Discrimination in the Lab: Experiments Exploring the Impact of Performance and Appearance on Sorting and Cooperation

Marco Castillo
Georgia Institute of Technology

Ragan Petrie*
Georgia State University

July 2006
(updated December 2006)

Abstract: We present experimental evidence consistent with statistical discrimination in a public good and group formation game. We find that behavior is correlated with race and, to a lesser extent, gender, and people use race to predict behavior when no other information is available. When information on behavior is provided, people disregard personal characteristics completely. These characteristics are also disregarded when individual behavior is induced to break the correlation between characteristics and behavior. That is, people disregard race and gender even when observed behavior is unusual but relevant to payoffs. Finally, our experiments show that sorting into groups has dramatic implications on cooperation. Overall payoffs are higher when sorting is possible than when groups are randomly assigned. This only occurs, however, when personal characteristics are known. Higher payoffs are attained at the cost of an equitable distribution.

*Petrie thanks the Office of the Advancement of Women and the Office of the Vice President for Research at Georgia State University for financial support. We also thank Krawee Ackaramongkolrotn, John List, Todd Swarthout, and seminar participants at the Southern Economics Association meetings, the International Economics Science Association meetings, Georgia State University, and Georgetown University.

1. Introduction

Without information on the reputation or performance of others, people may turn to physical appearance as the criteria for forming impressions or choosing associations. For example, people may withdraw from or never enter into interactions with certain segments of the population because of these impressions. Initial perceptions might persist even in the face of evidence contradicting them. In the long run, individuals in society may suffer persistent losses due to exclusion if enough sorting takes place.

How important are these types of (mis)perceptions in determining group composition and economic outcomes? This research uses laboratory experiments to examine how and why observable characteristics and salient performance signals are used by people to sort into groups. We conjecture that people may focus on personal characteristics, such as gender or race, to choose group members because they lack better information on future performance. The difficulty of identifying whether personal characteristics have an impact on who is chosen as a group member is that personal characteristics and performance can be correlated (or thought to be). For example, if men are better performers and we observe a preference for men as group members, is this because they are better performers or because people prefer to be in groups with men? Our experiments test this hypothesis by manipulating the information made available to subjects and by inducing behavior to break the correlation between performance and personal characteristics. We do this by making personal characteristics irrelevant or bad predictors of behavior and therefore irrelevant to payoffs. Making appearance a bad predictor of behavior is an important robustness test of statistical discrimination, since no room is left for self-confirming biases.

We use a repeated linear public goods game. Repeated public goods experiments are a natural 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. In our experimental treatments, subjects are shown, in a surprise sorting game, either the digital photographs of others in the experiment or information on past performance (or both). Subjects are asked to choose who they would like to have in their group. One treatment breaks the correlation between performance and appearance, and this allows us to identify discrimination for other than statistical reasons.

Our approach is novel in that it manipulates information or the equilibrium

at the experimental level within the same game to identify sources of discrimination. Not only can we cleanly distinguish if discrimination is statistical or taste-based, but by altering the incentives to create the necessary counterfactuals, we can directly observe the impact of sorting on performance and payoffs. Our experimental environment is strategic and relevant to understanding how groups or neighborhoods form.

Several experimental studies have shown the social context of decisions can affect outcomes (Andreoni and Petrie, 2004; Bohnet and Frey, 1999; Burnham, 2003; Henrich, Boyd, Bowles, Camerer, Fehr and Gintis, 2004). Indeed, specific characteristics of a partner, such as gender, beauty and ethnicity, can affect decisions (Andreoni and Petrie, 2006; Castillo and Carter, 2006; Eckel and Wilson, 2003; Ferraro and Cummings, 2006; Hammermesh and Biddle, 1994; Mobius and Rosenblat, 2006; and Petrie, 2004). There is also experimental evidence showing that strategic behavior is affected by gender (Croson and Gneezy, 2004). Finally, Slonim (2004) shows that there is significant against one's gender discrimination in a trust game where partners are chosen.

