Games and Economic Behavior 68 (2010) 50–59

Contents lists available at ScienceDirect

Games and Economic Behavior

www.elsevier.com/locate/geb

Discrimination in the lab: Does information trump appearance?

Marco Castillo, Ragan Petrie *

George Mason University, Interdisciplinary Center for Economic Science, Department of Economics, 4400 University Dr., MSN 1B2, Fairfax, VA 22030, United States

ARTICLE INFO

Article history:
Received 25 August 2007
Available online 3 May 2009

JEL classification:
C92
J15

ABSTRACT

Using a laboratory experiment, we find evidence consistent with statistical discrimination in a public good and group formation game. In the game, payoff relevant information is presented to subjects, thereby making it costly to discriminate when choosing group members. We find that behavior is correlated with race and people use race to predict behavior. However, race only matters when information on behavior is absent. These results are further confirmed when incentives are in place to encourage behavior that is counter to stereotypes. Not all subjects discriminate in the same way, suggesting unfamiliarity and some in-group, out-group bias. Overall, the evidence points to a lack of information rather than discriminatory preferences.

1. Introduction

Without information on the reputation or behavior of others, people may use physical appearance to form impressions or choose associations. They may withdraw from or never enter into interactions with certain segments of the population because of these impressions. Even in the face of contradicting evidence, initial perceptions may persist. Over time, if enough sorting takes place, some individuals in society could be excluded. Theories of statistical discrimination (Arrow, 1973; Phelps, 1972) point to a lack of information as the reason for this differential treatment, whereas theories of taste-based discrimination (Becker, 1975) point to preferences. A natural question to ask, then, is whether differential treatment in sorting will remain once relevant information is available. While the literature on discrimination is large, it is still unclear to what extent people are willing to pay a cost to discriminate in partner choice. Our research suggests that discrimination experienced in sorting is primarily due to statistical reasons rather than preferences.

We use a repeated linear public goods game to examine this. Public goods experiments are an ideal environment in which to study group formation because payments in the experiment are a function of both individual and group behavior. The more cooperative are the other group members, the more money a person makes. Public goods have broad applications and have been used to understand a variety of topics, from charitable giving (Andreoni, 1990) to the workings of religious groups and communities (Iannaccone, 1992; Benabou, 1996).

Our experimental design combines the public goods game with a sorting task. In this task, we manipulate the amount of payoff relevant information provided to subjects and generate incentives to behave counter to stereotypes. Subjects are shown either the digital photographs of others in the experiment and/or information on past behavior. They are then asked to choose who they would like to have in their group. The sorting task is a "surprise" and provides a cleaner measure on past behavior. Finally, since decisions can be highly correlated with appearance, we run a robustness treatment where we randomly and privately assign different returns to contributing to the public good to induce behavior and verify the results.

* Corresponding author.
E-mail addresses: mcastil8@gmu.edu (M. Castillo), rpetrie@gmu.edu (R. Petrie).

0899-8256/$ – see front matter
doi:10.1016/j.geb.2009.04.015

We find that behavior is correlated with race and people use race to predict behavior whenever other information is not available. When information is provided, however, people disregard personal characteristics and concentrate on behavior. This is also true when behavior is induced. The evidence is consistent with statistical discrimination.

Non-white subjects, including blacks and some other ethnicities, give significantly less and, therefore, make less desirable partners. Accordingly, black subjects are ranked two ranks lower than other potential partners when no information on past behavior is provided. This difference in ranking disappears when information on past behavior is available.[1]

There have been several studies to discern the nature of discrimination, i.e., whether it is taste-based (Becker, 1975) or statistical (Arrow, 1973; Phelps, 1972).[2] Closest to our approach, Fershtman and Gneezy (2001) used economic experiments to show evidence of statistical discrimination in Israel, and List (2004) provided evidence of statistical discrimination in a sport cards market by collecting additional evidence with experiments.

While we also use experiments, our approach does not rely on comparisons across different games nor does it require collecting private information on preferences or expectations. In our experiments, we manipulate the information available to subjects, since theories of statistical discrimination contend that discrimination is a reaction to imperfect information. If true, more information could eliminate any evidence of discrimination or confirm stereotypes, but the influence of appearance should disappear if it is not correlated with behavior. This is the main reason why we break the correlation between appearance and behavior in a robustness test. We consider this approach – manipulating information and behavior to test the nature of discrimination – to be one of the strengths of our design since measuring expectations and social preferences is not a trivial task (see Manski, 2004; Croson, 2000; Levitt and List, 2007).

Another advantage of our approach is that we reduce the possibility of experimental treatment effect biases or lack of comparability between games. It is not clear if the presence or absence of discrimination found in previous research using allocation exercises is the relevant information on preferences in the strategic environment we study. For instance, a person might be willing to pay a premium to live in a homogeneous neighborhood and simultaneously give to charities that target groups different than their own.

