)	1.61 (.80)	1.83 (.84)	1.76 (.84)
Outcome	3.69 (1.29)	3.68 (1.22)	3.62 (1.28)	3.74 (1.34)	3.81 (1.3)
Expected demand	3.43 (1.26)	3.422 (1.25)	3.24 (1.30)	3.71 (1.36)	3.77 (1.3)
Reward	.54 (.50)	.50 (.50)	.62 (.49)	.53 (.50)	.46 (.50)
Attribution	.55 (50)	.43 (.50)	.34 (.48)	.54 (.50)	.61 (.49)
Announced rule	-	-	-	4.66 (1.02)	-

Table S.5: Principal-level error-minimizing threshold in the outcome space categorizing the decision whether to award or not award a bonus by treatment (standard deviation is shown in parenthesis)

	Threshold in-group	Threshold out-group
Baseline	3.93 (1.24)	4.56 (1.22)
Don't See ID	3.90 (1.41)	3.89 (1.37)
Observable effort	4.06 (1.39)	4.35 (1.28)
Announce Rule	4.08 (1.20)	4.13 (1.27)
Non-identity	4.45 (1.56)	

Table S.6: Proportion of principals who are biased in their reward decisions; confidence bounds estimated from principal-level clustered bootstrap shown in parenthesis

	In-group biased		**No biased**		**Out-group biased**	
Baseline	.62	(.50,.73)	.33	(.22,.44)	.05	(.00,.09)
Don't see ID	.48	(.36,.61)	.48	(.35,.62)	.03	(-.02,.09)
Observable effort	.53	(.33,.72)	.47	(.28,.67)	0	.
Announce rule	.54	(.40,.67)	.40	(.26,.53)	.07	(.00,.15)

Table S.7: Proportion of principals who are biased in their attribution of good outcomes; confidence bounds estimated from principal-level clustered bootstrap shown in parenthesis

	In-group biased		**No biased**		**Out-group biased**	
Baseline	.60	(.51,.68)	.32	(.22,.43)	.08	(.03,.13)
Don't see ID	.38	(.24,.52)	.38	(.24,.51)	.25	(.11,.39)
Observable effort	.47	(.28,.66)	.47	(.27,.66)	.07	(-.04,.17)
Announce rule	.33	(.20,.46)	.48	(.33,.63)	.19	(.08,.29)

Table S.8: Proportion of agents who are biased in their effort; confidence bounds estimated from agent-level clustered bootstrap shown in parenthesis

	In-group biased		**No biased**		**Out-group biased**	
Baseline	.40	(.31,.49)	.47	(.38,.56)	.13	(.05,.20)
Don't see ID	.60	(.48,.71)	.35	(.24,.47)	.05	(.00,.11)
Observable effort	.41	(.31,.52)	.41	(.30,.53)	.17	(.07,.28)
Announce rule	.47	(.35,.59)	.41	(.28,.53)	.13	(.03,.22)

Table S.9: Proportion of agents who expect principals biased in their reward decisions; confidence bounds estimated from agent-level clustered bootstrap shown in parenthesis

	In-group biased		**No biased**		**Out-group biased**	
Baseline	.33	(.23,.42)	.60	(.50,.70)	.07	(.02,.12)
Don't see ID	.51	(.40,.63)	.46	(.34,.58)	.03	(-.01,.07)
Observable effort	.43	(.30,.56)	.50	(.37,.63)	.07	(.00,.15)
Announce rule	.60	(.45,.74)	.41	(.26,.55)	0	.

Table S.10: Difference in proportion of reward biased principals Baseline vs intervention treatments

	Don't see ID		**Obserable effort**		**Announce rule**	
Bias in rewards	**-.14**	**(-.29, .02)**	-.09	(-.29,.11)	-.08	(-.27,.11)
Bias in attribution	**- .21**	**(-.37, -.06)**	**-.16**	**(.32, .00)**	**-.20**	**(-.36, .04)**
Bias in effort	**.19**	**(.04, .35)**	.01	(-.17,.20)	.07	(-.07, .21)
Expected bias in rewards	**.19**	**(.01, .36)**	.09	(-.10, .27)	**.29**	**(.09, .44)**

Table S.11: Least square regression of agents' effort on covariates; standard errors are computed based on clustering by agent

