Alleviating Strategic Discrimination*

Dominik Duell† and Dimitri Landa‡

September 3, 2020
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 systemic 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 and agents' expectation of principals' identity-contingent bias.

*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

1 Introduction

Systemic 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), 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), or the rate of incarcerations of African-American vs Whites (Loury, 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, embedded in various incentives created by institutions. 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 the part of systemic discrimination that arises in *strategic* settings – settings, such as that of employers (principals) overseeing employees (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 systemic 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, both directly, and, given the strategic feedback, through 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 that the policy interventions we study are, in a laboratory setting, effective in both (1) 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, and (2) decreasing the expectation of principals' on the part of the agents.

2 Experimental design

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.[1] 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, that can be used to incentivize them to choose higher effort.

[1] Detailed information about the experimental protocol are given in Section 2 of the Appendix.

This formulation of the principal-agent interaction follows a large literature in political economy (Persson and Tabellini, 2002; Ashworth, 2005; Gehlbach, 2006; Ashworth and Bueno de Mesquita, 2017).

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 Stage 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.[2] Then, subjects participated in an activity within each identity group aimed at strengthening their attachment to these identities.[3] Unless noted otherwise, in the principal-agent game, the identities 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 not directly tied to subjects' payoffs, which allows us to elicit effects of identity, including subjects' responses to identity, without "feeding" them to the subjects.

[2]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.

[3]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.

2.2 Stage 2: Principal-Agent Games

2.2.1 The Baseline Treatment

In the game instantiated in the *baseline* treatment, 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 (see point 3 below) principals demand to see to give a bonus (*expected outcome demand*).

3. *Outcome* is realized as 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)*.

The principal's utility is determined by two considerations: the outcome realized and whether the attribution decision was correct. The agent's utility increases in outcome and bonus and decreases in effort:

$$\begin{cases} \beta\sqrt{\text{outcome} + \text{bonus}} - \alpha(\text{effort}) \text{ if the bonus is awarded} \\ \beta\sqrt{\text{outcome}} - \alpha(\text{effort}) \text{ if the bonus is not awarded.} \end{cases}$$

In the experiments described below, the parameters are as follows: bonus $= 1$, $\alpha = 1.95$, and $\beta = 6$.

By monetarily incentivizing subjects in the role of agents, we create concerns about outcomes because agents value receiving a bonus from the principals. Subjects in the role of principals benefit from high outcomes. While the principals do not bear a direct cost of awarding the bonus, the agents' choices respond to the principals' bonus-awarding strategy. Because those choices affect principals' payoffs, they create a benefit to the principals of adopting a bonus-rewarding strategy that induces higher choices by the agents, as is standard in moral hazard settings.

The formal analysis of the incentives and equilibria in this game is provided in Duell and Landa (forthcoming). Here we highlight some of the details that are critical for formulating our hypotheses.

Because any reward rule by the principal can be sustained in equilibrium, there are many Perfect Bayesian Equilibria of this game. The multiplicity of equilibrium-consistent behavioral expectations, which induces a strategic coordination problem for the players, 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. Second, one of our primary interests is in understanding how introducing additional information into this environment, the nature of which we vary in our different treatments, can reduce uncertainty over mutual expectations.

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 a strategy that calls for rewarding if and only if outcome $\geq$ z, z $\in \{3, 4, 5\}$, and the agent chooses a level of effort e^* such that $e^* + t = 4$. These are pooling equilibria, and in these equilibria, the principal's beliefs are such that she is indifferent between attributing the outcome to agent's effort or type. Under the given the payoff structure, the principal always prefers to obtain the highest possible expected outcome, in spite of the greater uncertainty about attribution that that entails, than to play an equilibrium in which it is easier to make a correct attribution but at the cost of a lower expected outcome.

The uncertainty on the part of the subjects-as-agents about which of these cut-point strategies subjects-as-principals with whom they are matched are playing underscores the value of focusing on the predictions from the best-responses given a distribution over principals' possible demands (cut-points). The key such prediction is that, assuming that agents' participation constraints hold, agents choose higher effort when they expect the principal to be more demanding, but when the promise of bonus becomes very remote, the incentives created for the agent may be such that the optimal effort actually drops (i.e., the principal is "too demanding").

Note that the baseline game described above treats players' identities as irrelevant. This is because, as we emphasized above, in the experiment, we induce individual identities without building them into the payoff structure of the game – that is, they are simply ignorable. One possible equilibrium behavioral expectation is, then, that identity has simply no effect on behavior. However,

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

2.2.2 Policy Intervention Treatments

We consider the following interventions:

The *observable effort* treatment is a game that provides principals with better (in fact, 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.

Although this game is strategically distinct from the game in the baseline treatment, the outcome-contingent equilibria that we focus on in the discussion of the baseline treatment are equilibria in this setting as well and provide the most natural point of comparison.[4]

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

[4]Especially so insofar as we prime the subjects with the elicitation of their beliefs about the outcome threshold necessary for obtaining the bonus. That said, in this setting, the principal does better still in the equilibria in which she ignores the outcome entirely and awards the bonus if and only if the agent chooses the maximal effort, which is not possible in the baseline environment. The focalness of the outcome-threshold equilibria and the fact that our measures of interest concern within-treatment differences between behavior in in- and out-group matches, alleviate the concern about comparability that might otherwise arise from the existence of effort-threshold equilibria in this setting but not in the baseline. In a survey after the end of the experiment, we queried subjects about the rationale behind their decision-making in a series of structured and open questions. We find that the frequency with which subjects reported that they based their decision on "effort" or "outcome" is balanced across treatments. See Table S.4.

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.

The principal's announcement in this game is cheap-talk, but it can be informative. It is straightforward that for any cut-point strategy that maximizes the principal's utility in the baseline game, there exists an equilibrium of this game with the same cut-point strategy and truthful announcement of the bonus award rule corresponding to that strategy, and further, that there is no other truthful announcement that would correspond to an expected utility-maximizing equilibrium for the principals. Further, given the strategic uncertainty in this game, the announcement of a bonus-awarding rule can, clearly, be efficiency maximizing – for instance, it can, if believed, alleviate the likelihood that the agent chooses a low effort because they expect the principal to be too demanding. However, setting aside the psychic costs of principals' deviating from their announcements, the cheap-talk nature of the announcement implies that there is nothing to prevent the principals from choosing different identity-contingent cut-points for the actual bonus-award decision.

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.

Every identity-independent behavioral prediction from the baseline game has an equivalent prediction for this game, as well. Further, the possibility of principals setting identity-contingent cut-points for awarding the bonus is similarly present in this game. However, because agents cannot condition their effort choices on principal's identity, principals' attribution decisions should be expected to be symmetric across identity matches. Insofar as they are not, they are plausibly interpretable as evidence of what is known in psychology scholarship as the ultimate attribution error (Pettigrew, 1979) – see more on this below.