Finally, we agree with Plott and Zeiler (2007), who show that trading asymmetries observed in the field might not be due to the endowment effect, that laboratory experiments can allow us to draw strong inference on behavior. For instance, in the absence of information on preferences and beliefs, it is difficult to directly identify whether differential outcomes in market negotiations are due to social preferences or information. Indeed, face-to-face negotiations might be influenced by emotions and competitive attitudes which are separate from preferences or expectations. The advantage of our laboratory protocol is that it provides a *colder* environment in which to make decisions, and this provides stronger identification of discrimination.

The paper is organized as follows. Section 2 describes the experiment. Section 3 shows results. Section 4 concludes.

2. Experiment

2.1. Design

We use a linear public goods game to explore discrimination in group formation. Each subject must decide how to divide a 25 token endowment between a private investment and a public investment. Each token placed in the private investment yields a return of 2 cents to the subject. Each token placed in the public investment yields a return of α_i to the subject and every other member of the group. In three of the four treatments, $\alpha_i = 1$ cent. There are 20 subjects in each experimental session. Subjects are randomly assigned to a five-person group and play 10 rounds with that same group. At the end of each round, subjects learn their payoff, π_i, and the total number of tokens contributed to the public investment by the group, G. Subjects made their decisions privately on a computer. In total, subjects play three 10-round sequences, and each 10-round sequence is with the same group. Subjects know this before they make decisions. At the end of the first 10-round sequence, subjects are again randomly assigned to a new five-person group, and at the end of the second 10-round sequence, subjects are asked to choose their group for the final 10 investment decisions. This is a surprise. Subjects do not know they will

[1] The only evidence we find consistent with taste-based discrimination is that white men have a 10% greater chance of making it to the most-preferred group (top four out of nineteen) in the ranking of black subjects when behavior is induced.

[2] Dickinson and Oaxaca (2006) examine distributional risk as it relates to detecting statistical discrimination. In sports economics, there is evidence of wage discrimination in basketball but not in baseball (Kahn, 1991). Szymanski (2000) shows that English soccer teams take losses by hiring too few black players. Audit studies suggest findings consistent with taste-based discrimination (Riach and Rich, 2002), but there are concerns about treatment effect biases (Heckman, 1998). Ayres and Siegelman (1995) use an audit method to look at race and gender discrimination in new car sales, and Gneezy and List (2004) look at discrimination against the disabled in car repair quotes. Bertrand and Mullainathan (2004) improve upon audit studies by creating fake resumes and find that those with black-sounding names tend to be discriminated against. Knowles et al. (2001) develop a test of taste-based discrimination in police car searches. They find evidence of statistical discrimination but not taste-based discrimination. A more robust test of taste-based discrimination was suggested by Anwar and Fang (2006). They also find evidence of statistical but not taste-based discrimination. Levitt (2004) exploits the changes in incentives in the *Weakest Link* television show to test for alternative theories of discrimination. He does not find evidence of race or gender discrimination but of age discrimination. List (2006) also finds evidence of age discrimination by examining partner choice in the television show *Friend or Foe*. Finally, Charless and Guryan (2007) show that wage gaps and racial preferences in the US vary across states in ways predicted by Becker's (1975) model of discrimination.

Table 1
Experimental treatments.

Average contribution shown when ranking?		Photo shown when ranking?	
		No	Yes
	No		Photo Only ($\alpha_i = 1$ cent)
	Yes	Contribution Only ($\alpha_i = 1$ cent)	Contribution and Photo ($\alpha_i = 1$ cent)
			Two Types ($\alpha_i \in \{0.25\text{cent}, 2.5 \text{ cents}\}$)

be asked to choose their group before this point in the experiment.[3] No personal information is revealed in the first 20 rounds of the experiment, and no information on individual contributions is revealed. We run two 10-round sequences before subjects choose their groups to give subjects experience with playing the game.

In order to create an incentive for people to reveal who they would prefer to be matched with, we create the following game. Subjects rank all the other 19 subjects in the session from most preferred to least preferred. We provide subjects with some information on the other subjects in the room to use for ranking. The information is either the average amount contributed to the public investment during the second 10-round sequence, the subject's photo, or both.[4] Subjects use that information to create a list from most preferred to least preferred. Digital photographs of subjects are taken at the beginning of the experiment, and photographs are head shots, similar to a passport or identification photo.[5]

Once all subjects submit their lists, groups are formed in four steps. First, one person is chosen at random. A group is formed that includes the randomly chosen person and her four best ranked partners. Second, one person from the remaining 15 people who have not been assigned to a group is randomly chosen. A group is formed with that person and her four best ranked partners from the remaining people who have not been previously assigned to a group. Third, one person from the remaining 10 people who have not been previously assigned to a group is randomly chosen. The first four people on that person's list among the remaining people are put in a group with that person. Fourth, anyone not already assigned to a group is put in a group together. Before playing the last 10 rounds, subjects see a screen with the information corresponding to the subjects in their new group and then play the last 10 rounds with that group. In these last 10 rounds, as in the previous 20 rounds, at the end of each round, a subject learns his payoff, π_i, and the total number of tokens contributed to the public investment by the group, G.