VARIABLES	**Baseline**	**Don't see ID**	**Observable effort**	**Announce rule**
in-group	0.17	0.92**	-0.36	-0.25
	(0.311)	(0.371)	(0.364)	(0.377)
type	-0.03	0.09	-0.16	-0.18
	(0.155)	(0.132)	(0.136)	(0.176)
expected demand	0.31***	0.40***	0.24*	0.11
	(0.085)	(0.078)	(0.124)	(0.116)
expected bias	0.05	0.10	-0.00	0.72
	(0.266)	(0.124)	(0.489)	(0.536)
type × *expecte demand*	-0.04	-0.10**	-0.04	-0.02
	(0.039)	(0.039)	(0.041)	(0.046)
expected demand × *expected bias*	-0.09	-0.03	-0.07	-0.09
	(0.057)	(0.070)	(0.099)	(0.101)
ingroup × *expected demand*	0.00	-0.23*	0.13	0.00
	(0.089)	(0.133)	(0.122)	(0.108)
ingroup × *type*	-0.06	-0.48**	0.07	-0.02
	(0.151)	(0.195)	(0.183)	(0.216)
ingroup × *expected bias*	-0.19	-0.24	0.03	-1.08**
	(0.255)	(0.271)	(0.541)	(0.466)
ingroup × *expected demand* × *type*	-0.01	0.12*	-0.03	0.03
	(0.043)	(0.066)	(0.056)	(0.058)
ingroup × *expected demand* × *expected bias*	0.13	0.01	0.02	0.20*
	(0.089)	(0.081)	(0.135)	(0.101)
round	-0.00	-0.01*	-0.01*	-0.01*
	(0.005)	(0.005)	(0.007)	(0.007)
constant	1.10***	0.93***	1.60***	2.14***
	(0.288)	(0.265)	(0.443)	(0.483)
Observations	1,029	697	535	600
R-squared	0.165	0.210	0.181	0.112

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

2 Figures

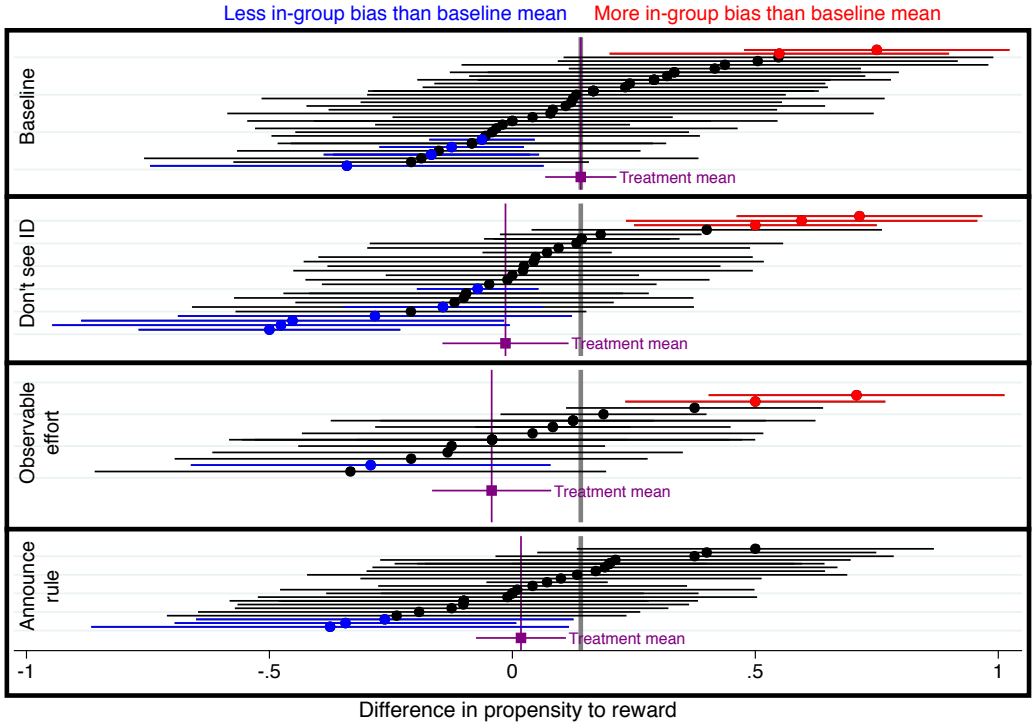

Figure S.4: **Distribution of bias:** Principal-level difference in reward in in-group and out-group matches (=*in-group bias in rewards*) with confidence bounds computed from an principal-level bootstrap of the difference; confidence bounds for *treatment mean* of in-group bias in reward are computed from an principal-level clustered bootstrap.