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). The literature on sport economics is mixed. In sports, performance is measured more accurately and therefore regressions of wages on race are less susceptible to omitted variable problems. However, there is evidence of wage discrimination in basketball but not in baseball (Kahn, 1991). Audit studies suggest findings consistent with taste-based discrimination (Riach and Rich, 2002), but there are concerns about treatment effect biases (Heckman, 1998). 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. While their study finds evidence of discrimination, it cannot identify its nature since resumes are imperfect measures of performance.

Knowles, Persico and Todd (2001) develop a test of taste-based discrimination in police car searches. They observe that the success rate of police searches must be equalized across drivers' races in a matching-pennies model of police interdiction. 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. However, as the authors acknowledge, their test favors the hypothesis of no 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*.

An alternative way to test for taste-based discrimination is the use of economic experiments. Fershtman and Gneezy (2001) show evidence of statistical discrimination in Israel. They observed that people mistrusted men of Eastern origin but otherwise did not make a difference when given the opportunity to make transfers to them. List (2004) also provides evidence of statistical discrimination in a sport cards market by collecting additional evidence with experiments. He finds that difference in bargaining behavior can be explained by difference in the distribution of reservation valuation and willingness to pay. Similar to Fershtman and Gneezy, he uses allocation exercises to test for taste-based discrimination and finds no evidence of it.

Our approach is similar to List (2006) in that we look at discrimination in sorting and to Fershtman and Gneezy and List (2004) in that we utilize economic experiments to gather information otherwise not readily available in economic transactions. To directly test for alternative theories of discrimination, our approach is different in that we use the same game and manipulate the information available to subjects. Previous approaches indirectly test for statistical discrimination since no payoff-relevant information is provided to subjects. We adopt this approach because theories of statistical discrimination contend that discrimination is a reaction to imperfect information. More information could eliminate any evidence of discrimination or confirm stereotypes. This is why breaking the correlation between appearance and performance is so important. By creating a counterfactual on expected behavior, if discrimination remains, it must be taste-based.

Our approach has some advantages. First, by keeping the game constant we reduce the possibility of experimental treatment effect biases or lack of comparability between strategic games and allocation exercises. Second, even in the absence of experimental treatment effects, it is not clear if the presence or absence of discrimination found in allocation exercises is the relevant information on preferences in the strategic environment we study. For instance, people might be indifferent to the race of those receiving their charity but not indifferent to the composition of their neighborhood. A person might be willing to pay a premium to live in a homogenous neighborhood and simultaneously give to charities that target groups different than their own. For this reason, we directly test for discrimination in the sorting task.

Our experiment is based on a series of repeated public good games with a *surprise* group formation stage before playing the last rounds of the public goods game. In the surprise group formation stage subjects are allowed to rank potential partners using an incentive-based mechanism. It is in this stage where information availability is manipulated. Subjects are shown the average past behavior of potential partners, the photo of potential partners or both. Finally, since behavior could be highly correlated with appearance, we induce behavior to break this correlation and test the effect of information when behavior is unusual. This treatment makes information on appearance a poor predictor of behavior and therefore irrelavant under the null hypothesis of statistical discrimination.[1] Note that taste-based discrimination theories would have a hard time explaining why people disregard information on appearance once the relevant information on performance is available. That is, we argue that regardless of the fact that a subject's beliefs on future behavior might be correct or not or observed by the researcher or not, taste-based discrimination should be immune to information on performance. We consider this approach – manipulating information and performance to test the nature of discrimination – to be one of the strengths of our design since measuring expectations is not a trivial task (see Manski, 2004).

We find evidence consistent with statistical discrimination or stereotyping, but not taste-based discrimination. First, we find that people of different backgrounds do behave differently. In particular, non-white subjects, including blacks and some other ethnicities, give significantly less than other subjects and, therefore, make less desirable partners. Second, we find that subjects use information on appearance to rank potential partners. Black subjects are ranked two ranks lower than other subjects when no information on past behavior is provided. Even though other non-white ethnicities give statistically the same amount as Blacks, they are not ranked lower. Under closer inspection, we find that only white subjects rank black subjects lower than others. On average, black subjects are ranked four ranks lower than other subjects by white subjects. Finally, we find that appearance is irrelevant for ranking people once information on average past behavior is available. This is true even when behavior is induced.