2.3 Hypotheses

Before stating our hypotheses, we formulate important variables and quantities of interest that are highlighted in the hypotheses. From principals' bonus award choices, we compute a principal-specific threshold of outcome that minimize errors in categorizing their respective reward decisions (*outcome demand).* The inferred principal-specific reward thresholds, whose distribution vary from 2 to 7.

When the principal thinks that effort was higher than type in generating the observed outcome, we call this behavior *attribution to effort.* When principals systematically excuse lower outcomes in in-group matches by reference to the agents' type (an unchosen agent characteristic) and explain the higher outcomes by agents' choice of effort, while doing the reverse in out-group matches – are marginally more likely to associate good outcomes from the out-group agent with factors that are not in the agent's control, and bad outcomes with such agent's choice of effort – they exhibit a familiar kind of bias. This bias is consistent with the "the ultimate attribution error" that is a standard feature of discriminatory behavior (Pettigrew, 1979; Hewstone, 1990).

We will refer to principals' higher attribution to effort of good outcomes in in-group than in out-group matches (positive in-group bias in attribution for good outcomes), as well as to principals' lower attribution to effort of bad outcomes in in-group than in out-group matches (negative in-group bias in attribution for bad outcomes) as *in-group biased attribution* to effort.

When agents choose higher levels of effort in in-group than out-group matches we refer to such behavior as *in-group biased effort*. When agents expect higher demands from out-group than in-group principals we refer to such beliefs as the *expectation of in-group biased outcome demands*.

As the baseline treatment has been analyzed in detail elsewhere, our analysis focuses on the intervention treatments, and in particular, on the comparison of those treatments to the baseline. We summarize the behavioral properties of the baseline in the following two observations developed in Duell and Landa (2021):

Observation 1 *(Principals' in-group bias in award and attribution decisions) Principals are more likely to believe that good outcomes are the result of agents' higher effort in in-group than out-group and are more likely to award agent a bonus in in-group than in out-group matches.*

Observation 2 *(Asymmetry in agents' expectation of principals' bias) Agents expect principals to be more lenient in their demanded outcome for awarding a bonus in in-group matches.*

As we explained above, out treatment interventions aim at undermining the potential for asymmetric attribution by the principal and therefore for decreasing the incidence of principals' biased reward choices. Our hypotheses are, thus, formulated in terms of the expectations of those effects relative to the baseline treatment. While, as we explained above, there are multiple equilibria in this setting, including equilibria indexed by different degrees of identity-contingent bias, we formulate our hypotheses as descriptions of what we expect to be the average tendency on the part of the subjects. Of course, none 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.

Hypothesis 1 *Making agents' effort observable reduces principals' asymmetric attribution and in-group favorable reward choices.*

Effort observability has two complimentary effects on principals' attribution choices: (1) it eliminates principals' uncertainty about the extent of the agent effort and (2) because it does so, it enables principals to arrive at "cleaner" expectations of agents' type (the expected difference between outcome and effort is no longer a pooling signal of agent's type). Consequently, we expect principals'

attribution judgments in this treatment to be not only less asymmetric (a "behavioral" prediction), but also more correct (an equilibrium prediction). This gives us an additional hypothesis:

Hypothesis 2 *Making agents' effort observable increases the correctness of principals' attribution choices.*

Our next two hypotheses are counterparts of the first hypothesis with respect to the other two interventions:

Hypothesis 3 *Allowing principals to announce their reward rule to agents before agents' actions reduces principals' asymmetric attribution and in-group favorable reward choices.*

Hypothesis 4 *Withholding identity information from agents reduces principals' asymmetric attribution and in-group favorable reward choices.*

As found in the previous work, agents' effort choices in response to expectations of outcome thresholds for receiving the bonus are highly heterogeneous, owing to their individual-specific senses of what makes for an expected demand that is "too high" (though we detail some of the variation in those choices below). However, the above hypotheses do have a "clean" counterpart in the agents' expectations of identity-dependent asymmetries in principals' demands:

Hypothesis 5 *Treatment interventions reduce the agents' expectation of the identity-contingent asymmetry in principals' outcome thresholds for awarding a bonus.*

3 Empirical strategy

3.1 Measuring bias and discrimination

In our experiment, principals are monetarily incentivized to obtain a higher payoff – that is, they earn more money when their rule for awarding a bonus to the matched agent is such that the agent responds to it by raising her effort, where that is feasible. The reward behavior of principals who set their bonus award threshold in the outcome space is potentially in-group biased through setting different demands for in-group and out-group agents. Principals who to not tie their bonus award decision to performance, in contrast, can only show in-group bias when they reward every in-group

agent but do not award a bonus to any out-group agent. We have shown elsewhere that in the principal-agent interaction modelled here, most subject in the role of principals play such threshold strategy (with a subset of principals showing in-group bias) while no subject differentiates between in-group and out-group agents in their reward decision independent of outcome (Duell and Landa, 2021).[5] We measure bias in principals' reward decisions by first eliciting their outcome demands and then compute whether these thresholds in the outcome space are lower for in-group than out-group agents. More specifically, we elicit an unique threshold for each principal from the 20 choices they are making over the course of the experiment. We find the cut-point in the outcome-space that best explains principals' decision to award a bonus. We call principals who are playing such cut-point strategy, *incentivizing* principals.

The threshold principals set separates outcomes into those they deem sufficient for awarding a bonus from those for which they do not reward the agent. We will refer to the outcomes at or above the principal-level threshold as *good outcomes* and to those below the threshold as *bad outcomes*. We then assess differences between instances of in-group and out-group matches in principals' attribution of outcomes to effort upon observing outcomes, separated by whether the outcome is good or bad. The extent of such differences is our measure of principals' beliefs whether the observed outcome emerged due to higher values of agents' effort than agents' assigned type.

We find evidence of principals' bias when principals' reward thresholds are lower for in-group than out-group agents as well as when their attribution to effort is more frequent for in-group than out-group agents upon observing good outcomes and less frequent upon observing bad outcomes.

To measure agents' perception of bias, we elicit agents' expectations of their principals' demands in each round of the experiment and check for differences in in-group vs out-group matches.

3.2 Testing for the capacity to alleviate discrimination

We quantify the alleviation of discrimination when comparing baseline and intervention treatments by a reduction in principals' biases in beliefs (attribution) and behaviors (reward rules) and the alleviation of the agent's expectation of discrimination by a reduction, in the treatments, in the gap between the agents' expectations of the principals' reward thresholds. In particular, our evidence

[5] In the experiments shown here, 76% of principals in the baseline-, 78% in the don't See ID-, 66% in the observable effort-, and 88% in the announce rule-treatment play a threshold strategy.

will support our main Hypotheses 1, 3, and 4 when the difference in principals' thresholds in in-group and out-group matches decreases significantly with an intervention treatment over the baseline, and when the principals' in-group bias in attribution to effort is significantly smaller with an intervention than in the baseline treatment.