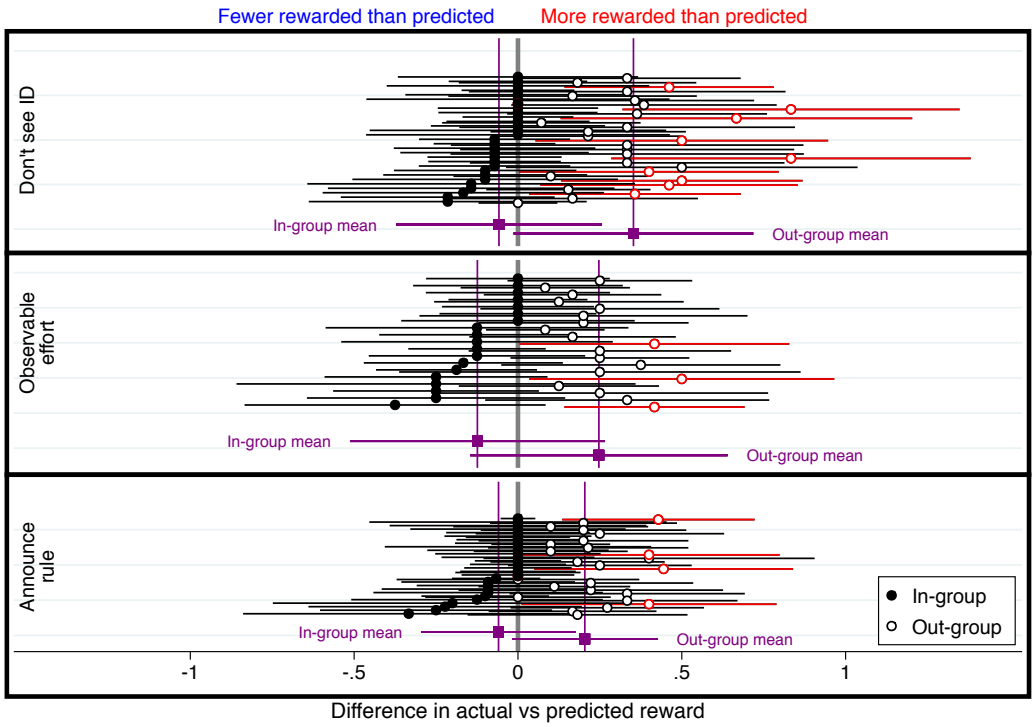

Figure S.5: **Treatment effects driven by a change in in-group or out-group behavior?** Principal-level difference in extrapolated reward predictions (=predicted reward) and in-sample reward predictions (=actual reward) in in-group and out-group matches with confidence bounds computed from an principal-level bootstrap of the difference. For each subject we show two estimates, one for behavior in in-group matches (solid markers) and one for choices in out-group matches (hollow markers). Red markers indicate that the particular subject rewards to effort more often and blue markers that the particular subject rewards less often than a similar subject in the baseline. Extrapolated reward predictions are computed based on the coefficients of a baseline treatment logistic regression of reward on outcome, in-group status, and their interaction and round of play applied to observations in the intervention treatments. In-sample reward predictions are based on the same regressions run on the intervention treatment sample itself.

