Beautiful or White? Discrimination in Group Formation

Marco Castillo*
Interdisciplinary Center for Economic Science (ICES)
George Mason University
mcastil8@gmu.edu

Ragan Petrie
Interdisciplinary Center for Economic Science (ICES)
George Mason University
rpetrie1@gmu.edu

Maximo Torero
International Food Policy Research Institute (IFPRI)
m.torero@cgiar.org

November 2011

Abstract

We explore the importance of appearance in the endogenous formation of groups using a series of experiments. Participants get to choose who they want in their group, and we manipulate the amount of payoff-relevant information on behavior, thereby making it costly to discriminate based on appearance. We draw participants from a representative sample of a demographically and economically diverse population. This allows broader applicability of our results. We find that beauty predicts desirability as a group member, yet it might mask racial preferences. Payoff-relevant information reduces discrimination a great deal, yet discrimination based on appearance remains. Although their behavior is the same, unattractive participants have a one in ten chance of making it to the most preferred group, whereas attractive participants have a one in three chance. Our results are most consistent with taste-based, rather than statistical, discrimination.

* We thank the Inter-American Development Bank for research funding. Kevin Ackaramongkolrotn programmed the experiments, and Jorge de la Roca, David Solis, and Néstor Valdivia provided research assistance. This paper has benefited greatly by comments from John List, Andrea Moro, seminar participants at University of Pittsburgh, the LACEA Meetings in Mexico City, Mexico and in Bogota, Colombia, the Economic Science Association Meetings, and the Southern Economics Association Meetings. This work was completed while Castillo and Petrie were on leave at University of Pittsburgh. We are grateful to Lise Vesterlund and the Economics Department for their hospitality.

1. Introduction

Many activities are done in groups, from making purchase decisions in a household to working on a team at the office. While some groups may be assigned by a third-party, many are formed by choice. In the latter case, who is included in a group will be affected by a variety of factors, including the nature of the group task and the ability of a potential member to contribute to group production. Sometimes, though, relevant information on this ability is not readily available. Absent that, people may resort to physical cues, such as gender, beauty or race, to make inferences on who would be a contributing member to the group. Perceptions (or misperceptions) of a person's ability to be a good member may result in exclusion of some people based solely on their physical appearance. In the case of extreme sorting, people may never interact with and learn of the abilities and talents of those outside the group.

Theories of statistical discrimination (Arrow, 1972; Phelps, 1972) argue that differential treatment is due to lack of information. Once relevant information is available, ignoring it and discriminating based on physical cues can be costly. This suggests that by providing relevant information on past behavior, discrimination should be reduced or eliminated.[1] In this paper we use a series of artefactual field experiments (Harrison and List 2004) where the necessary counterfactuals to test the *nature* of discrimination are provided and agents likely to engage in discrimination face one another.

To test for the nature of discrimination, we present payoff-relevant information about a potential group member's past decisions. By doing so, we make it costly to discriminate solely by appearance when choosing who will be in the group. In addition, since good signals from minority groups might be noisy (Aigner and Cain, 1997; Cornell and Welch, 1996; Heckman 1998) and therefore discounted, we add a robustness test on the nature of discrimination: we provide incentives for people to behave counter to stereotypes and increase the cost to discriminate.[2] In this case, ignoring bad signals from majority groups cannot be consistent with statistical discrimination.

To get a robust measure of discrimination in the population, we recruit from groups that normally do not interact. Participants are recruited from the working population and are between the ages of 20 and 35. The sample, while small, is similar to the population at large. By

[1] Altonji and Pierret (2001) provide evidence for this type of phenomenon.
[2] The idea is to test for taste-based discrimination, as defined by Becker (1975).

restricting ourselves to the working population we diminish the critique that convenient populations (such as students) might be quite different than the general population (especially in developing countries where college education is uncommon) and might bias the results towards no discrimination.[3]

Our results are intriguing. Even with additional information available, we find that people still choose group members based on their appearance. This is borne out strongly in the data: conditioning on payoff-relevant information in our treatment where the costs to discriminate are high, an unattractive person has a one in ten chance to make it to the most preferred group, while a randomly chosen person should have a one in four chance. Also, discrimination is costly. With discrimination completely eliminated, we calculate that earnings would increase by 13%.[4] Finally, we find that race and beauty are strongly correlated in the minds and preferences of participants. This suggests that a revealed preference for beauty might mask racial discrimination.