This mechanism is similar to the one suggested in Bogomolnaia and Jackson (2002). The mechanism is incentive compatible if preferences over groups are additive in the preferences over its members.[6] It would also be incentive compatible, regardless of preferences over groups, if people are able to rank all possible groups that one could be paired with. Unfortunately, this option would be impractical since the number of groups to be ranked would be exceedingly large.[7]

There are four experimental treatments: Contribution Only, Photo Only, Contribution and Photo, and Two Types. Treatments differ in the α_i assigned to each person and the information that is shown to subjects when they are asked to rank the other subjects. Treatments are summarized in Table 1.

In the Contribution Only, Photo Only and Contribution and Photo treatments, all subjects are assigned an $\alpha_i = 1$ cent. This means that the effective price of contributing to the public good is $p = 2$ cents. In the Contribution Only treatment, when subjects are asked to rank others, they see the average amount contributed to the public good in the second 10-round sequence by all other subjects in the room. In the Photo Only treatment, subjects see the photos of all other subjects. And, in the Contribution and Photo treatment, subjects see the photo and the average amount contributed to the public good in the second 10-round sequence. The average is listed below each subject's respective photo.

In the Two Types treatment, $\alpha_i \in \{0.25 \text{ cent}, 2.5 \text{ cents}\}$. Half of the subjects are randomly assigned a value of 0.25 cent and half are randomly assigned a value of 2.5 cents. Subjects keep the same value for all 30 rounds of play. All subjects know these pieces of information before making decisions. A subject with an $\alpha_i = 0.25$ cent has a price of giving of $p = 8$ cents, making investment in the public good very expensive. A subject with an $\alpha_i = 2.5$ cents has a price of giving of $p = 0.8$ cent and should invest her entire endowment in the public good. We expect subjects assigned the low value to invest little to nothing in the public good. We expect subjects with a high value to invest all of their endowment in the public good. Since types are randomly assigned to subjects, correlation between personal characteristics and contribution

[3] This is important to produce an accurate signal of a subject's true behavior. If a subject had known at the beginning of the experiment they would be asked to do this, their behavior might have been different. Our design eliminates these anticipation effects.

[4] We use average amount invested in the second sequence, rather than the first, because we want a measure on behavior that mitigates the noise that comes with initially learning the game. We acknowledge that behavior is influenced by learning and the behavior of others in the group, however, since group assignment was random, we have no reason to believe that any subject's group experience is any more likely than another subject's.

[5] Using photos breaks confidentiality between the subject and his group. This is necessary to test for gender and racial discrimination. As one referee notes, however, this opens the door for side payments post experiment. This is unlikely to have played a role in subject decisions though because 97% of subjects reported they "never met" the other subjects in the experiment.

[6] Additivity in this context means that if James prefers Jill's company to Jane's company, then James always prefers a group than exchanges Jane by Jill, regardless of who the other members of the group are. Under these conditions, revealing the ordering of others is a weakly dominant strategy for James. If James is not chosen, he is indifferent in the ranking he reveals. If he is chosen, he is better off by revealing his true rankings. Since preferences over others' company is additive, it does not matter whether he is chosen first or last.

[7] In a session of 20 subjects, each subject would need to rank 3,876 possible groups.

levels is expected to be low or nil. If subjects in this treatment are ranked according to gender or race, then this must be taste-based discrimination.[8]

2.2. Implementation

The Contribution Only and Two Types treatments were run twice. The Photo Only and Contribution and Photo treatments were run three times. Each experimental session had 20 subjects. An experimental session lasted one hour and a half. In total, 200 subjects participated in the four treatments. Subjects were recruited from introductory courses in economics and political science.[9] All experiments were run in the computer lab at the Experimental Economics Center (ExCEN) at Georgia State University.[10]

Fifty-four percent of the subjects are women. For race, 44.5% are self-classified as African American or Black, 32.5% are Caucasian or White, 8.0% are Indian, 6.5% are Asian (not Indian), and 8.5% are other categories (this includes Hispanic, Bi-racial, one Arab, and one Pakistani).[11] Because of few observations in groups other than Black and White, we collapse all the non-black, non-white groups into one group called Other. All the main results in the paper hold if the groups combined into the Other category are disaggregated into Indian, Asian and other categories. Average age is 21.0 years (standard deviation 3.8 years). In the Contribution Only, Photo Only, and Contribution and Photo treatments, average payoffs are $21.97 (standard deviation $2.63). In the Two Types treatment, average payoffs are $47.13 (standard deviation $11.12).

3. Results

We would like to know if people discriminate by gender or race when sorting into groups and if that discrimination is statistical or taste-based. To do so, we need to first look at behavior by gender and race and then at how people rank others. Most of the discussion regarding the Two Type treatment is presented in a separate section at the end.

3.1. Average behavior

Fig. 1 shows average contributions to the public good across rounds and treatments for the second sequence. Looking at the three treatments where the price of giving is 2 (Contribution Only, Photo Only, Contribution and Photo), we see that contributions start around 40% and decline over rounds to around 15% in the last round. This compares well with other research in linear public goods games (see Ledyard's, 1995, review). We conclude that our instrument is good. In the Two Types treatment, across rounds, low types contribute about 10% and high types contribute around 80%. High types did not contribute 100% as the theory would predict. We discuss this further in the next section.