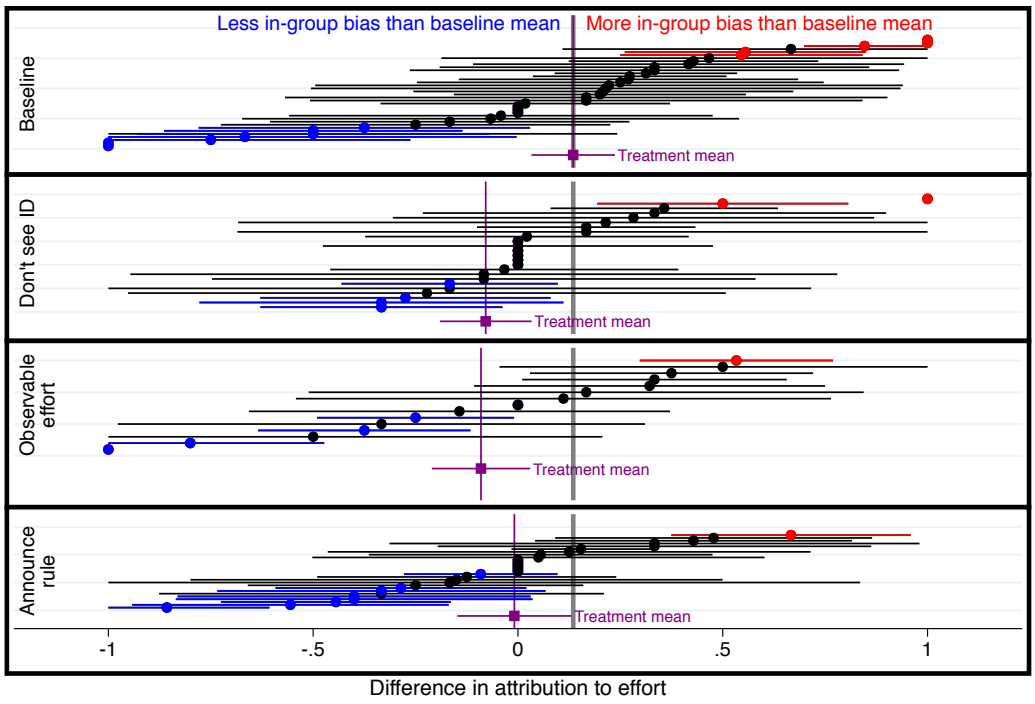

Figure S.6: **Distribution of bias:** Principal-level difference in attribution of good outcomes to effort in in-group and out-group matches (=*in-group bias in attribution*) with confidence bounds computed from an principal-level bootstrap of the difference; confidence bounds for *treatment mean* of in-group bias in reward are computed from an principal-level clustered bootstrap.

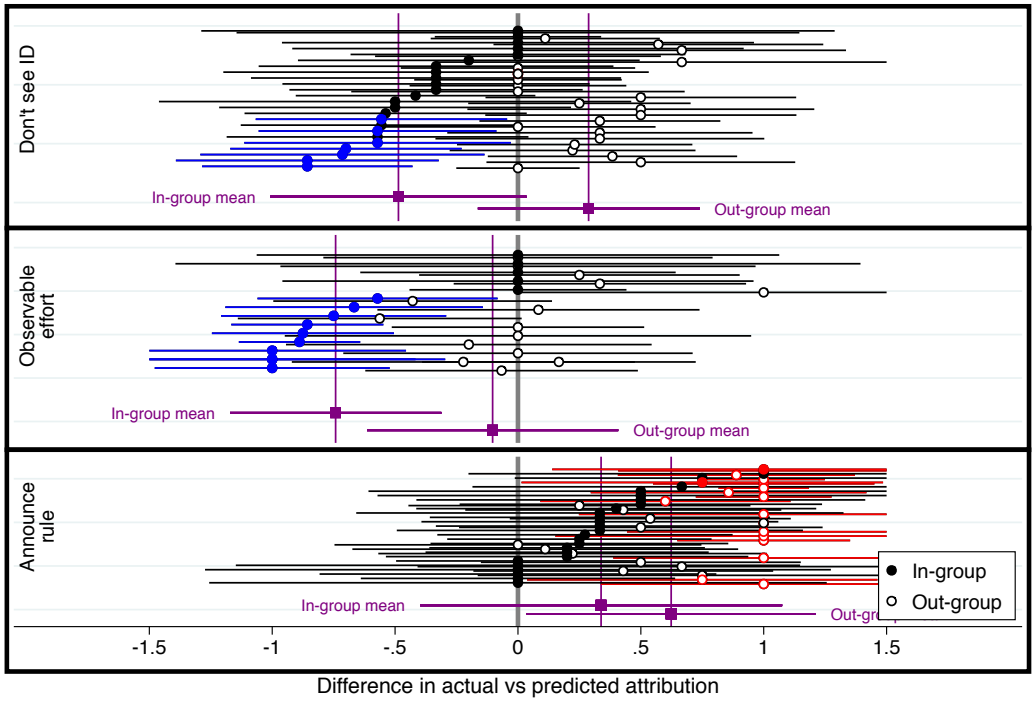

Figure S.7: **Treatment effects driven by a change in in-group or out-group behavior?** Principal-level difference in extrapolated attribution predictions (=predicted reward) and in-sample attribution predictions (=actual reward) in in-group and out-group matches with confidence bounds computed from an principal-level bootstrap of the difference. For each subject we show two estimates, one for behavior in in-group matches (solid markers) and one for choices in out-group matches (hollow markers). Red markers indicate that the particular subject attributes to effort more often and blue markers that the particular subject attributes to effort less often than a similar subject in the baseline. Extrapolated attribution to effort predictions are computed based on the coefficients of a baseline treatment logistic regression of attribution to effort on outcome, in-group status, and their interaction and round of play applied to observations in the intervention treatments for observations above the principal-specific reward threshold.