We conduct the experiments in Peru because its rich multiracial heritage provides an ideal environment to test for discrimination. Peru's ethnic diversity and history of inter-marriage allow us to check whether people share stereotypes based on race and appearance. Anthropologists argue that race in Latin America is based on phenotypical (appearance) rather than genotypical (ancestry) characteristics.[5] Therefore, the measurement or determination of race is important. Racial discrimination might manifest itself as a beauty premium, which is harder to detect and monitor.[6] Also, in segmented societies, like Peru, interactions across certain groups will likely be limited and hierarchical as populations are sorted into different occupations.[7] Considering such limitations, if perfect sorting into professions is observed, there is little hope in saying much about the extent of discrimination without resorting to strong exclusion restrictions.

[3] Harrison and List (2004) discuss these issues extensively.
[4] Earnings would increase for some and decrease for others. With discrimination completely eliminated, the net gain in earnings in our laboratory economy is zero. On average, the absolute change in earnings is 13%, which means there would be a large change in earnings (both positive and negative).
[5] Goldsmith, Hamilton, and Darity (2005) show evidence against the idea that race in America is a cultural trait by showing that light skinned blacks do not experience wage gaps as brown-skinned and dark blacks do. Gyimah-Brempong and Price (2005) show that skin tone also affect transition into crime and sentence durations. Darity, Dietrich, Hamilton (2005) argue against the idea that race in Latin America is phenotypical. They present evidence of strong preferences for whiteness among people of mixed blood. The fact that racial mixing and cultural adaptation are potential strategies to escape discrimination makes the issue of measuring race more salient.
[6] There may be some truth to this, in that the Peruvian government passed a law in 2000 (Law No. 27270 against discriminatory acts) partially as a reaction to hidden forms of discrimination, such as requiring "good presence" for clerical positions. This requirement was interpreted as code for "not being indigenous."
[7] See Moreno, Nopo, Saavedra and Torero (2007) for a discussion of gender and race occupational sorting into professions in Peru. Blau and Ferber (1992) show that occupational sorting is much more pronounced in Latin America than in other regions.

Artefactual economic experiments (Harrison and List, 2005) allow us to explore these issues more fully.[8]

We use a repeated linear public goods game to explore these issues. Repeated public goods experiments represent a natural environment to study the formation of groups because payments in the experiment are a function of both individual and group behavior and mimic many social situations. Indeed, the tension between cooperation and free riding found in a public goods game is analogous to that in team production (Alchian and Demstz, 1972; Holmstrom, 1982). In a public goods game, the more cooperative are other group members, the more money a person makes. Also, as in any team situation, there is an incentive to free ride. Therefore, when choosing who is in one's group, there is a strong incentive to choose people who are more cooperative. In our experiment, participants choose who they would like to have in their group in a surprise task before the last rounds of play. Treatments determine the type of information made available. Participants are shown either digital photographs of others in the experiment or information on past performance (or both). In our robustness check treatment, because performance and appearance may be correlated and signals might be considered noisy, we create a counterfactual situation where behavior does not perfectly adhere to held beliefs.