With respect to the hypothesis we formulated about agents' beliefs (Hypothesis 5), we test by estimating the difference in agents' elicited beliefs about principals' in-group bias in rewards (difference in principals' threshold in in-group and out-group matches).

Beyond the effect on bias resulting from changes to the strategic rationale, the manipulation of the information environment in the *don't see ID* treatment (agents do not learn principals identity) should already prevent agents from differentiating between in- and out-group principals. Indeed, in this treatment, agents' effort choices and the expectation of principals' bias are not significantly different in in-group than out-group matches.

We implement hypothesis tests over treatment effects in a regression framework. The precise regression specification on the different outcome measures is described in the result section in more detail whenever we report a result from such regression. Regression results are, additionally, presented in the appendix (Tables S.6-S.9).

4 Experimental Results

We collect 6120 subject-round observations on 306 subjects in the baseline and three intervention treatment conditions in 17 experimental sessions with 14-22 subjects each.[6]

4.1 Detecting discrimination by the principals

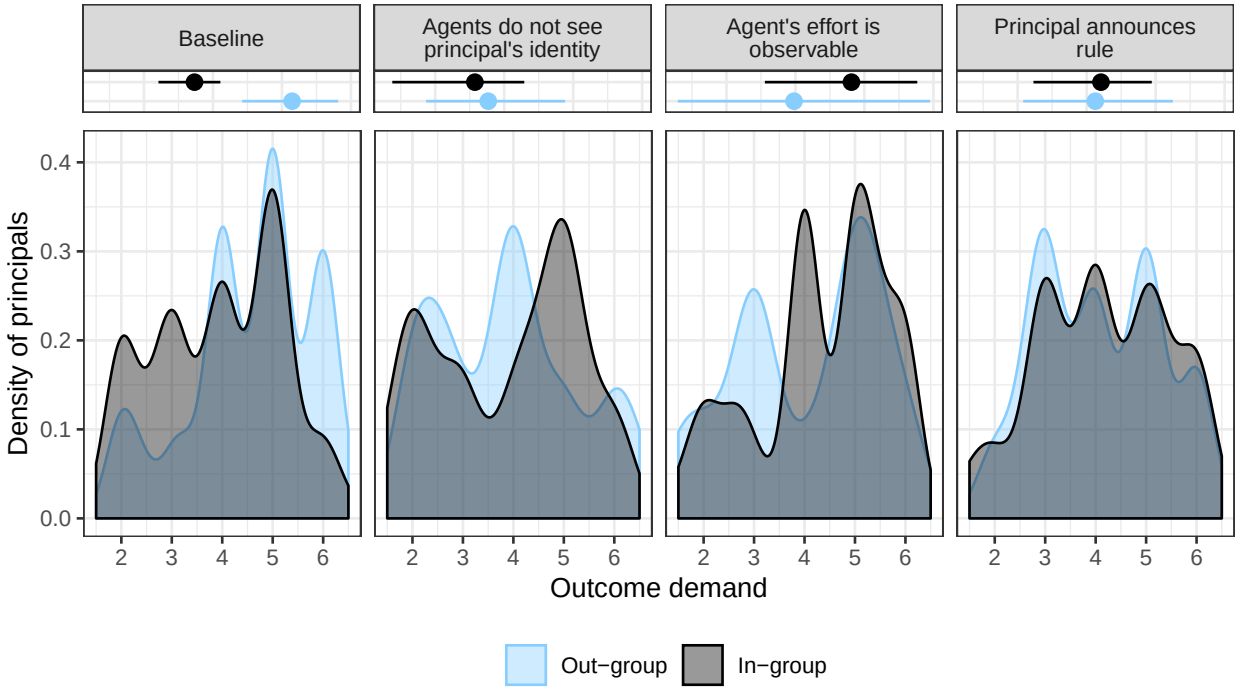
As summarized in Observation 1 above, in the *baseline* treatment condition, incentivizing principals are more likely to reward in-group agents than out-group agents, and they are more likely to show in-group bias in attribution of outcomes to effort. In particular, the outcome principals demand to see to award a bonus is, on average, 3.96 for in-group agents but 4.53 for out-group agents,

[6] Assignment to treatment is balanced in subjects' level of risk aversion (elicited in a Holt and Laury (2002)-list prior to the experiment, in their positive group experience in the collaborative painter quiz, and the share of white subjects. Fewer women and older subjects were randomly assigned to the intervention than the baseline treatment. Further, the share of Asian students was higher and the share of economics major students lower in the *don't see ID* vs the baseline treatment. See session, balance, and summary statistics by treatment condition in Section 1.1 of the appendix.

with the statistically significant difference of .57 (.02., 1.11) 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) with $p < .01$ in a difference-in-means test). Consistent with Observation 2, agents in the *baseline* treatment expect principals' reward thresholds to be lower for in-group than for out-group agents: the average expected in-group bias by principals is positive (.14 (.05, .27)).

All interventions decrease principals' bias in their reward decisions, represented in Figure 1 by mean and distribution of principals' outcome demands for in-group and out-group agents. In all intervention treatments, principals' in-group bias in reward decisions disappears; that is, the difference in outcome demand is only significantly different from zero in the baseline ($p < .05$).

Figure 1: Distribution of principals' outcome demands by in-group status and treatment