The experiments also show that the ability to sort into groups can have sig-

[1] It is possible that some evidence of discrimination remains if measures of performance are available and performace is made orthogonal to appearance. This can happen if subjects believe that the performance of a group is measured relatively poorly (Phelps, 1972). Our experiments gather enough information to test for this alternative hypothesis. We do not find evidence of this.

nificant effects on the distribution of earnings. We find that sorting increases contributions across all groups but more so for those groups composed of the most preferred subjects. This amounts to an increase in average earnings but also in the disparity of earnings. Finally, our experiments show that sorting into groups is effective in increasing cooperation only when people know the identity of their partners. Sorting based on past behavior only has no discernible effect on either average payoffs or the distribution of payoffs.

2. Experimental 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_s, and the total number of tokens contributed to the public investment by the group, G. In total, subjects play three 10-round sequences, and each 10-round sequence is with the same group. 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 be asked to choose their group before this point in the experiment. No personal information is revealed in the first 20 rounds of the experiment.

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. 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.

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 persons 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. 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.

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.[2] 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.[3]

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.

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

[2] 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.

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

of 2.5 cents. Subjects keep the same value for all 30 rounds of play. 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 and each type has a completely different predicted contribution level, there is no correlation between personal characteristics and contribution levels. If subjects in this treatment are ranked according to gender or race, then this must be taste-based discrimination.[4]

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.[5] All experiments were run in the computer lab at the Experimental Economics Center (EXCEN) at Georgia State University.[6]

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 Hispanics, Mulatos, one Arab, and one Pakistani).[7] 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).

[4] As mentioned in the introduction, 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 implication is testable, however, given the data our experiments collect.

[5] 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 course is filled with a variety of majors.

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

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

3. Instrument Check

To test whether our experimental design yields results similar to previous research on repeated linear public goods games, we look at the average contribution across the three treatments where the price of giving is 2 (Contribution Only, Photo Only, Contribution and Photo). Subjects played two 10-round sequences of the public goods game with two different randomly assigned groups. In the first sequence (when subjects are inexperienced), the average contribution is 32.8% of the endowment over all ten rounds. This is similar to that of Andreoni (1988), 33.2%, and to Croson (1996), 35.7%, both of whom also use inexperienced subjects, maintain subjects in the same group, and have a price of giving of 2. Contributions also show a steady downward trend over the 10 rounds. Average contributions start out at 41.6% in the first round and decline to 23.9% in the tenth round. The last rounds are slightly higher than 11.6% in Andreoni and 18.2% in Croson. Subjects in our experiment knew they were playing three 10-round sequences, so the last round in the first sequence was not the last round of the experiment. Subjects in Andreoni and Croson's experiments thought they would only play 10 rounds in total.

The average contribution and trend behavior from our standard public goods game is similar to previous work. We conclude that our instrument is good and proceed to looking at behavior across treatments.

4. 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. Finally, we look at who does relatively better when people sort into groups.

4.1. Average Behavior

Table 1 shows a 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. 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. As discussed in the previous section, we combine all the non-white, non-black subjects into one group because

of the limited number of observations per group. Our main results are the same if we further disaggregate the Other group into separate categories, including Indians, Asians and the remaining in one group. The second column in Table 1 shows the results from the Contribution Only, Photo Only and Contribution and Photo treatments. Columns 3 shows results from the Two Types treatment. We look first at column 2.

Table 1

Dependent Variable: Percent Contributed in Sequence 2

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 Effects	yes	yes
Individual Effects	yes	yes
within-R2	0.10	0.00
N	1600	400
p-values in parentheses		

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.