The strength of our experimental design is that it allows us to systematically test to what extent personal characteristics, such as gender, beauty and race, and relevant information on the past contributions of a potential group member are used when forming groups or teams. Also, in our robustness check, because the correlation between behavior in the group task and appearance is, by design, weak, we can see if people still prefer, say, a beautiful person on the team who would not contribute much to group production over a not-so-beautiful person who would contribute a lot. If taste-based discrimination is present in preferences, it should remain even when relevant information on past behavior in a group is revealed. We consider this approach -- manipulating information at the experimental level within the same game to test the nature of discrimination -- to be one of the strengths of our design since measuring expectations is not trivial (see Manski, 2004), nor is collecting expectations neutral to the experimental task (Croson, 2000). Finally, separate questions were asked about how much White, Indigenous,

[8] There is little economic research aimed at detecting discrimination in Peru. Nopo, Saavedra and Torero (2007) study wage gaps between white and indigenous workers and find it to be around 12%. Moreno, Nopo, Saavedra and Torero (2004) do not find robust differences in the probability of being hired for job seekers of different ethnic backgrounds in a small audit study in Lima, Peru.

Black, or Asian can be seen in a person's face. This task was performed by participants not involved in the experiment, but recruited from the same population as the experimental participants. This gives us an independent measure of race and beauty.

There have been other studies that examine the nature of discrimination.[9] In the experimental literature, 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 differences in bargaining behavior can be explained by difference in the distribution of reservation valuations 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.[10]

This paper differs in that we test for discrimination by manipulating the incentives directly within the experiment, using the experimental design from Castillo and Petrie (2010). Castillo and Petrie apply the design to a student population in the U.S. and found behavior consistent with statistical discrimination. In this study, we sample from the working-age population in a multi-racial setting, where both types of discrimination are more likely to be present. Indeed, we find evidence of taste-based discrimination based on beauty and race in this richer sample.

Evidence from previous research give reason to believe that people may prefer attractive people because of expectations and productivity. Hammermesh and Biddle (1994) find that attractive people earn more than unattractive people, and beauty can be productivity enhancing (Biddle and Hammermesh, 1998). Mobius and Rosenblat (2006) examine the channels through

[9] The research on discrimination is extensive and rich (for reviews, see Altonji and Blank, 1999, and Cain, 1986), yet to what extent people are willing to incur costs to discriminate in their choice over group members is still an open question. Khan (1991) presents evidence of wage discrimination in basketball but not in baseball in the U.S. Audit studies suggest findings that are consistent with taste-based discrimination (Riach and Rich, 2002), but there are concerns about treatment effect biases (Heckman, 1998). Bertrand and Mullainathan (2004) find that those with black-sounding names tend to be discriminated against in a study using fake resumes. Knowles, Persico and Todd (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. Finally, List (2006) finds evidence of age discrimination in choosing partners in the television show *Friend or Foe*. Dickinson and Oaxaca (2009) examine distributional risk as it relates to detecting statistical discrimination.

[10] In recent papers, Bardsley (2007) and List (2007) show that allocation exercises are fragile to the decision support and may not be a good control for preferences.

which beauty affects wages and finds confidence to be important, as well as expectations that attractive workers will be more productive. Andreoni and Petrie (2008) find that people expect attractive people to be more cooperative in a public goods game. In an on-line market for credit, Ravina (2008) finds that beautiful people are more likely to be awarded a loan, whereas Pope and Sydnor (2011), using data from the same market but for a longer time period, find no significant effect of beauty, but do find that African American loan applicants are less likely to get funded. Collectively, these results suggest that beauty, and race, may interact with selection of group members.[11]

We find that the answer to the issue of discrimination in partner choice is a complicated one. People do use others' personal characteristics to make decisions. However, attractiveness, rather than race, is a much better predictor of unequal treatment. Our estimates of the effects of others' appearance on behavior are large and robust. We also find that race and attractiveness are strongly correlated. The probability of being considered unattractive given that a person is indigenous is 78%, but only 22% for a person classified as white. Although the availability of information on behavior eliminates most discrimination based on appearance, differential treatment remains even when the costs to discrimination are high.