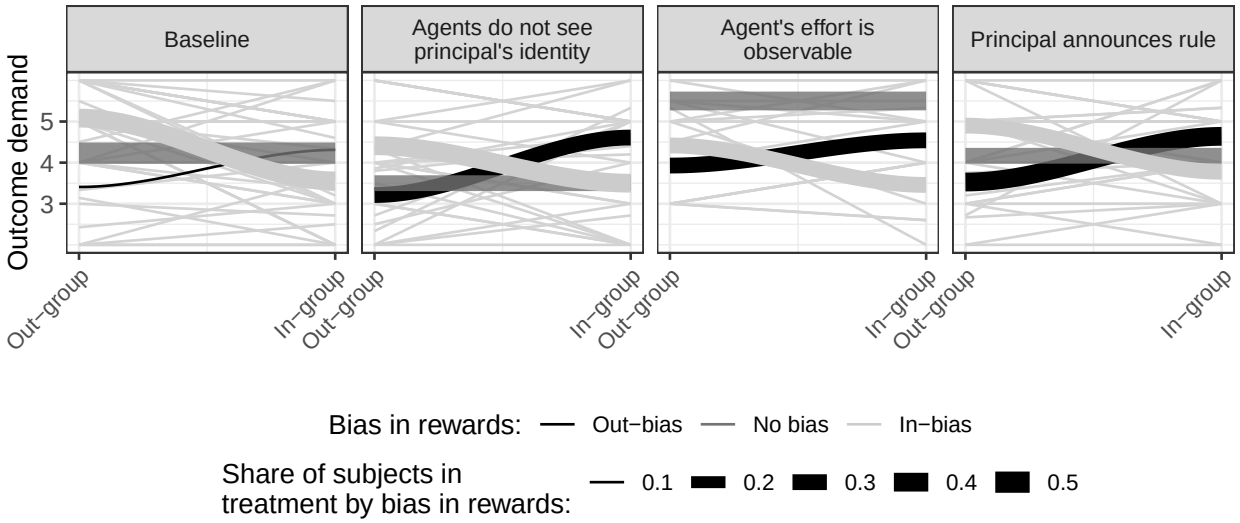


Compared to the difference in reward thresholds between in-group and out-group principals of .57 in the *baseline*, it is .10 (–.64, .85) in the *don't see ID* treatment, 0.12 (–.79, 1.07) in the *observable effort*, and 0.01 (–.65, .68) in the *announce rule* treatment. In other words, there is no discernible difference in reward thresholds in the interventions while there was in the *baseline* treatment. The reduction in in-group bias in reward thresholds in the intervention treatments over the *baseline*

is modest with p = .09, p = .09, and p < .05 for the comparison of *baseline* with *don't see ID*, *observable effort*, and *announce rule* treatments, respectively.[7]

Exploring outcome demands by in-group status and treatment condition more thoroughly, Figure 2 reveals, first, that there is variation across principals in whether they are in-group biased (their reward thresholds are lower for in-group than for out-group agents) or out-group biased (they demand lower outcomes from out-group than in-group agents to award a bonus, or they do not show a bias in rewards). Second, looking across treatment conditions, we see that the reduction in reward bias is driven, in all interventions, by the more favorable treatment of out-group agents than in the *baseline*; the mean and distribution of outcome demands for out-group agents move downwards in all interventions in contrast to the baseline while they increase for in-group agents only in the *observable effort* treatment. Further, the share of out-group biased principals – principals demonstrating favorable behavior towards out-group than in-group agents – is almost non-existent in the *baseline* but constitutes more than 30% of principals in each intervention treatment.

Figure 2: Subject-level and treatment average of principals' outcome demands over in-group status by treatment.

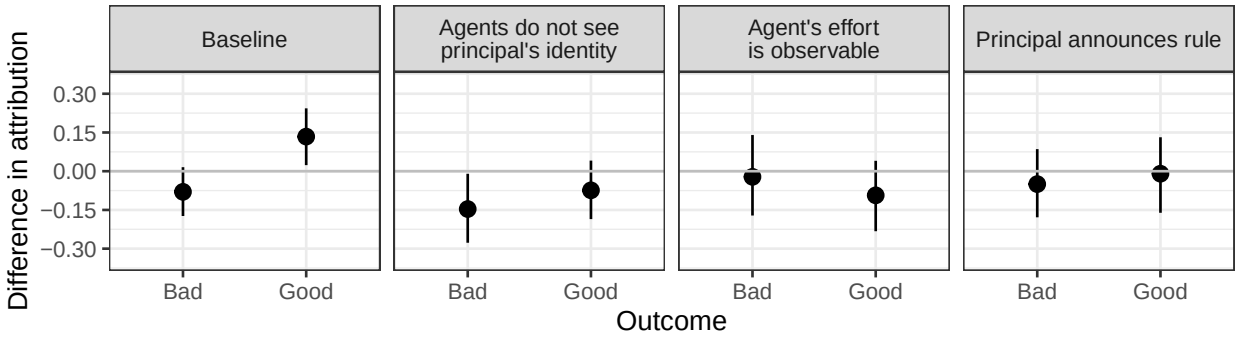


[7] p-values reported here are recovered from a principal-level regression of the difference in reward threshold on treatment indicators). Also note that in the *announce rule* treatment, while there is a significant relationship between the outcome demand principals announce and the reward threshold they are setting, we do not find evidence for a correlation between the level of in-group bias in reward thresholds and differences between in-group and out-group agents in principals pre-announced rule; these relationships are evaluated in a regression of outcome demands on the demand the principal announced before agent's effort, in-group status, and the interaction of the two variables and a regression of the difference in reward thresholds and the difference in announced rule in in-group and out-group matches (with standard errors clustered at the session-level). Regressions are reported in Table S.6 and S.7.

It is also worth noting that principals who do not impose different outcome demands on in-group and out-group agents set a higher reward threshold when agents' effort is observable than when it is not.[8]

With respect to eliciting principals' bias in attribution to effort, recall that in-group bias exists when the difference in attribution to effort between in-group and out-group matches upon observing a good outcome is positive – that, is when the attribution to effort is higher when the principal observes a good outcome generated by in-group than out-group agents. For bad outcomes, in contrast, we would observe in-group bias when the difference in attribution to effort between in-group and out-group matches is negative; in this case, principals' would be less likely to attribute a bad outcome to effort when matched with in-group than out-group agents. Figure 3 shows the differences in attribution to effort for good and bad outcomes generated by in-group vs out-group agents. (Recall that whether an outcome is good or bad is determined by whether it is above or below the principal's reward threshold.)

Figure 3: Difference in principals' attribution to effort between in-group and out-group matches by outcome and treatment.



All three treatments lead to lower in-group bias in attribution to effort by the incentivizing principals who observe a good outcome – that is, the in-group bias in attribution to effort is closer to zero than in the baseline. To the extent that there is a standout, it is the treatment in which agents do not see principals' identity: here the incentivizing principals show significant in-group bias in attribution upon observing a bad outcome; that is, they are less likely to attribute bad

[8]We find no evidence of any principal setting an effort threshold – reward at and above a particular level of effort in the observable effort treatment and not reward below – but not an (lower) outcome threshold. The level of effort the incentivizing principals demand to see correspond with the outcome threshold they set.

outcomes to effort in in-group than out-group matches. The in-group bias in attribution to effort when observing a good outcome decreases from .13 in the baseline to 0.07 (–.04, .19) in the *don't see ID*, to 0.09 (–0.04, 0.23) in the *observable effort*, and to –0.01 (–.13, 0.16) in the *announce rule* treatment. Observing a bad outcome, the difference in principals' attribution between in-group and out-group agents was .08 in the baseline but declines to 0.02 (–.14, –.17) and 0.05 (–.09, 0.18), in the *observable effort* and *announce rule* treatments, respectively. In the *don't see ID* treatment, in-group bias in attribution to effort for bad outcomes increases. That is, the attribution to effort is .15 (.01, .28) lower in in-group than out-group matches. The reduction in in-group bias upon observing a good outcome is significant for the comparison baseline vs. *don't see ID* treatment and baseline vs. *observable effort* treatment ($p < .05$), while the p-value for a test over the change from *baseline* to *announce rule* treatment is $p = .13$; the shift in differences in attribution to effort for bad outcomes when moving from *baseline* treatment to interventions is not significantly different from zero.[9]