Table 2 shows an individual-level random-effects regression of the percent contributed to the public good in sequence 2, controlling for gender, race, gender/race interaction terms, round, individual effects and group effects.[12] For race, we use a dummy variable for Blacks and a dummy variable called Other which includes all non-white, non-black groups. The omitted category is white women. The second column in Table 2 shows the results from the Contribution Only, Photo Only and Contribution and Photo treatments. We combine the three treatments into one regression because this gives us a more robust measure of average behavior for different demographic groups. Column 3 shows results from the Two Types treatment. We look first at column 2.

Blacks and Others contribute 8–12 percentage points less than Whites. There are no gender effects on contributions. Contributions decline 2.4 percentage points per round. These results are robust to alternative specifications, including OLS with clustered errors and random-effects Tobit. Also, if we average contributions across all rounds for individuals in sequence two and compare across race and gender, both rank-sum and t-tests come to the same conclusions.[13]

Looking at the Two Types treatment in Table 2, the percent contributed for those who were assigned a high type is 62.4 percentage points higher than those who were assigned a low type. There are no significant round effects. High types do contribute significantly more than low types across all racial and gender groups, and it is this divergence that is key to our ability to distinguish whether the effect of information is due to information itself or self-confirming biases.

[8] As mentioned in Section 1, given that future performance is measured with error (i.e., by previous contributions), it is possible that personal characteristics still play a role in the ranking decision. This is not the case in our data.

[9] Almost all students take these courses at some point in their undergraduate career (either as a required course or one that satisfies a general education requirement), so the courses are filled with a variety of majors.

[10] Georgia State University is a racially-diverse, urban campus in Atlanta.

[11] We checked both gender and race self-classifications made by the subjects to ensure that there were no obvious misclassifications. There were not.

[12] We use fixed effects for groups and random effects for individuals. We do this to assess the importance of race.

[13] It is not clear why Blacks and Others contribute relatively less (or why Whites contribute relatively more). The results are robust to the inclusion of income-related variables (i.e. education of the subject's father and mother, the number of older and younger siblings and parent's marital status). This difference is not likely due to confusion, as subjects gain experience in the first sequence. The behavior by Blacks and Others in our game is consistent with less trusting behavior found in Trust games (Eckel and Petrie, 2009). One explanation might come from sociology. Bobo and Hutchings (1996) find that Blacks and Latinos are more likely to find other groups as competitive resources, whereas Whites are less likely. The public goods game, by its nature, presents some competition over scarce resources.

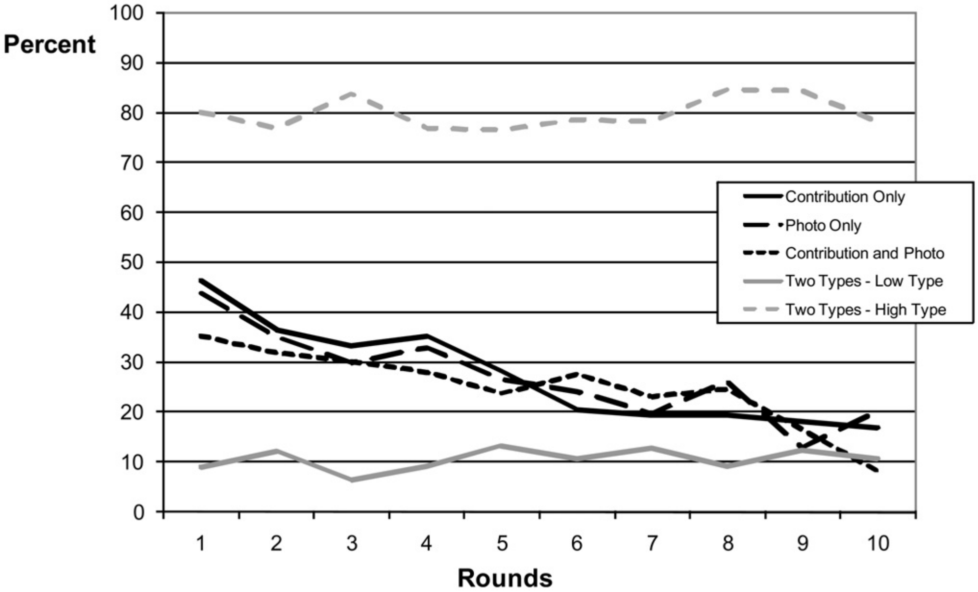

Fig. 1. Average contribution to the public good – second sequence.

Table 2
Dependent variable: Percent contributed in sequence 2 individual-level random-effect regression.

	Combined Treatments (Contribution Only, Photo Only, and Contribution & Photo)	Two Types
Constant	33.34***	−11.33
	(0.000)	(0.439)
Male	−1.09	9.70
	(0.838)	(0.374)
Black	−8.71**	9.72
	(0.042)	(0.493)
Other	−12.76**	7.30
	(0.033)	(0.562)
Black* Male	−1.25	−21.92
	(0.845)	(0.200)
Other* Male	3.81	−3.78
	(0.644)	(0.828)
High Type		62.40***
		(0.000)
Round	−2.43***	0.34
	(0.000)	(0.266)
Group dummies	yes	yes
Individual effects	yes	yes
Within-R2	0.10	0.00
N	1600	400

Note. p-Values in parentheses.
* p-Value < 0.10. ** p-Value < 0.05. *** p-Value < 0.01.