To our knowledge, our work is the first to present evidence consistent with taste-based discrimination in the experimental literature. By changing the information structure within the game and raising the costs to discriminate by exogenously assigning behavioral incentives, our research can more cleanly identify the nature of discrimination in the endogenous formation of groups. Finally, our results confirm that people expect beautiful people to be more productive without information on behavior, and we add to the literature by showing that this preference is complex and can be confounded with racial preferences.

The paper proceeds as follows. The next section describes our sample, section three the experiment and section four the beauty and race classifications. Section five discusses results, and section six quantifies the costs to discrimination. Section seven concludes.

2. The Sample

[11] In a comprehensive review of the literature on beauty, Langlois et al. (2000) find that attractive people are not only judged and treated more favorably but they also behave differently.

The experiments were conducted in urban metropolitan Lima in Peru. We chose this site because of the racial diversity and because we want a broadly representative, non-student sample of the population that is familiar with computers and the internet.[12] By drawing upon this broader population, we are able to look more accurately at the extent of discrimination.

Our sampling strategy is guided by our research questions. First, we want to create a sample of people of various social distances who might not normally interact with one another. Second, at the same time, we want to have a sample of participants which is representative of the young working population in metropolitan Lima. To this end, eligible participants are between 20-35 years of age, live in a variety of neighborhoods in Metropolitan Lima, have labor market experience, are currently working, know how to use the Internet, and have an e-mail account. In addition we sought to keep a gender and income balance. To get our sample, we worked with two companies specialized in surveys and recruiting to help us secure a diverse population in the experiments.[13] Also, we sampled from clusters of owners of small, medium and micro-enterprises.[14]

The protocol used for the experiments was simple enough to include large segments of the population. The interface was graphical and required simply that the participants know how to use a computer mouse. It is important to note, however, that because our experiments rely on internet protocols and the knowledge of using a computer, we likely excluded some segments of the population that might suffer more marked patterns of discrimination. Therefore our results may be viewed as giving a lower bound estimate to the extent of discrimination.

According to the population census of 1993, our sample essentially covers most of the districts in Metropolitan Lima and is highly correlated with the distribution of the population with complete or incomplete higher education.[15] To investigate the comparability of our sample to the population in other dimensions, we compare our experimental participants to a sub-sample from the *Encuesta Nacional de Hogares* (ENAHO) 2004. The sub-sample complies with the

[12] Lima is replete with internet cafes, and there is a high proportion of the non-student population with expertise using computers and the internet.

[13] In general, this mechanism ensures that the opportunity to participate in the experiment is distributed equally across the population. From these databases we sampled all the potential participants that comply with all of our criteria. From the resulting sub-sample we performed a random lottery and selected the individuals to be part of the experiment.

[14] We also recruited from Gamarra (an industrial area in metropolitan Lima). We drew upon a pre-census of all the establishments in Gamarra and this allowed us to randomly select buildings from which to invite participants. Also, this area is one of the largest small- to medium-sized enterprise clusters in metropolitan Lima and represents a rich mix of population in terms of place of origin and socio-economic background.

[15] This includes the following categories: incomplete non-university tertiary, complete non-university tertiary, incomplete university tertiary, and complete university tertiary.

eligibility criteria for all of our participants. The advantage of using the ENAHO as a comparison group is that it is representative of Metropolitan Lima and therefore could help us identify any selection bias in our sample. Our experimental participants and the ENAHO comparison group have a similar distribution among almost all the variables (i.e. age, gender, monthly income, average education, and language distribution), but our experimental participants are slightly more educated. This is most likely a reflection of the requirement in our experiment that participants know how to use the internet. This comparison gives us confidence that the participants in our experiment are a good representation of the larger population in metropolitan Lima.

3. Experimental Design

We use a linear public goods game to explore discrimination in the formation of teams or groups.[16] In the game, there is a tension between self-interest and what is best for the group. No matter what the participant decides to do, he always does better if his fellow group members are cooperative. We exploit this element of the game to examine how appearance and past cooperation affects a participant's choice of members he would like in his group.