Finally, consistent with Hypothesis 2, observing agents' effort significantly increases the correctness of principals' attribution of an outcome to effort. In particular, principals correctly attribute an outcome to higher effort than type at a rate of .33 in the baseline treatment, .37 in the *announce rule* treatment, and .39 in the *don't see ID* treatment. In the *observable effort* treatment, correctness is significantly higher ($p < .01$) at .66 (Recall, observed outcomes are still a function of noise in this treatment as well, explaining attribution mistakes beyond intentional discriminatory attribution).

We summarize the preceding discussion in the following conclusion:

Result 1 *interventions mostly reduce principals' bias in rewards and in their attribution of good outcomes to effort.*

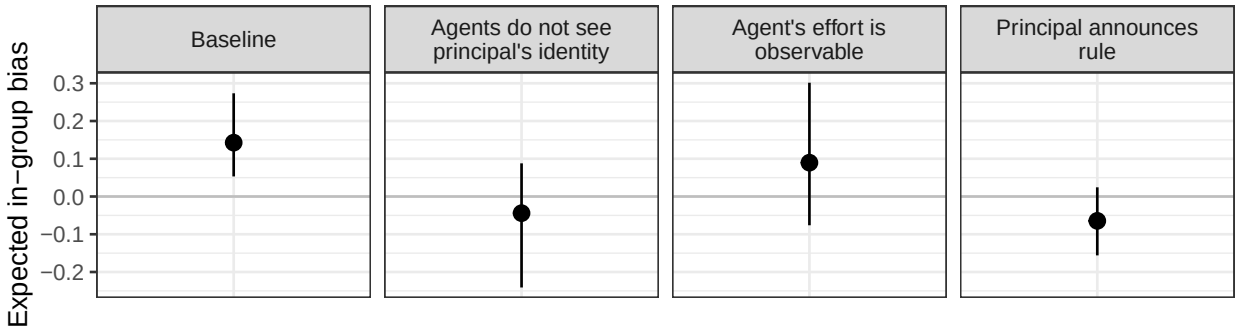
4.2 Interventions – agents

While our intervention treatments prove to be rather promising in reducing discrimination in principals' choices that may be attributable to situational factors, agents' responses are more subtle.

[9] p-values are recovered from a principal-round-level regression of attribution to effort on treatment indicator, in-group status, and the interaction of those two variables; more specifically, the p-value is associated with the hypothesis test of the coefficient on the interaction being zero. Standard errors in this regression are clustered at the principal-level.

Agents expectation of principals' bias disappears in the interventions, as shown in Figure 4; it was .14 in the baseline treatment and reduces to –.04 (–.24, .09) in the *don't see ID* treatment, to .09 (–.08, .30) when effort is observable, and –.06 (–.16, .02) in the *announce rule* treatment. The drop in expected in-group bias from baseline treatment to *announce rule* treatment is significantly different from zero at $\alpha = .05$ while the reductions in in-group bias we see in the *don't see ID* ($p = .11$) and *observable effort* treatments ($p = .72$) are not.[10]

Figure 4: Agents' expectation of in-group bias in principals' reward thresholds by treatment.



To summarize:

Result 2 *Agents' expectation of bias in principals' reward decisions disappears with the interventions.*

5 Discussion

The interventions implemented in this study aim to reduce discrimination in settings where differences in performance and asymmetric attribution thereof may arise from self-reinforcing beliefs about the choices of others and may be independent of the underlying distribution of group attributes. A reduction in discriminatory behavior and beliefs, then, may be thought of as the kind of discrimination that is driven by situational factors and not distinct group statistics or prejudice. Should differences in how groups are treated and believe to act remain, it must be driven by a taste for discrimination among individual actors.

[10] We test for the difference between treatments by a agent-level regression of expected in-group bias on an treatment indicator with standard errors clustered at the session-level. The regression result is tabled in the appendix (Table S.9).

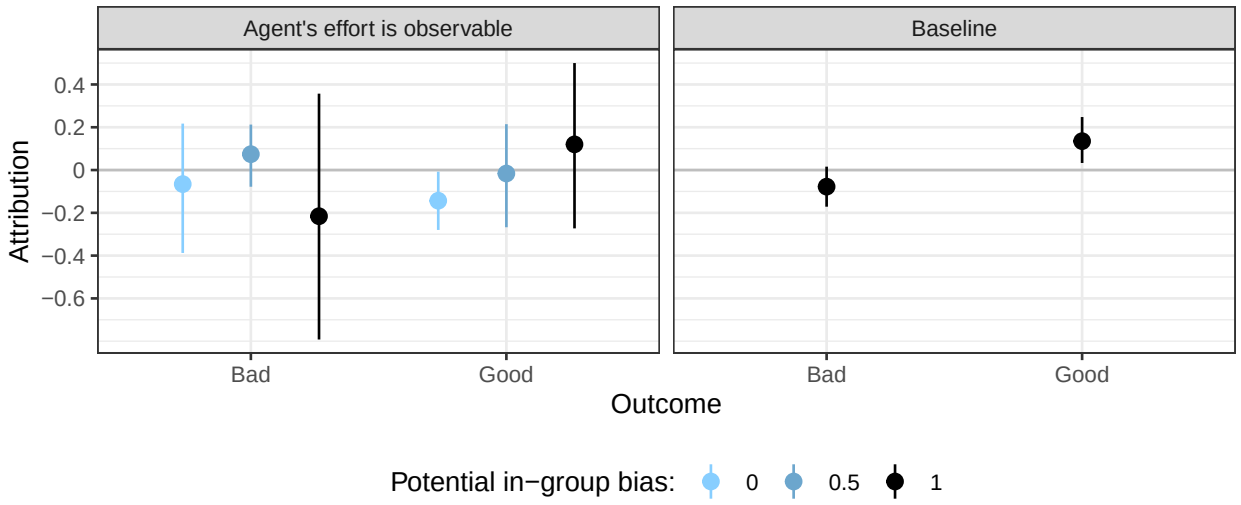
Principals: asymmetric attribution or prejudice? The baseline and observable effort treatments allow us to further parse to what degree group-contingent behavior is driven by the strategic incentives of the principal-agent interaction and to what degree behavior arises from prejudicial in-group favoritism and out-group animosity. Specifically, the observable effort treatment eliminates only parts of the potential for asymmetric attribution. Should in-group bias in attribution to effort remain to exist, it must be driven by prejudice (that is a taste-based explanation for discriminatory beliefs). Surely, principals are much more likely to attribute an observed outcome to effort when effort is observable than when it is not and, as we showed earlier, those principals are more often correctly attributing an outcome to effort than principals in the other treatment conditions.[11] Such attribution behavior obviously follows from receiving better information about agents' effort choices. Further, principals' bias in attribution that existed in the baseline treatment reduces most, when the observed effort, in combination with the observed outcome, is fully informative about the relationship between agent's effort and her assigned type. Certain combinations of observed effort and outcome, though, are not informative. For example, if the principal observes outcome 4 and is also told that the effort exerted by the agent is 2, agent's assigned type may be 1, 2, or 3, leaving the principal in the dark about how to attribute. Figure 5 illustrates that when their is room for making a biased attribution even if effort is observable, principals' attribution choices are similar to the baseline and observable effort treatments (black marker).