Note that because Blacks contribute less on average, providing information and photos is self-confirming of prejudice. The counterfactual created by the Two Types treatment is an important further test.

Given contributions in Photo Only, Contribution Only, and Contribution and Photo, we would expect Blacks and Others to be ranked lower since they contribute the least. We do not expect women and men to be ranked differently.

3.2. Ranking

In the experiments, we allow people to rank who they would like to have in their group for further rounds of investments. The different treatments allow us to tease apart differences in ranking due to past performance and due to the gender and race of the person being ranked. Each person ranked the other nineteen subjects in order from most preferred

Table 3
Dependent variable: rank (1 = highest, 19 = lowest) fixed-effects regression.

	Contribution Only	Photo Only	Contribution & Photo
Constant	0.85***	8.71***	1.14***
	(0.000)	(0.000)	(0.000)
Expected rank	0.92***		0.90***
	(0.000)		(0.000)
Male		0.57	−0.12
		(0.320)	(0.630)
Black		2.23***	−0.21
		(0.000)	(0.373)
Other		−0.48	−0.41
		(0.517)	(0.244)
Black* male		0.15	0.24
		(0.843)	(0.480)
Other* male		−0.48	0.20
		(0.632)	(0.652)
Individual effects	yes	yes	yes
Within-R2	0.84	0.06	0.80
N	760	1140	1140

Note. p-Values in parentheses.
* p-Value < 0.10. ** p-Value < 0.05. *** p-Value < 0.01.

to least preferred as fellow group members. We use fixed-effects regressions to see how the rank received is affected by demographics as well as past performance.

As mentioned in Section 2, treatments Contribution Only, Contribution and Photos and Two Types revealed subjects average past contributions. However, contributions are not strictly comparable across experimental sessions due to the fact that the distribution of average contributions varied across sessions for any treatment. To make comparisons across sessions meaningful, we use the subject's expected rank for that session's distribution of contributions. So, if a subject had the highest average contribution, her expected rank would be one, and if it was the lowest, it would be nineteen. Ties were assigned the average rank.

Table 3 reports fixed-effects regressions of ranking on expected rank, gender, race, and gender/race interaction terms. The omitted gender/race category is white women. Because ranks went from one to nineteen, with one being the highest rank, a lower rank means that the person was more preferred to be in a group. Regressions include fixed effects on the person doing the ranking since each individual ranked 19 people. All results are robust to alternative specifications and to controls for possible group effects on behavior.[14] This latter control assures us that behavior is independent of the group to which one was randomly assigned in sequence two. We present the fixed-effects regressions because of the ease of interpreting the parameters. Regressions are run separately for each treatment. The Contribution Only treatment allows us to see if past performance alone affects rank. The Photo Only treatment shows if people discriminate based on race and gender, and the Contribution and Photo treatment shows how rank is affected when both performance and physical characteristics are known.

Looking at the results for the Contribution Only and Photo Only treatments, we confirm that, in general, it is difficult to identify the separate effect of personal characteristics on sorting. In Contribution Only, subjects only saw past average contributions when ranking. Not surprisingly, ranking is strongly affected by the subject's expected rank. The relationship is not one to one, but it is very close. A one rank increase in predicted rank increases a person's actual rank by 0.92.[15,16]

In Photo Only, subjects only saw pictures of the other subjects when ranking. They did not know what any other subject contributed on average. In this treatment, black subjects are ranked 2.2 ranks lower, but Others are ranked no lower than Whites. The result on Blacks is robust to alternative specifications.[17] If we compare this to the actual rank a subject would be given based on their actual contribution, Blacks should be ranked 1.9 ranks lower and Others should be ranked 2.3 ranks lower.[18] So, on average, the rank that Blacks are given in Photo Only is 0.3 ranks lower than they should be ranked based

[14] The results are the same if we use random-effects OLS, random-effects Tobit, OLS regressions with standard errors clustered on the individual doing the ranking, rank-ordered logit, and fixed-effect regressions with dummies for the group the person being ranked was in sequence two. In the Two Types treatment, if we regress contributions on the number of high types in the person's group (either 2 or 3) in sequence 2, there is no significant effect on behavior.

[15] A t-test rejects the null hypothesis that the coefficient equals one (p-value = 0.000).

[16] Considering the contribution regression results reported in Table 2, even when no personal characteristics were revealed to subjects in the Contribution Only treatment, ex-post groups would likely be segregated by race. Indeed, this is the case. Using the Contribution Only data and regressing rank on personal characteristics of the person being ranked, Blacks and Others are ranked lower. This is the identification problem.

[17] Similar results hold if we run the regressions and interact race and gender with predicted rank. If we classify people as Black, Asian, Indian or Other, with the omitted category being White, the same results hold. Blacks are still ranked at least two ranks lower.

[18] This comes from a regression of the actual rank of a person's contribution on race.