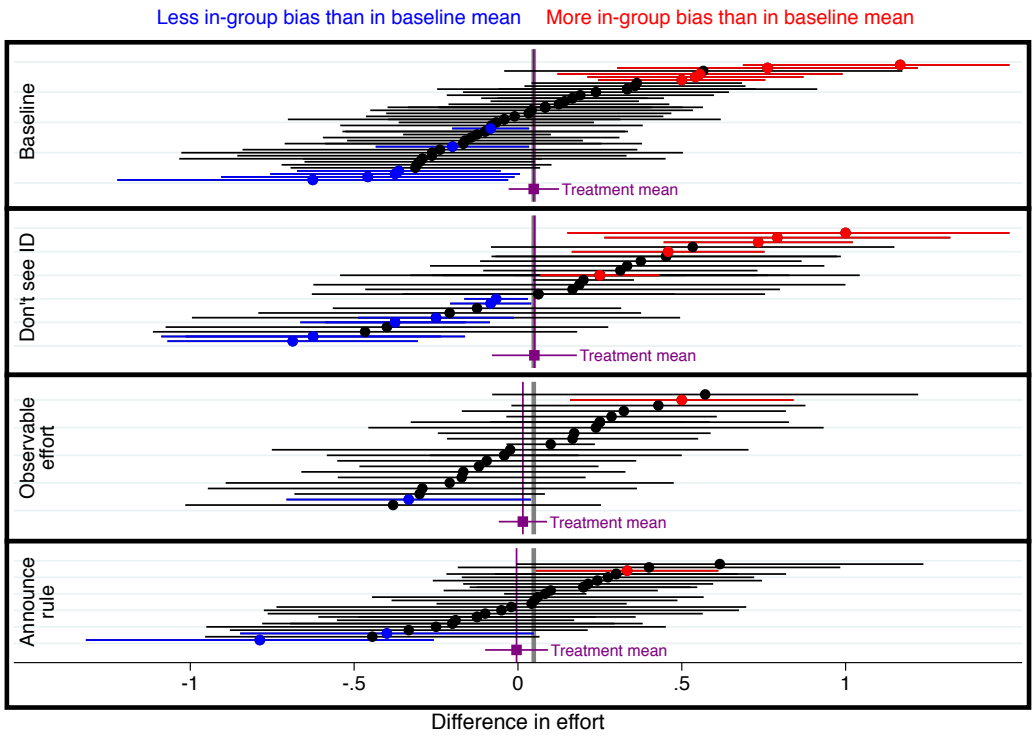

Figure S.8: Agent-level difference in effort in in-group and out-group matches (=*in-group bias in effort*) with confidence bounds computed from an agent-level bootstrap of the difference; confidence bounds for *treatment mean* of difference in effort are computed from an agent-level clustered bootstrap.