Looking at the Two Types treatment in Table 1, 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 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 between statistical and taste-based discrimination. Note that high types did not contribute 100% as the theory would predict. We discuss this further in the next section.

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.

4.2. Ranking

In the experiments, we allow people to rank who they would like to have in their group for further rounds of investments so we can see if past behavior (percent contributed to the public good), gender, and race affect ranking. 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 to least preferred as fellow group members. We use fixed-effects regressions to see how the rank received is affected by the gender and race of the person being ranked as well as the 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. An average contribution of 10 tokens may be the highest average contribution in one session but the median average contribution in another session. 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 2
Dependent Variable: Rank (1=Highest, 19=Lowest)
Fixed-Effects Regression

	Contribution Only	Photo Only	Contribution & Photo	Two Types
constant	0.85	8.71	1.14	1.46
	(0.000)	(0.000)	(0.000)	(0.000)
expected rank	0.92		0.90	0.86
	(0.000)		(0.000)	(0.000)
male		0.57	-0.12	-0.20
		(0.320)	(0.630)	(0.578)
black		2.23	-0.21	-0.37
		(0.000)	(0.373)	(0.369)
other		-0.48	-0.41	0.04
		(0.517)	(0.244)	(0.928)
black*male		0.15	0.24	0.47
		(0.843)	(0.480)	(0.408)
other*male		-0.48	0.20	0.24
		(0.632)	(0.652)	(0.631)
Individual Effects	yes	yes	yes	yes
within-R2	0.84	0.06	0.80	0.75
N	760	1140	1140	760

Note: p-values in parentheses.

Table 2 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.[8] 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

[8] The results are the same if we use 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 in the sequence two. This last regression assures us that the group the subject was randomly assigned to in sequence two had no effect on rankings.

Only treatment shows if people discriminate based on race and gender, and the Contribution and Photo and Two Types treatments show 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 quite one to one, but it is very close. A one rank increase in predicted rank increases a person's actual rank by 0.92.

Considering the contribution regression results reported in Table 1, 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.

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.[9] Recall that both Blacks and Others gave significantly less on average and should be ranked lower if ranking is solely a function of performance. That Others are not ranked lower may be a function of misaligned expectations on their behavior. We will see later that if that is the case, they are re-aligned with information on performance.

Not everyone agrees on the rankings in Photo Only. Table 3 shows regressions of rank on personal characteristics for different groups of rankers in the Photo Only treatment. Table 3 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 do rank Others higher than Whites or Blacks, and Whites rank Blacks 4.9 ranks lower. Blacks and Others do not rank Blacks lower.

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

[9] 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.

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.[10] The regression in Table 2 hides this. This evidence alone cannot distinguish the sources of discrimination. The results are consistent with white subjects discriminating against black subjects, perhaps statistically. But, it is also consistent with an in-group hypothesis that black subjects favor other black subjects, since the payoff-maximizing strategy, given the results shown in Table 1, is to rank black and non-white, non-black subjects lower.

Table 3
Dependent Variable: Rank (1=Highest, 19=Lowest)
Photo Only Treatment
Fixed-Effects Regressions

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

Note: p-values in parentheses.

Is the observed differential ranking by race in Photo Only due to taste-based discrimination? We cannot ascertain this from the 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

[10] Eckel and Petrie (2006) found a similar result, in that Whites were more willing to buy the photo of their partner in a trust game and they were more likely to use the personal characteristics in the photo to differentiate their trust.

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 treatments Contribution and Photo and Two Types permit us to see how personal characteristics affect ranking when information on performance is also provided. In the Two Types treatment, performance is uncorrelated with personal characteristics by design.

Looking at the results in the Contribution and Photo and Two Types treatments in Table 2, 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 and by 0.87 in Two Types. The results from the Contribution and Photo treatment suggest that the differential ranking by race observed in the Photo Only treatment is due to statistical discrimination. The Two Types treatment confirms this.

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 2 that this is not the case.[11] Finally, it is interesting to see that black subjects do not use the information efficiently in the Photos Only treatment. Barring taste-based discrimination, they overestimate the performance of black subjects and non-white, non-black subjects.