3.1 The Basic Game

Each participant is given a 25 token endowment and must decide how to divide the endowment between a private investment and a public investment. Each token placed in the private investment yields a return of 4 centimos to the participant.[17] Each token placed in the public investment yields a return of α_i to the participant and every other member of the group. The return to the public investment, α_i, is 2 centimos in three of the four treatments. There are 20 participants in each experimental session. Participants are randomly assigned to a five-person group and play 10 rounds with that same group. At the end of each round, participants learn their payoff, π_i, and the total number of tokens contributed to the public investment by the group, G. Participants make decisions privately on a computer and do not talk to one another. They do not interact with other participants in any way other than through decisions on the computer.

[16] This design was first developed and used by Castillo and Petrie (2010) with a student population in Atlanta. The design is described in detail here for the convenience of the reader and because the design is crucial to our ability to identify the nature of discrimination in the formation of groups.
[17] There are 100 centimos in 1 sol (the Peruvian currency). At the time of the study, US$1 = 3.2 soles.

In total, participants play three 10-round sequences, and each 10-round sequence is with the same group. At the end of the first 10-round sequence, participants are again randomly assigned to a new five-person group, and at the end of the second 10-round sequence, participants are asked to choose their group for the final 10 investment decisions. Participants do not know they will be asked to choose their group before this point in the experiment. This is a surprise. This design element is important to avoid biasing participant behavior. No personal or individual contribution information is revealed in the first 20 rounds of the game. We run two 10-round sequences before participants choose their groups to give participants experience with playing the game.

3.2 Group Formation

In order to create an incentive for people to reveal who they would prefer to be in their group, we create the following procedure. Participants rank all the other 19 participants in the session from 1 (most preferred) to 19 (least preferred). We provide participants with some information on the other participants 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 participant's photo, or both. Participants use that information to create a list from most preferred to least preferred. Digital photographs of participants are taken at the beginning of the experiment, and photographs are head shots, similar to a passport or identification photo.

Once all participants 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 the top four people on his list. 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 the first four people on that person’s list 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.

Once groups are formed by the procedure described above, participants then see a screen with the information corresponding to the participants in their new group. Participants click a button to acknowledge they have seen this information and then play the last 10 rounds with that

group. During these last 10 rounds, at the end of each round, they see the same information they saw during the previous 20 rounds: their payoff, π_i, and the total number of tokens contributed to the public investment by the group, G. No other information is revealed either when making decisions or at the end of each round.

This sorting 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. Additivity in this context means that if Pablo prefers Maria's company to Gabriela's company, then Pablo always prefers a group than exchanges Gabriela for Maria, regardless of who the other members of the group are. Under these conditions, revealing the ordering of others is a weakly dominant strategy for Pablo. If Pablo is not chosen, he is indifferent in the ranking he reveals, but 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.

Some may argue that additivity of preferences over others' company may be a strong assumption. Some combinations of people might be less successful than others. For instance, women might be very cooperative with other women but not so with men. Therefore, a woman might be chosen to be part of a group when other women are available, but not when mostly men are available.

There is another mechanism that is incentive compatible, regardless of preferences over groups. If people are able to rank all possible groups that one could be paired with, we would not need to be concerned with the additivity assumption. Unfortunately, this option would be impractical since the number of groups to be ranked would be exceedingly large.[18] For this reason, we opted for the mechanism described above because it is intuitive, easy to explain to participants and can be implemented quickly once participants have submitted their lists of rankings.

3.3 Description of the Treatments: Contribution Only, Photo Only, Contribution and Photo

There are four experimental treatments in total: Contribution Only, Photo Only, Contribution and Photo, and Two Types. We discuss the first three in this section and the last in

[18] With 20 participants, each participant would need to rank 3,876 groups.

the next section. Treatments differ in the α_i assigned to each person and the information that is shown to participants when they are asked to rank the other participants.