The in-group bias in attribution to effort (that is a higher attribution to effort upon observing a good outcome in in-group than out-group matches and a lower attribution to effort for bad outcomes from in-group agents than from out-group agents) diminishes in size when the observed effort is actually informative with respect to the relationship between effort and assigned type in generating the outcome (blue, lighter markers). In this comparison, any deviation from no asymmetry in attribution for those situations where effort is informative, should be read as resulting from a taste for discrimination and not emerging from strategic incentives. We see here that there is no evidence for the taste for discrimination account.

What situations are approximated by our observable effort treatment? Observable effort means

[11] Further, the marginal effect of effort on attribution to effort is .36 (.32, .40) in the observable effort treatment but not distinguishable from zero in the baseline treatment. The estimates of marginal effects are taken from a regression of attribution to effort on effort, in-group status, the interaction of the two variables, and round of play. Standard errors are clustered at the principal-level. Estimated marginal effects are reported in Table S.10.

Figure 5: Difference in principals' attribution to effort between in-group and out-group matches by outcome and potential for bias in attribution for baseline and observable effort treatment.



better information about agents in general. Such better information may arise when the principal, say an employer, evaluates his/her agent, and employee and that agent shares some meaningful social characteristics with the principal. Assessing the agent with better information implies less room for the need to resort to, for example, priors about group statistics; referring to such priors bears the threat of arising statistical discrimination.

Agents: reduced discrimination or increased transactional motivations? As mentioned in hypothesis section above, based on previous work, agents effort choices are heterogeneous and rooted in various rationales. The following analysis and interpretation is therefore rather speculative but still informative. While agents showed an in-group bias in their effort choices in the baseline treatment, exerted higher effort when matched with an in-group than out-group principal, when they expected that principal to behave favorable towards the in-group agent, there is no such in-group bias left in any of the interventions, although the agents response is heterogeneous across treatments. For agents with the expectation of in-group favoritism in principals' reward decisions, effort is by .13 significantly higher for in-group than out-group matches in the baseline but .07 (–.16, .29) lower when the agent does not learn the principals' group identity ($p = .57$), only .09 (–.12, .29) higher when the agent's effort is observable ($p = .41$), and even .23 (–.03, .50) lower in the announce rule treatment ($p = .08$). Also, as with the baseline treatment, agents' effort is indistinguishable in in-

group and out-group matches when they do not expect the principal to be in-group favorable in their reward choices for all intervention treatments; the difference between when the agent is matched with an in-group and not an out-group principal is .08 (–.06, .23), –.01 (–.14, .12), and .02 (–.13, .16) in don't see ID, observable effort, and announce rule treatment, respectively.[12]

This means agents' effort identity-continent choices are very much a function of differences in their beliefs about the behavior of in-group vs out-group principals. It is precisely in the nature of this conditional relationship where agents' behavior in the baseline treatment is favoring the own-identity group. We show that, conditioning on agents' beliefs, interventions reduce agents' in-group bias in effort choices. To enable a more detailed interpretation of the effect of the interventions on agents' effort choices, we further investigate this conditional relationship. Figure 6 gives the levels of agents' predicted effort over the in-group bias in principals' reward decisions and the outcome demand by the principal they anticipate. In the baseline, we see that effort generally increases with both of those expectations. If agents expect the in-group bias in principals' outcome demands to be low, their effort levels favor in-group principals less and sometimes even favor out-group principals.

[12]The estimate of the differences in effort between in-group and out-group matches are taken from a regression of effort on in-group status, agent's assign type, agent's expectation of principals outcome demand, the interaction of the three variables, agent's expectation of principals' bias in outcome demands, the interaction of that expectation with in-group status and expected outcome demand, and round of play (standard errors are clustered at the agent-level. The results of this regression, run separately for each treatment, are shown in Table S.8 in the appendix. The marginal effects reported here are visualised in Figure S.7 in the appendix.