4.3. Efficiency

The results thus far give evidence for statistical discrimination in how people rank others. We would also like to know how sorting affects payoffs and efficiency. After subjects submitted their rankings of others, four groups were formed. Group 1 is the first group formed, and Group 4 is the last group formed. Because the group formation mechanism forms groups in order, Group 1 is more likely to be composed of the most preferred people and Group 4 of the least preferred. Indeed,

[11] 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 average rank of people in Group 1 is higher than that of people in Group 2, and so on, for all treatments. Once groups are formed, subjects played another ten rounds with the selected group.

Looking at payoffs first, Table 4 shows the average payoff in each group when groups were randomly assigned and when they were chosen. The first panel shows payoffs in each treatment in sequence 2, when groups were randomly assigned. The last row in the panel shows the difference between the highest and lowest average payoff for the treatment. This difference is between \$1.52-\$1.65 in the first three treatments and \$8.56 in the Two Types treatment. Comparing this to the second panel, which shows payoffs in sequence 3 when groups were chosen, there is a much larger difference when groups are chosen. The difference in highest and lowest payoffs ranges between \$2.18-\$3.70 for the first three treatments and is \$26.23 in the Two Types treatment.

Table 4

Average Payoff per Person for Entire Sequence by Group

Random Groups (Sequence 2)				
	Contribution Only	Photo Only	Contribution & Photo	Two Types
Group 1	7.75	7.11	6.33	10.88
Group 2	6.45	6.35	6.82	16.64
Group 3	7.77	7.94	7.98	13.83
Group 4	6.24	6.49	7.14	19.44
Highest - Lowest	1.52	1.59	1.65	8.56
Chosen Groups (Sequence 3)				
	Contribution Only	Photo Only	Contribution & Photo	Two Types
Group 1	7.66	9.54	9.65	31.14
Group 2	7.73	6.57	8.26	23.20
Group 3	5.94	8.03	7.08	7.82
Group 4	5.55	7.00	5.95	4.92
Highest - Lowest	2.18	2.97	3.70	26.23

These differences show that there is increasing income inequality when people are allowed to sort into groups. The difference between the lowest average individual payoff in a group and the highest increases by 33%-300% when people are

allowed to choose groups. This is significant in all treatments but Contribution Only.[12] The increase in inequality is due to changes on both ends of the distribution. The lowest payoff declines and the highest increases, except in the Contribution Only treatment. Similar results hold if we compare the first sequence, instead of the second sequence, to the third, and there is no significant difference between lowest and highest payoff from sequence one to sequence two. This means that the income inequality comes from sorting, not necessarily learning.

Table 5

Average Money Generated in a Session by Sequence and Treatment

	Contribution Only	Photo Only	Contribution & Photo	Two Types	Total
Random Groups (Seq 2)	141.07	139.47	141.36	303.94	725.83
Chosen Groups (Seq 3)	134.34	155.66	154.73	335.43	780.15
Percentage change	-5%	12%	9%	10%	7%

Does this increase in income inequality from sorting also imply a loss of efficiency? Not necessarily. Table 5 shows the average amount of money generated in an experimental session when groups are randomly assigned and when groups are chosen. Overall, allowing people to sort into groups generates more money across the treatment sessions. The overall amount generated increases by 7%. The biggest gains come in sessions where subjects can see the photograph of their fellow group members. In these sessions, the money generated increases by 9-12%. There is a net decrease in Contribution Only. Thus, while sorting does increase income inequality across subjects in different groups, it does *not* necessarily decrease efficiency in the session. The overall amount of money generated decreases slightly from sequence one to sequence two, but there is a still a net gain in money generated comparing sequence one to sequence three. This further confirms that the efficiency gains are due to sorting and not learning per se.

We have seen an important difference in how people are ranked by others based on race. Does this differential ranking also imply that people make different amounts of money in the last ten rounds? Looking at the difference between payoffs for Whites and Blacks, Blacks do make significantly less money in the

[12] An interquantile regression between the 10th and 90th percentile shows a significantly larger dispersion of payoffs when playing with a chosen group in all treatments but Contribution Only.