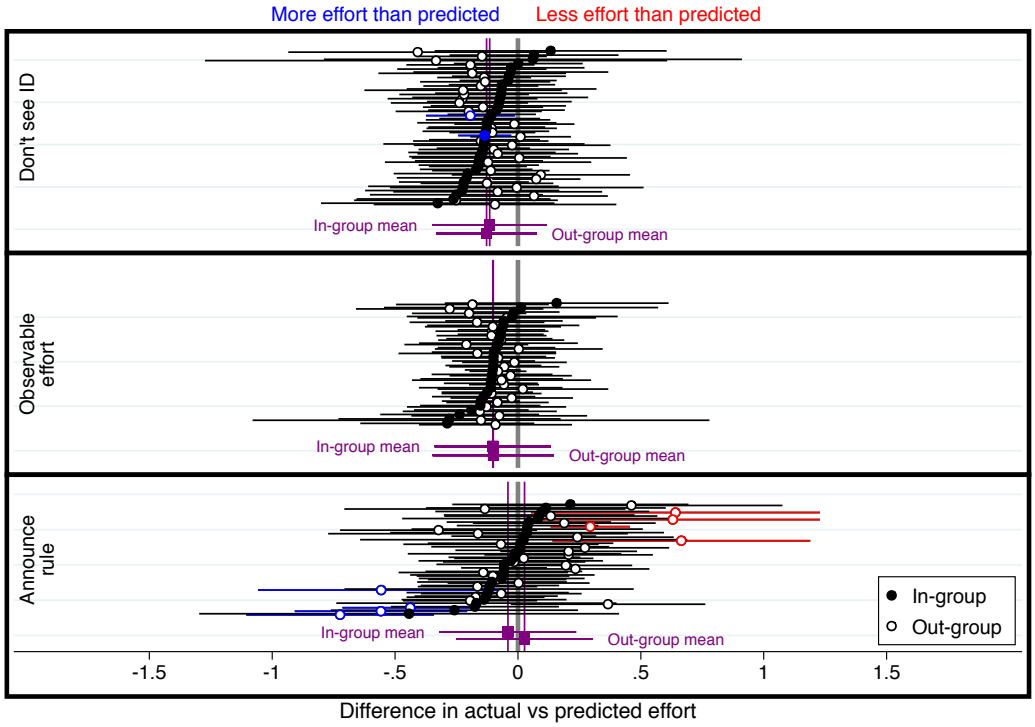

Figure S.9: **Treatment effects driven by a change in in-group or out-group behavior?** Agent-level difference in extrapolated effort predictions (=predicted effort) and in-sample effort predictions (=actual effort) in in-group and out-group matches with confidence bounds computed from an agent-level bootstrap of the difference. For each subject we show two estimates, one for behavior in in-group matches (solid markers) and one for choices in out-group matches (hollow markers). Red markers indicate that the particular subject chooses higher effort and blue markers that the particular subject chooses lower effort than a similar subject in the baseline. Extrapolated effort predictions are computed based on the coefficients of a baseline treatment linear least square regression of effort on type, in-group status, expected demanded outcome, expected in-group bias in rewards, and their interaction and round of play applied to the observations in the intervention treatments. In-sample effort predictions are based on the same regression run on the intervention treatment sample itself.

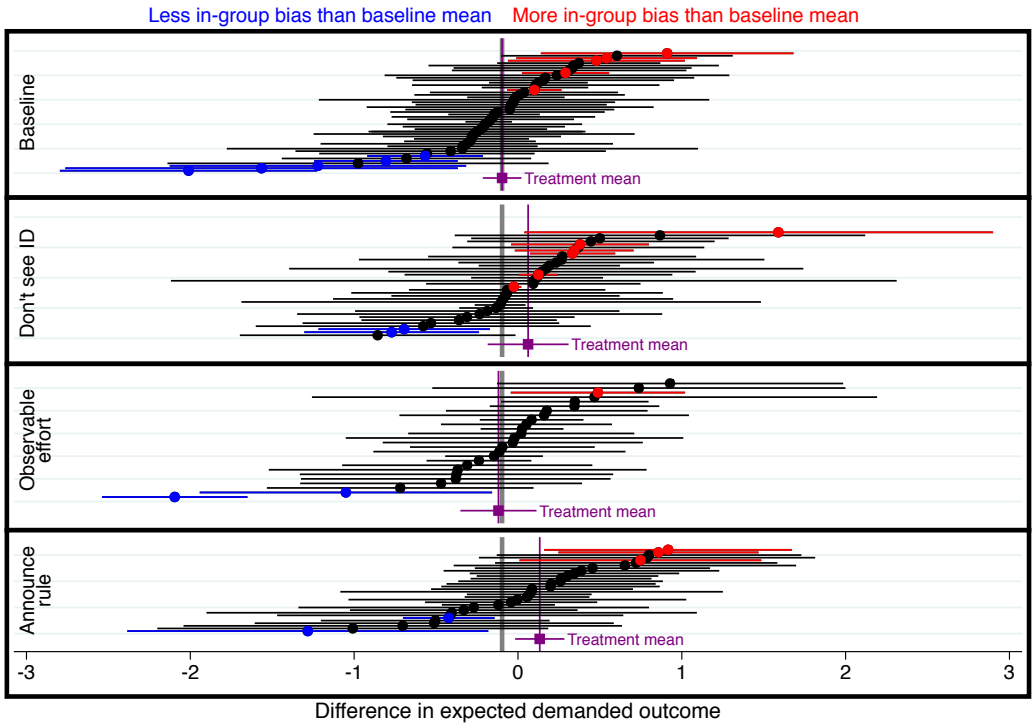

Figure S.10: Agent-level expectation of principal's bias in rewards (=*expected in-group bias in rewards*) with confidence bounds computed from an agent-level bootstrap of the difference; confidence bounds for *treatment mean* of expected bias in rewards are computed from an agent-level clustered bootstrap.