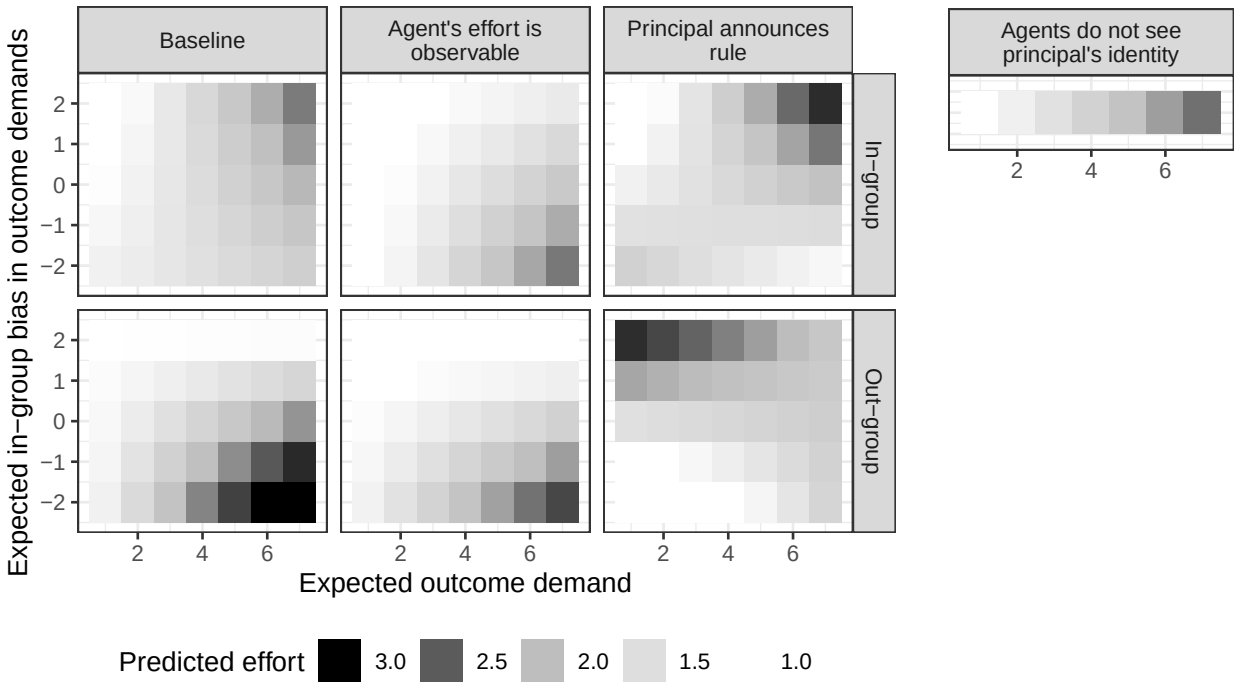


Figure 6: Predicted levels of effort over expected in-group bias in demands and expected demands for in- and out-group matches; estimates are taken from a regression of effort on agent's type, whether they share an identity with the matched principal, expected demands from the principal, expected in-group bias in principal's demands, and the interaction of these variables.

When effort is observable or when principals can announce a non-binding rule, the discrepancies in effort between in-group and out-group matches indicate that effort choices are, to a greater extent, driven by transactional motivations. In the intervention that makes the agents' effort observable, effort in in-group matches decreases with their expectation of in-group bias from the principals, and in out-group matches, the relationship between expected bias and effort is also weaker than in the baseline. In the intervention in which the principal announces a prior rule, the relationship between expected bias and effort in the in-group matches is weaker than in the baseline, and in the out-group matches, lower in-group (and so higher out-group) bias leads to weakly lower effort. In short, in both interventions, there is less evidence that agents' effort is positively aligned with an expectation of identity-based bias in their favor (or lack of such bias against them). In the intervention in which agents do not see principals' identity, agent's effort choices cannot conditioned on that factor, but they, naturally, increase with the expected outcome demand, as they do when the agent's effort is observable. In sum, 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 expected adverse bias due to identity dis-affinity has a negative effect on effort and of the favorable bias due to affinity a positive affect, in the interventions, the effects of expected bias go away.

6 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 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 interventions we evaluate 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.

A key policy implication from our analysis is that successfully reducing discrimination may benefit from focusing on both psychological and situational factors that may contribute to it.

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.

Ashworth, Scott. 2005. "Reputational Dynamics and Political Careers." *Journal of Law, Economics, and Organization* 21(2):441–66.

Ashworth, Scott and Ethan Bueno de Mesquita. 2017. "Unified versus divided political authority." *The Journal of Politics* 79(4):1372–1385.

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.

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

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.

Duell, Dominik and Dimitri Landa. 2021. "Strategic Discrimination in Hierarchies." *The Journal of Politics* .

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.

Gehlbach, Scott. 2006. "Electoral Institutions and the National Provision of Local Public Goods." *Quarterly Journal of Political Science* 2:5–25.

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* .

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

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

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

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.

Loury, Glenn C. 2008. *Race, incarceration, and American values*. MIT Press.

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.

Persson, Torsten and Guido Enrico Tabellini. 2002. *Political economics: explaining economic policy*. MIT press.

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.

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 Statistical appendix

1.1 Session and summary statistics

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
Baseline	Klees	55	1100
	Kandinskys	55	1100
	Total	110	2200
Agents do not see principals' identity	Klees	44	740
	Kandinskys	30	740
	Total	74	1480
Agents' effort is observable	Klees	33	580
	Kandinskys	25	580
	Total	58	1160
Principal announces rule	Klees	31	640
	Kandinskys	33	640
	Total	64	1280
		306	6120

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
Agents do not see principals' identity	.04 (-.07,.16)	.45
Agents' effort is observable	.02 (-.11,.15)	.77
Principal announces 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	Agents do not see principals' identity	Agents' effort is observable	Principal announces 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: Relative frequency of mentions of "effort" or "outcome" in answering the exit survey question about subjects' decision rationale by treatment. .

Mention	Baseline	Agents do not see principals' identity	Agents' effort is observable	Principal announces rule	Difference Baseline vs treatments
Effort	0.47	0.38	0.51	0.31	0.103
Outcome	0.16	0.22	0.29	0.25	0.245

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

Variable	Baseline	Agents do not see principals' identity	Agents' effort is observable	Principal announces 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)	-

1.2 Additional analysis and robustness

Table S.6: Linear least squares regression of reward thresholds on outcome demand announced by principals' before agents choose effort, in-group status, and the interaction of the two variables. Standard errors are clustered at the session-level.

VARIABLES	
announced outcome demand	0.449***
	(0.082)
in-group	2.233*
	(1.310)
announced outcome demand × *in-group*	–0.467*
	(0.238)
constant	2.005***
	(0.523)
Observations	56
R^2	0.035
Adjusted R^2	–0.020

*p<0.1; **p<0.05; ***p<0.01

Table S.7: Linear least squares regression of in-group bias in reward thresholds on difference in outcome demand announced by principals' before agents choose effort out-group and in-group matches (in-group bias in announced rule). Standard errors are clustered at the session-level.

VARIABLES	
difference in announced rule	0.234
to out-group minus in-group agent	(0.510)
constant	0.039
	(0.287)
Observations	28
R^2	0.005
Adjusted R^2	–0.034

*p<0.1; **p<0.05; ***p<0.01

Table S.8: Least square regression of agents' effort on covariates; standard errors are computed based on clustering by agent

VARIABLES	**Baseline**	**Agents do not see principals' identity**	**Agents' effort is observable**	**Principal announces rule**
type	0.168	0.922**	–0.355	–0.249
	(0.321)	(0.369)	(0.465)	(0.532)
in-group	0.310***	0.401***	0.237**	0.108
	(0.077)	(0.073)	(0.102)	(0.099)
expected demand	0.045	0.099	–0.003	0.722**
	(0.140)	(0.099)	(0.255)	(0.288)
expected bias	–0.041	–0.104***	–0.043	–0.021
	(0.037)	(0.031)	(0.041)	(0.046)
type × *in-group*	0.003	–0.233*	0.130	0.002
	(0.095)	(0.120)	(0.144)	(0.137)
type × *expected demand*	–0.186	–0.242	0.032	–1.079***
	(0.209)	(0.233)	(0.381)	(0.343)
in-group × *expected demand*	–0.090***	–0.027	–0.065	–0.094
	(0.033)	(0.046)	(0.053)	(0.063)
in-group × *expected bias*	–0.007	0.122**	–0.028	0.025
	(0.047)	(0.052)	(0.062)	(0.065)
expected demand × *expected bias*	0.134**	0.008	0.024	0.204**
	(0.060)	(0.075)	(0.103)	(0.079)
type × *in-group* × *expected demand*	1.105***	0.930***	1.604***	2.138***
	(0.267)	(0.229)	(0.350)	(0.381)
in-group × *expected demand* × *expected bias*	–0.061	–0.482***	0.066	–0.017
	(0.169)	(0.163)	(0.207)	(0.257)
round	–0.034	0.090	–0.157	–0.177
	(0.134)	(0.097)	(0.138)	(0.181)
constant	–0.003	–0.010**	–0.013**	–0.012**
	(0.004)	(0.004)	(0.006)	(0.006)
Observations	1,029	697	535	600
R^2	0.165	0.210	0.181	0.112
Adjusted R^2	0.155	0.196	0.162	0.093