Contribution Only and Photo Only treatments, but there is no significant difference in payoffs in the Contribution & Photo and Two Types treatments.[13] While Blacks make less money in the first two treatments, they are also more likely to be in Group 4 in those treatments. Everyone in Group 4 makes less money than those in Group 1.[14]

Conditioning on being in the top group, however, the percentage gain in income from being in a randomly assigned group to being in a chosen group does not differ between Blacks and Whites. The gain in income is 0-5% in Contribution Only, 55-58% in Photo Only, 24-36% in Contribution and Photo, and 102-104% in Two Types. That is, once someone makes it to the most preferred group, payoffs are similar. This means that the difference in payoff from sorting is not necessarily due to personal characteristics but to the group one belongs to.

Where do these efficiency gains come from? Table 6 shows the percentage change in contributions from sequence 2 to sequence 3. In the treatments where photos are shown, there is an increasing differentiation in behavior. Members of Group 1 change their behavior the most, whereas Group 4 changes their behavior the least. There is a significant increase in contribution behavior in Group 1 in Photo Only and Contribution and Photo.[15] There is no significant change in behavior in the Two Types treatment, but in the top group, there was little room for change since most were already contributing close to their full endowment. This shows that it is not sorting per se that changes behavior but seeing who is in one's group.[16]

[13] The p-value for the Wilcoxon rank sum test for difference in payoffs in Contribution Only is 0.0266, in Photo Only is 0.0670, in Contribution and Photo is 0.3530, and in Two Types is 0.7505.

[14] Recall that subjects were not told whether they were in the first group or the last group. We use this classification for expositional purposes only.

[15] A regression on the average change in contributions from being in a randomly-assigned group to a chosen group is regressed on dummy variables for each group for the last 10-round sequence.

[16] Andreoni and Petrie (2004) found a similar result in that individual contributions to the public good increased significantly only when group members could identify other group members.

Table 6
Change in Individual Budget Share Contributions
from Sequence 2 to Sequence 3

	Contribution Only	Photo Only	Contribution & Photo	Two Types
Group 1	-5%	28%	18%	7%
Group 2	0%	-4%	10%	15%
Group 3	-10%	14%	6%	2%
Group 4	-3%	5%	1%	3%

4.4. Preferred Partners and Contribution Behavior

Contribution behavior in the chosen group may also be affected by whether a person was in a group with preferred partners. It might be the case that an individual is more cooperative in a group with people she wanted to be with. We find some evidence to support this hypothesis only in the Contribution and Photo treatment. In all treatments, conditioning on a person's own average rank, the average rank of all group members, and the average contribution in the previous ten rounds, we find that past behavior significantly explains average contributions in the final ten rounds in all treatments.[17] If we include variables on the gender or race composition of the group, these variables do not explain contribution behavior.

In the Contribution and Photo treatment, however, people do also contribute more if they are in a more-preferred group. As the average ranking of all group members increases by one, average contributions increase by 1.14 tokens. This means that people increase their cooperation in groups they want to be in and decrease their contributions when they are with people they do not want to be with.

5. Conclusion

We present a new experimental design that permits us to analyze the nature and consequence of discrimination in group formation and cooperation. Our design allows us to cleanly distinguish between statistical and taste-based discrimination

[17] The correlation between expected rank from average behavior in sequence 2 and expected rank from average behavior in sequence 3 is $\rho = 0.85$. This suggests that past contributions were a good predictor of future behavior.

within the same game by manipulating the information made available to subjects and by breaking the correlation between performance 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 performance of others, the 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. Any differential ranking of others by gender or race 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. Our design is robust in distinguishing statistical from taste-based discrimination. While we find no evidence of taste-based discrimination in this study, we do in a non-student population in Peru (Castillo, Petrie and Torero, 2006). Because payoffs in public goods games are increasing in the contributions of others, a payoff-maximizing individual would do best by choosing cooperative people to be in her group. This is precisely what we find. Past performance is a consistently strong predictor of how someone is ranked by others.