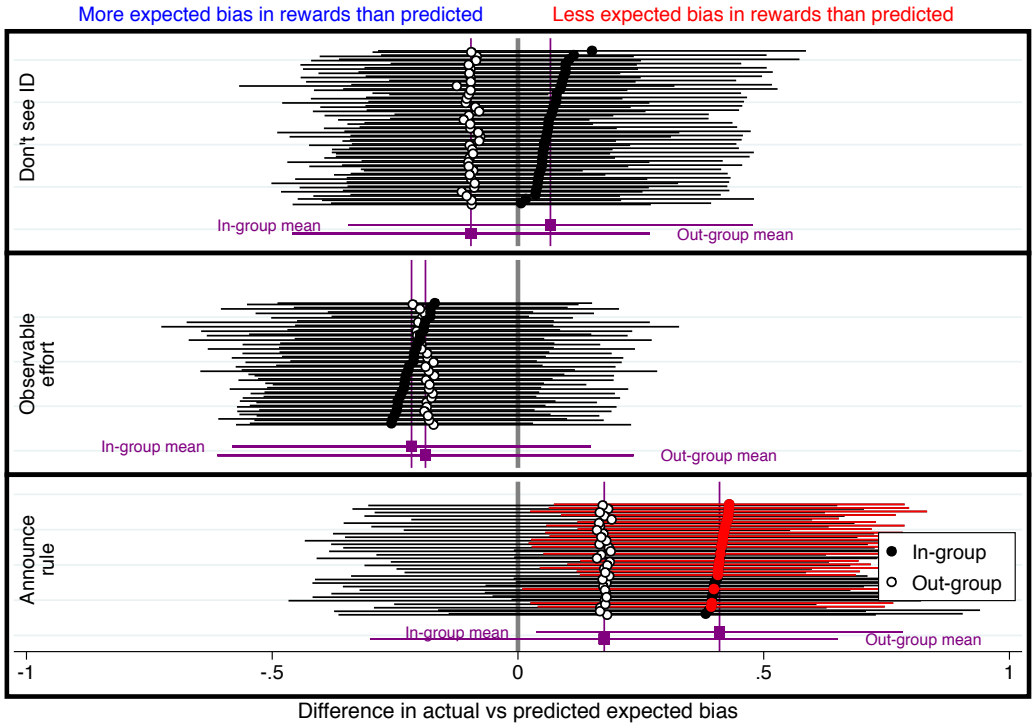

Figure S.11: **Treatment effects driven by a change in in-group or out-group behavior?** Agent-level difference in extrapolated expected bias predictions (=predicted expected bias) and in-sample expected bias predictions (=actual expected bias) in in-group and out-group matches with confidence bounds computed from an agent-level bootstrap of the difference. For each subject we show two estimates, one for behavior in in-group matches (solid markers) and one for choices in out-group matches (hollow markers). Red markers indicate that the particular subject expects higher bias in rewards from principals and blue markers that the particular subject expects lower bias in rewards from principals than a similar subject in the baseline. Extrapolated expected bias predictions are computed based on the coefficients of a baseline treatment linear least square regression of expected bias on type, in-group status, and their interaction and round of play applied to the observations in the intervention treatments. In-sample effort predictions are based on the same regression run on the intervention treatment sample itself.

3 Experimental set-up, instructions, and screens

3.1 Elicitation of principals' evaluation of agents' performance

From variation in principals reward choices with outcome, we are able to infer how they evaluate performance even if what constitutes *good* performance will depend on what the individual principal is trying to incentivize. To get at a valid measure of such performance evaluation for each of the incentivizing principals, we compute their individual-specific threshold values of outcome that minimize errors in categorizing bonus reward decisions. These threshold values provide natural individual-specific definitions of what outcomes a given principals perceives as good performance (at and above the threshold) versus bad performance (below the threshold). The inferred principal-specific thresholds vary from 2 to 7, with the average threshold being lower (3.93) in in-group matches than in out-group matches (4.56) in the *Baseline*. Table S.5 in the SI shows the mean of principals' thresholds in in-group and out-group matches across treatments.