Table 4
Dependent variable: rank (1 = highest, 19 = lowest) Photo Only treatment fixed-effects regressions.

	Whites	Blacks	Others	Men	Women
Constant	7.17***	9.73***	9.00***	8.33***	9.04***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Male	1.25	0.62	−1.08	1.96**	−0.54
	(0.177)	(0.456)	(0.477)	(0.024)	(0.483)
Black	4.96**	0.69	1.05	2.23**	2.27***
	(0.000)	(0.329)	(0.387)	(0.002)	(0.001)
Other	0.66	−2.44**	2.95	0.10	−0.98
	(0.593)	(0.019)	(0.120)	(0.928)	(0.321)
Black* male	−1.03	0.23	2.53	−0.91	0.94
	(0.389)	(0.829)	(0.176)	(0.407)	(0.348)
Other* male	−0.67	0.27	−2.13	−1.69	0.48
	(0.687)	(0.847)	(0.417)	(0.257)	(0.717)
Individual effects	yes	yes	yes	yes	yes
Within-R2	0.16	0.04	0.05	0.05	0.07
N	380	570	190	532	608

Note. p-Values in parentheses.
* p-Value < 0.10. ** p-Value < 0.05. *** p-Value < 0.01.

on their contributions. That Others are not ranked lower may be a function of a general unfamiliarity with what this group would do or misaligned expectations on their behavior.[19] We will see later that if either of these are the case, expectations are re-aligned with information on performance.

Is the observed differential ranking by race in Photo Only due to taste-based discrimination? We cannot ascertain this from the Photo Only regression results alone. The results from Contribution Only clearly indicate that people want higher contributors in their group. If Blacks and Others contribute less on average, then we would expect them to be ranked lower. Because race is correlated with contributions, we cannot determine if the differential ranking in Photo Only is because people know who are the high and low contributors or because they do not like a particular group. That is, without assuming that subjects have rational expectations, we cannot distinguish why people use the information this way.

The treatment Contribution and Photo permits us to see how personal characteristics affect ranking when information on performance is also provided.

Looking at the results in the Contribution and Photo treatment in Table 3, we see that when both photos and past performance are known, the only significant predictor of how people are ranked is their past performance. An increase by one of predicted rank increases actual rank by 0.90 in Contribution and Photo.[20] Also, in a t-test of the null hypothesis of equal coefficients, the coefficient on Black in the Photo Only treatment (2.23) compared to that in the Contribution and Photo (−0.21) is significantly different (p-value = 0.000). Race only matters when information on behavior is absent. These combined results suggest that the differential ranking by race observed in the Photo Only treatment is due to statistical discrimination or misaligned expectations (especially for Others).

One may wonder if there is agreement across demographic groups on the rankings in Photo Only. Table 4 shows regressions of rank on personal characteristics for different groups of rankers in the Photo Only treatment. Table 4 looks at broad categories of the data to have sufficient observations. Conditioning on the gender or race of the person doing the ranking, we find that both men and women rank Blacks lower. Looking at race, there are differences. Blacks rank Others higher than Whites or Blacks, and Whites rank Blacks 4.96 ranks lower. Blacks and Others do not rank Blacks lower. In a t-test with the null hypothesis of equal coefficients, the coefficient on Black for Whites (4.96) compared to that for Blacks (0.69) is significantly different (p-value = 0.000). It is also significantly different (p-value = 0.000) from that for Others (1.05). That is, without payoff relevant information on behavior, race is important to ranking for Whites, but not for Blacks or Others.

This result is remarkable and shows that information is not equally important (or used in the same way) for everyone. Indeed, for Whites, the characteristics of others explains 16% of the variation in ranking, but for Blacks, it only explains 4%. That is, Whites use the information on personal characteristics more than Blacks.[21] The regression in Table 3 hides this.

On average, the regressions in Table 3 show that the rank placed on Blacks is close to being correct. However, when broken down by race in Table 4, we see that this is a result of Whites overreacting and Blacks underreacting to information on appearance. Why is this the case? It could be due to in-group bias, where black subjects favor other black subjects and white subjects favor other white subjects, unfamiliarity, incorrect beliefs, or some type of stereotyping. The payoff-maximizing strategy, given the results shown in Table 2, is to rank black and other subjects lower. Other subjects is a broadly defined group and there are few subjects per subgroup, so it is difficult to draw definitive conclusions about rankings for

19 Many of the subjects in the Other category are foreign born. People may have been unfamiliar with what to expect from this group.
20 A t-test of the null hypothesis that the coefficient equals one is rejected (p-value = 0.000).
21 Eckel and Petrie (2009) find that Whites are more willing to buy the photo of their partner in a trust game.

Table 5
Two Types treatment dependent variable: rank (1 = highest, 19 = lowest) fixed-effects regression.

	Two Types
Constant	1.46***
	(0.000)
Expected rank	0.86***
	(0.000)
Male	−0.20
	(0.578)
Black	−0.37
	(0.369)
Other	0.04
	(0.928)
Black* male	0.47
	(0.408)
Other* male	0.24
	(0.631)
Individual effects	yes
Within-R2	0.75
N	760

Note. p-Values in parentheses.
* p-Value < 0.10. ** p-Value < 0.05. *** p-Value < 0.01.