*p<0.1; **p<0.05; ***p<0.01

Table S.9: Least square regression of agents' expectation of principals' in-group bias in reward thresholds; standard errors are computed based on clustering by session

VARIABLES	
agents do not see principal's identity	–0.184
	(0.101)
agent's effort is observable	–0.052
	(0.135)
principal announces rule	0.213**
	(0.075)
constant	–0.145*
	(0.061)
Observations	152
R^2	0.032
Adjusted R^2	0.012

*p<0.1; **p<0.05; ***p<0.01

Table S.10: Average marginal effect of effort on attribution to effort estimated from a linear least squares regression of attribution to effort on in-group status, the interaction of effort and in-group status, and round of play. Principal-level standard errors are reported.

	AME	SE	z	p
Baseline	-0.01	0.03	-0.56	0.58
Agents do not see principal's identity	0.01	0.03	0.30	0.77
Agent's effort is observable	0.36	0.04	8.83	0.00
Principal announces rule	0.05	0.03	1.90	0.06

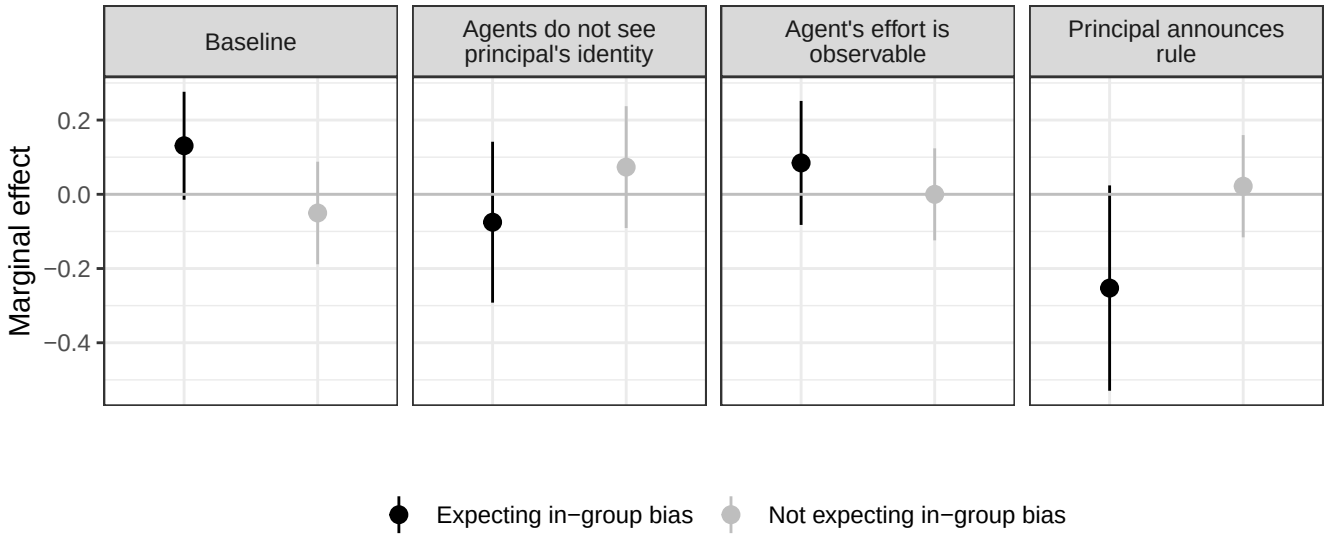


Figure S.7: Marginal effect of in-group status on agents' effort by treatment. Estimates from a regression of effort on type, in-group statues, agents' expected outcome demand, the interaction of those variables, agents' expected bias, the interaction of expected bias with in-group status and agents' expected outcome demand. Standard errors clustered at the agent-level.

2 Experimental design appendix

2.1 Set-up

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.

2.2 Elicitation of principals' evaluation of agents' performance: reward thresholds

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.

2.3 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.5 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.

2.4 Experimental instructions (*baseline* treatment)

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**

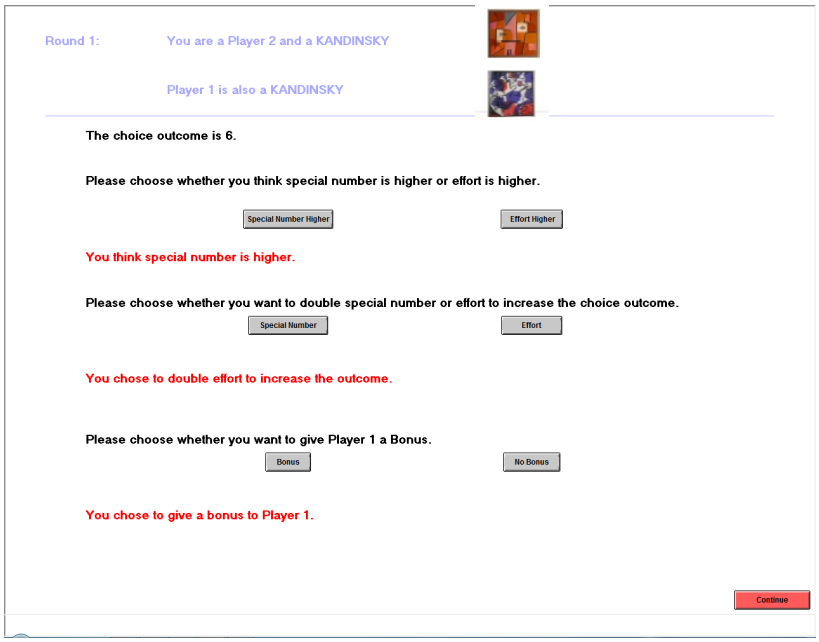


Figure S.8: Principal's decision screen

Figure S.9: Agent's decision screen (Shown in instructions to subjects)