In the Contribution Only, Photo Only and Contribution and Photo treatments, all participants are assigned an $\alpha_i = 2$ centimos, so the price of contributing to the public good is 2.[19] In other words, the return of contributing two tokens to the public good is equivalent to contributing one token to the private good. In these treatments, payoffs are as follows, $\pi_i = 4*(25-g_i)+2*\sum_{j=1}^{5} g_j$, where g is the amount contributed to the public good. It is in the group's interest for everyone to contribute their full endowment to the public investment, but each individual in the group has a selfish incentive (maximizes payoffs) to put all his tokens in the individual investment.

In the Contribution Only treatment, when participants are asked to rank others, they see the average amount contributed to the public good in the second 10-round sequence by all other participants in the room. Because groups are randomly assigned in the first and second sequences, all participants have an equal probability of being assigned to any given group. Therefore, while contributions in a public goods game are a function of preferences, learning and group behavior, no participant is any more likely to be in a "good" or "bad" group. Average contribution behavior in the second sequence should reflect average performance in a public goods game and minimize the effects of learning.

In the Photo Only treatment, when participants are asked to rank others, they see the photos of all other participants. And, in the Contribution and Photo treatment, participants see the photo and the average amount contributed to the public good in the second 10-round sequence. The average is listed below each participant's photo.

3.4 Description of the Treatments: Two Types

In the Two Types treatment, as in the Contribution and Photo treatment, when participants are asked to rank others, participants see the photo and average contribution to the public good in the second 10-round sequence. In the Two Types treatment, however, $\alpha_i \in \{0.5, 5.0\}$ centimos. Half of the participants are randomly assigned a value of 0.5 and half are randomly assigned a value of 5.0. Payoffs are as follows,

[19] This can also be thought of as a cost. One token contributed to the public investment cuts the return to the participant by one half.

$\pi_i = 4*(25-g_i)+\alpha_i g_i + 0.5*\sum_{j\neq i, j\neq k}^{n_j} g_j + 5.0*\sum_{k\neq i,k\neq j}^{n_k} g_k$, where g is the amount contributed to the public good, n_j is the number of participants in the group with a low return to the public good, n_k is the number of participants in the group with a high return to the public good, and $n_j + n_k = 5$. This means that a participant earns a private return to what he contributed to the private good, plus a return from the total amount contributed by low-return participants in the group, and plus a return from the total amount contributed by the high-return participants in the group. We designed payoffs this way to give everyone an incentive to choose high-return participants to be in the group. With this payoff structure, the game no longer has the tension inherent in typical public goods games (an individual incentive to free ride, but the social optimum is for all to contribute). Participants with a high return to the public good (5.0) have a private incentive to contribute all of their endowment to the public good, and for those with a low return to the public good (0.5), being selfish is the dominant incentive.

Participants keep the same value for all 30 rounds of play. All participants know this information before making decisions. A participant with an $\alpha_i = 5.0$ has a very low cost of contributing to the public good. If he is selfish or altruistic, he should invest his entire endowment in the public good. If he is spiteful or is inequality averse, however, he might not contribute his full endowment, despite the low cost.[20] For a participant with an $\alpha_i = 0.5$, the cost of contributing to the public good is very expensive. We would expect participants assigned the low α_i to invest little to nothing in the public good. In all cases, we expect there to be a clear separation in the contribution behavior between those assigned a low and a high price of giving. Because participants are randomly assigned incentives, however, performance and appearance are not perfectly correlated. The Two Types treatment raises the cost to discriminate in group member selection and, therefore, provides a strong test for taste-based discrimination.