Absent information on past performance, though, do subjects use personal characteristics to rank others? We find that they do. The most consistent result is that black subjects are ranked lower than others when past performance is unknown. We find some evidence consistent with in-group/out-group behavior in this ranking, as only white subjects rank Blacks lower.

This differential ranking cannot be attributed to taste-based discrimination. Both black and non-white, non-black subjects contribute less than any other group, although the latter group is not ranked lower. Because contribution behavior and personal characteristics are correlated (or thought to be), from this result alone we cannot determine if this is due to statistical or taste-based discrimination.

Once we control for performance and personal characteristics, the only explanation for how one is ranked is past performance. This result is strongly confirmed with our last treatment that breaks the correlation between performance and characteristics. This implies that observed discrimination in the absence of information on performance is statistical. Also, given clear signals of performance, any discrimination is eliminated.

There are efficiency gains to sorting but at the expense of income equality. By allowing people to choose their groups, there is a 7% increase in money generated in the experiment. The largest gains happen in the treatments where people can

identify their fellow group members. At the same time that efficiency increases with sorting, there is increasing income inequality between the most-preferred and least-preferred groups. That is, while the experimental economy as a whole benefits by allowing people to sort into groups, there are people who clearly benefit relatively more than others.

We find no evidence that personal characteristics affect payoffs for those in the most-preferred group. Blacks do earn less than Whites when there is sorting, but that difference disappears when subjects know past performance and see the person's picture. Indeed, the gains in earnings for being in the top group are no different between Blacks and Whites.

Our results suggest that people do hold perceptions about others based on their personal characteristics. And, it appears that people use them as a basis to sort into groups. Information on personal characteristics, however, is used differently by different groups. Blacks disregard personal information, even when they should not, and Whites seem to use this information more.

It remains to be learned why black and certain non-white, non-black subjects consistently contribute less. It may be an expression of difference in expectations across subjects and suggests that heterogeneity in behavior and beliefs is important in our sample.

Mechanisms that give clear signals on performance, though, eliminate any evidence of differentiation. This is good news for policy makers who may be seeking institutions that diminish discrimination. The best results are obtained, however, 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.

6. References

Andreoni, J., "Why Free Ride? Strategies and Learning in Public Goods Experiments," *Journal of Public Economics*, 37, 291–304, 1988.

Andreoni, J. and R. Petrie, "Beauty, Gender and Stereotypes: Evidence from Laboratory Experiments," Working Paper, Georgia State University, 2006.

Andreoni, J. and R. Petrie, "Public Goods Experiments Without Confidentiality: A Glimpse Into Fund-Raising," *Journal of Public Economics*, 88(7-8), 1605-1623, 2004.

Andreoni, J., and L. Vesterlund, "Which Is the Fair Sex? Gender Differences in Altruism," *Quarterly Journal of Economics*, 116, 1, pp. 293-312, 2001.

Anwar, S. and H. Fang, "An Alternative Test of Racial Prejudice in Motor Vehicle Searches: Theory and Evidence," *American Economic Review*, 96 127-51, 2006.

Arrow, K. "The Theory of Discrimination," in Orley Ashenfelter and Albert Rees, eds., *Discrimination in Labor Markets*, Princeton University Press, Princeton NJ, 3-33, 1973.

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

Bertrand, M. and S. Mullainathan, "Are Emily and Greg More Employable Than Lakisha and Jamal: A Field Experiment on Labor Market Discrimination," *American Economic Review*, 94, 991-1013, 2004.

Bogomolnaia, A. and M. Jackson, "The Stability of Hedonic Coalition Structures," *Games and Economic Behavior*, 38, 2, pp. 201-30, 2002.

Bohnet, I. and B. Frey, "The Sound of Silence in Prisoner's Dilemma and Dictator Games," *Journal of Economic Behavior and Organization*, 38, 43-57, 1999.