Others. Unfamiliarity may be an explanation for some of the subgroups in Others since the majority groups, Whites and Blacks, did not rank them consistent with their behavior.[22]

Finally, as a robustness check for the results in Table 3, we look at non-linearities in rankings. For instance, it might be the case that ranking all 19 other subjects is a cognitively difficult task. Subjects may concentrate on the top four rankings or the bottom four rankings. Looking at the probability of making it to the top four in the rankings, controlling for actually being in the top four, race and gender, all results hold for Contribution Only, Photo Only and Contribution and Photo. All results also hold for making it to the bottom four.

3.3. Induced behavior – two types treatment

It is possible for sorting behavior to hide taste-based discrimination if behavior and personal appearance are highly correlated. In this case, taste-based discriminators could pass as statistical discriminators even if they are not. A basic robustness test is, therefore, to check if information trumps appearance when, by design, incentives are randomly assigned and performance is uncorrelated with personal characteristics. This is the purpose of the Two Types treatment.

Table 5 reproduces Table 3 for the Two Types treatment. We observe that an increase by one of predicted rank increases actual rank by 0.87 in Two Types.[23] We confirm with a t-test of the null hypothesis that the coefficient on Black in the Photo Only treatment (2.23) is equal to that in the Two Types (−0.37) treatment that the two coefficients are different (p-value = 0.000). Race only matters when information on behavior is absent.

In the Two Types treatment, past performance and race are no longer correlated. Each type was randomly assigned to subjects. If there is any differential ranking in this treatment, it must be due to taste-based discrimination. We see in Table 5 that this is not the case. This is further confirmed by looking at the extremes of behavior. Recall that not all high types gave 100% of their endowment, and not all low types were free riders. So, there was some variation in behavior by high types and by low types. Looking at individuals whose average contribution was 25 tokens or 0 tokens, we still find no evidence of differential ranking by gender or race.

The only other robust finding is that white men are 10% more likely to make it to the top four in the rankings of Blacks. This does not hold when we consider the rankings by Whites or Others. Because performance is uncorrelated with race and sex in this treatment, this suggests that there may be some taste-based discrimination at the upper extreme of the rankings by black subjects. There is no evidence of this, however, in the probability of making it to the bottom four. This discrimination is not in evidence in the full ranking regressions in Table 3.

[22] There is some evidence to suggest that Whites and Blacks may be unfamiliar with the behavior of other subjects, particularly Indians. In regressions with Asians and Indians as separate groups Whites rank Indians significantly higher, and Blacks rank them no differently than white women (the omitted category). Compared to behavior, they are incorrect. Indians tend to contribute less and should be ranked lower. Both Blacks and Whites rank Asians correctly.

[23] A t-test of the null hypothesis that the coefficient equals one is rejected (p-value = 0.000).

4. Conclusion

We present a new experimental design that permits us to analyze discrimination in group formation. Our design allows us to distinguish between statistical and taste-based discrimination by manipulating the payoff relevant information made available to subjects and by breaking the correlation between behavior and personal characteristics. Subjects play a repeated linear public goods game and are allowed to rank others as potential group members for the last ten rounds of play. We systematically vary the information available for ranking. Either subjects see the past behavior of others, their photo, or both. A final treatment randomly assigns either a low or high price of giving so that contribution behavior is not correlated with a person's gender or race and the cost of ignoring this payoff relevant information is high. Any differential ranking of others in this treatment must be due to taste-based discrimination.

We find that there is differential ranking of others by personal characteristics, but this discrimination is mainly statistical and clear incentives/signals eliminate it. Black subjects are ranked two ranks lower than any other group when payoff relevant information is unknown. We find some evidence consistent with in-group/out-group bias in this ranking, as only white subjects rank Blacks lower.

Once we provide payoff relevant information, however, the only explanation for how one is ranked is past behavior. This result is confirmed with our final treatment that makes race and gender poor predictors of behavior. Looking at the extremes of the rankings, the only evidence we find consistent with taste-based discrimination is that white men have a higher chance to make it to the top group in rankings by black subjects. This does not hold in the rankings by any other group. This implies that observed discrimination in the absence of information on behavior is mainly statistical. Also, given clear signals on behavior, most discrimination is eliminated.

Our design is robust in distinguishing statistical from taste-based discrimination. While we do not find strong evidence for taste-based discrimination in this study, we do in a non-student population in Peru (Castillo et al., 2008). The subjects in Peru are from a diverse random sample of young adults who are currently working, whereas the sample in this paper, while diverse, are college students. The difference in the degree of taste-based discrimination we find may be due to student status, age, or different cultures. Since nothing similar to the civil rights movement or affirmative action has been implemented in Peru, it is difficult to draw conclusions on the importance of legislation on discrimination.[24] We do not take this as detrimental to our results, but as an argument in favor of collecting experimental data across diverse populations. It is important to notice, however, that in both studies differential treatment by race or gender is greatly reduced when verifiable payoff relevant information is available, even though some lingering discrimination may remain.