3.2 Eliciting agents' beliefs about principals' reward decisions

We elicit agents beliefs of principals' reward rules. Before agents make their investment decision and after they observe their randomly assigned type, they are asked: "What minimal outcome do you think Player 2 will demand to give you a bonus?" Contingent on their answer and their type, they are given payoffs conditional on the level of effort they may choose and the possible values of noise. This information enables agents to aim for a more highly rewarded choice and is therefore indirectly incentivized monetarily. We take as measure of agents' beliefs the mean expected demanded outcome of all clicks they make in each round. Table S.4 in the SI gives the mean of agents' expected demanded outcome across treatments.

In 91% of subject-round observations, agents check at least one minimal outcome they expected to be demanded by their matched principals (95% in the first and still 85% in the last round). In 30% of subject-round observations, agents also investigate the payoff consequences of a second minimal outcome demanded and in 23% a third value. In the modal case – in 30% of the subject-rounds where agents check the first outcome – they obtain information about payoffs for a minimally required outcome of 4, the next highest-frequency outcome value checked is 3 (25%). 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 indicating that they are very specific in their expectation of the payoff information they want to obtain with variation in their behavior not in the number of clicks but only in which outcomes the investigate.

3.3 Instructions

Handed out to each subject in paper and read out aloud:

Introduction

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

This is an experiment on group decision-making. In this experiment you will make a series of choices. At the end of the experiment, you will be paid depending on the specific choices that

you made during the experiment and the choices made by other participants. If you follow the instructions and make appropriate decisions, you may make an appreciable amount of money.

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

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

Part 1

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

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

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

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

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

for Part 2.

Part 2

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

Matched group

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

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

Choices within each round of the experiment

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

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

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

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

After *the choice outcome* is computed, it will be shown to Player 2. However, Player 2 will not see Player 1's *special number* nor her choice of *effort* nor the realized value of the *random bump*.

After seeing *the choice outcome,* Player 2 will be given an opportunity to *increase* the outcome by doubling the contribution to outcome of either Player 1's *effort* or of her *special number* – whichever of those two Player 2 decides to increase. A new outcome will, then, be computed,

based on the corresponding *choice outcome*, but now increased because of the doubled contribution of *effort* or *special number*, as indicated by Player 2. We will refer to this new resulting outcome as *the increased outcome*.

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

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

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

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

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

Payoffs

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

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

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

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

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

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

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

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

increasing *effort*, then your payoff in that round is

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

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

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

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

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

If you have any questions, please ask them now.

Table 1: Player 1's round payoff

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

Table 2: Player 2's round payoff

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

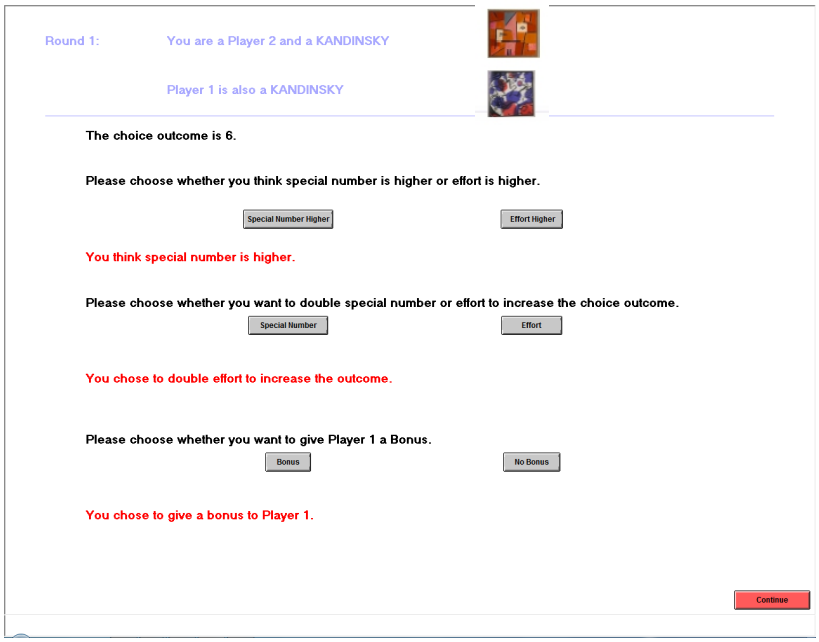

Figure S.12: Principal's decision screen

Figure S.13: Agent's decision screen (Shown in instructions to subjects)