Burnham, T., "Engineering Altruism: a Theoretical and Experimental Investigation of Anonymity and Gift Giving," *Journal of Economic Behavior and Organization*, 50, 133-144, 2003.

Castillo, M. and M. Carter, "Identifying Social Effects with Economic Experiments," Working Paper, Georgia Institute of Technology, 2006.

Castillo, M., R. Petrie and M. Torero, "Ethnic and Social Barriers to Cooperation: Experiments Studying the Extent and Nature of Discrimination in Urban Peru," Working Paper, Georgia State University, 2006.

Croson, R., "Partners and Strangers Revisited," *Economics Letters*, 53, 25-32, 1996.

Croson, R. and U. Gneezy, "Gender Differences in Preferences," Working paper, University of Chicago, 2004.

Eckel, C. and P. Grossman, "Are Women Less Selfish Than Men? Evidence from Dictator Experiments," *Economic Journal*, 108, 448, pp. 726-35, 1998.

Eckel, C. and P. Grossman, "Chivalry and Solidarity in Ultimatum Games," *Economic Inquiry*, v. 39, 2, pp. 171-88, 2001.

Eckel, C. and R. Petrie, "Face Value," Working Paper, Georgia State University, 2006.

Eckel, C. and R. Wilson, "Is There a Mechanism for Detecting Trustworthiness?" Working Paper, Virginia Tech, 2003

Ferraro, P.J. and R.G. Cummings, "Cultural Diversity, Discrimination and Economic Outcomes: an experimental analysis," forthcoming *Economic Inquiry*, 2006.

Fershtman, C. and U. Gneezy, "Discrimination in a Segmented Society: An Experimental Approach," *Quarterly Journal of Economics*, 116, 1, pp. 351-77, 2001.

Gneezy, U., M. Niederle, and A. Rustichini, "Performance in Competitive Environments: Gender Differences," *Quarterly Journal of Economics*, 118, 3, pp. 1049-74, 2003.

Hamermesh, D. and J. Biddle, "Beauty and the Labor Market," *American Economic Review*, 84(5), 1174-94, 1994.

Heckman, J. "Detecting Discrimination," *Journal of Economic Perspectives*, 12, pp. 101-16, 1998.

Joseph Henrich, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr and Herbert Gintis, *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies*, Oxford University Press, 2004.

Kahn, L. "Discrimination in Professional Sports: A Survey of the Literature," *Industrial and Labor Relations Review*, 44, 395-418, 1991.

Knowles, J., Persico, N., and P. Todd. "Racial Bias in Motor Vehicle Searches: Theory and Evidence." *Journal of Political Economy*, 109, 203–29, 2001.

Levitt, S., "Testing Theories of Discrimination: Evidence from Weakest Link," *Journal of Law and Economics*, 47, 431-52, 2004.

List, J. "The Nature and the Extent of Discrimination in the Marketplace: Evidence from the Field," *Quarterly Journal of Economics*, 119, 1, pp. 49-89, 2004.

List, John A., "Friend or Foe? A Natural Experiment of the Prisoner's

Dilemma," *Review of Economics and Statistics*, forthcoming, 2006.

Manski, C., "Measuring Expectations," *Econometrica*, 72(5), 1329-76, 2004.

Mobius, M. and T. Rosenblat, "Why Beauty Matters," Working Paper, Wesleyan University, forthcoming *American Economic Review*, 2006.

Petrie, R., "Trusting Appearances and Reciprocating Looks: Experimental Evidence on Gender and Race Preferences," Working Paper, Georgia State University, 2004.

Phelps, E. "The Statistical Theory of Racism and Sexism," *American Economic Review*, 62, 659-661, 1972.

Riach, P. A., and J. Rich, "Field Experiments of Discrimination in the Market Place," *Economic Journal*, 112, pp. 480-518, 2002.

Slonin, R. "Gender Selection Discrimination: Evidence from a Trust Game," Working Paper, Case Western Reserve University, 2004.