Our results suggest that mechanisms that give clear signals on behavior may go far to reduce evidence of differentiation. This is important for policy makers who may be seeking institutions that address discrimination.[25] The best results are obtained when incentives are strong. In that case, there is no real difference in performance and therefore no risk in perpetuating inequality. The challenge is how to design mechanisms in ways that are believable and efficient.

References

Andreoni, J., 1990. Impure altruism and donations to public goods: A theory of warm-glow giving. Econ. J. 100, 464–477.

Anwar, S., Fang, H., 2006. An alternative test of racial prejudice in motor vehicle searches: Theory and evidence. Amer. Econ. Rev. 96, 127–151.

Arrow, K., 1973. The theory of discrimination. In: Ashenfelter, O., Rees, A. (Eds.), Discrimination in Labor Markets. Princeton Univ. Press, Princeton, NJ, pp. 3–33.

Ayres, I., Siegelman, P., 1995. Race and gender discrimination in bargaining for a new car. Amer. Econ. Rev. 85 (3), 304–323.

Becker, G., 1975. The Economics of Discrimination, 2nd ed. Chicago, Univ. Chicago Press.

Benabou, R., 1996. Heterogeneity, stratification, and growth: Macroeconomic implications of community structure and school finance. Amer. Econ. Rev. 86 (3), 584–609.

Bertrand, M., Mullainathan, S., 2004. Are Emily and Greg more employable than Lakisha and Jamal: A field experiment on labor market discrimination. Amer. Econ. Rev. 94, 991–1013.

Bobo, L., Hutchings, V., 1996. Perceptions of racial group competition. Amer. Sociolog. Rev. 61, 951–972.

Bogomolnaia, A., Jackson, M., 2002. The stability of hedonic coalition structures. Games Econom. Behav. 38 (2), 201–230.

Castillo, M., Petrie, R., Torero, M., 2008. Beautiful or White? Discrimination in group formation. Working paper, Georgia State University.

Charless, K., Guryan, J., 2007. Prejudice and the economics of discrimination. NBER working paper 13661.

Croson, R., 2000. Thinking like a game theorist: Factors affecting the frequency of equilibrium play. J. Econ. Behav. Organ. 41, 299–314.

Dickinson, D., Oaxaca, R., 2006. Statistical discrimination in labor markets: An experimental analysis. Working paper, Appalachian State University.

Eckel, C., Petrie, R., 2009. Face value. Working paper, Georgia State University.

Fershtman, C., Gneezy, U., 2001. Discrimination in a segmented society: An experimental approach. Quart. J. Econ. 116 (1), 351–377.

Gneezy, U., List, J. 2004. Are the disabled discriminated against in product markets? Evidence from field experiments. Working paper, University of Chicago.

Heckman, J., 1998. Detecting discrimination. J. Econ. Perspect. 12, 101–116.

Iannaccone, L.R., 1992. Sacrifice and stigma: Reducing free-riding in cults, communes, and other collectives. J. Polit. Economy 100 (2), 271–291.

Kahn, L., 1991. Discrimination in professional sports: A survey of the literature. Ind. Lab. Relat. Rev. 44, 395–418.

Knowles, J., Persico, N., Todd, P., 2001. Racial bias in motor vehicle searches: Theory and evidence. J. Polit. Economy 109, 203–229.

[24] Only recently has some legislation been passed to address discrimination. In 2000, the Peruvian government passed a law (Law No. 27270 against discriminatory acts) partially as a reaction to hidden forms of discrimination, such as requiring "good presence" for clerical positions. For this, we believe that what we observe in Peru is taste-based discrimination.

[25] Equal opportunity legislation might create environments where these clear signals are possible. However, there might be circumstances where measurement of performance is too difficult to provide clear signals.

Ledyard, J.O., 1995. Public goods: A survey of experimental research. In: Kagel, J.H., Roth, A.E. (Eds.), The Handbook of Experimental Economics. Princeton Univ. Press, Princeton, NJ.
Levitt, S., 2004. Testing theories of discrimination: Evidence from weakest link. J. Law Econ. 47, 431–452.
Levitt, S., List, J., 2007. What do laboratory experiments measuring social preferences reveal about the real world? J. Econ. Perspect. 21, 153–174.
List, J., 2004. The nature and the extent of discrimination in the marketplace: Evidence from the field. Quart. J. Econ. 119 (1), 49–89.
List, J.A., 2006. Friend or foe? A natural experiment of the Prisoner's Dilemma. Rev. Econ. Statist. 88 (3), 463–471.
Manski, C., 2004. Measuring expectations. Econometrica 72 (5), 1329–1376.
Phelps, E., 1972. The statistical theory of racism and sexism. Amer. Econ. Rev. 62, 659–661.
Plott, C., Zeiler, K., 2007. Exchange asymmetries incorrectly interpreted as evidence of endowment effect theory and prospect theory? Amer. Econ. Rev. 97, 1449–1466.
Riach, P.A., Rich, J., 2002. Field experiments of discrimination in the market place. Econ. J. 112, 480–518.
Szymanski, S., 2000. A market test for discrimination in the English professional soccer leagues. J. Polit. Economy 108 (3), 590–603.