3.5 Experiment Implementation

Each treatment was run twice, and each experimental session had 20 participants. An experimental session lasted approximately two hours. In total, 160 participants participated in the four treatments. Each session ended with an extensive post-experiment questionnaire. The

[20] Palfrey and Prisbey (1997) show evidence consistent with participants not contributing their full endowment, even when it is payoff dominant to do so.

experiments were conducted on computers in two computer labs at the Pacific University in Lima, Peru. Two treatments were run at the same time, so participants were randomly assigned to treatments. Since most participants worked full time, the experiments were conducted on weekend afternoons.

In the Contribution Only, Photo Only, and Contribution and Photo treatments, average payoffs are $19.65 (standard deviation $1.36). In the Two Types treatment, average payoffs are $33.75 (standard deviation $6.87).[21]

4. Race and Beauty Classifications

4.1 Race and Beauty Rating Experiment

We are interested in knowing if people sort into teams or groups based on physical characteristics. While a person's sex is easy to determine, a person's race or beauty is not. We want an independent measure of race and beauty that reflects the general perception of a person. Therefore, we use raters, people who did not participate in the public goods experiment but who are drawn from the same cohort as the participants in the experiment, to rate the photos of the participants in terms of race and beauty.[22] A rater only rated the photo in terms of one characteristic, race or beauty, not both.[23]

For race ratings, because the most popular self-classification of race in Peru is *mestizo* (mixed race), it is important for us to have a measure of race that can adequately capture this mixing. For this reason, we use the race classification method developed by Torero et al. (2004) and Nopo et al. (2007). Instead of asking raters to rate a participant along one dimension only, e.g. "white" or "mestizo," raters evaluate participants along their racial intensity in four categories: white, indigenous, black and asian. These are groups that people readily recognize as distinct racial groups. This gives a more nuanced measure of race and more accurately captures the racial mixing in Peru.

[21] The minimum wage in Peru is about $1/hour.
[22] This technique has been used in other experimental research on beauty, including Andreoni and Petrie (2008), Eckel and Wilson (2006), and Mobius and Rosenblat (2006). Hamermesh and Biddle (1994) had interviewers rate the interviewees in terms of beauty.
[23] Half of the raters were men, with an average age of 27.4 years. For education, 21.7% had incomplete non-university tertiary, 16.7% had complete non-university tertiary, 25.0% had incomplete university, and 21.7% had complete university education.

Twenty people (10 women and 10 men), not involved in the group formation experiment, rated each participant along each of the four race dimensions. Each dimension was rated from zero to ten, with zero being complete absence of the dimension and 10 being the most intense. The four numbers did not need to add up to 10. The raters were also told that if they thought that a person belonged to only one racial group, then that person should be given a 10 for that racial dimension and a zero for all other dimensions. Raters were shown the photos one by one on a computer screen and chose the intensity of each dimension by clicking on a button. The order in which the photos were presented to raters was random. Raters could easily move back and forth between the photos to check or change their answers. Ratings took about one hour, and each rater was paid $9.38 (30 soles) for their time.

For the beauty rating, we followed the same procedure as with the race ratings. The only difference is that the ten men and ten women were asked to rate the physical attractiveness of each person in the photo on a scale of one to nine, with one being very unattractive and nine being very attractive.

4.2 Race and Beauty Standardized Measures

Since we will be using these race and beauty measures to see how they affect group member selection, we need to be sure there is a high degree of agreement among raters in terms of attractiveness and race. For beauty, pairwise correlations among raters ranged from 0.13-0.75, with an average of 0.50.[24] For race, along the white dimension, pairwise correlations among raters ranged from 0.31-0.76, with an average of 0.57. For the indigenous dimension, correlations ranged from 0.02-0.64, with an average of 0.41. For the black dimension, correlations ranged from 0.19-0.82, with an average of 0.50, and for the asian dimension, correlations ranged from -0.02-0.81, with an average of 0.37.[25],[26]

While there are some participants that display intensities in the dimensions of black and asian, the majority of participants display the greatest intensities in the dimensions of white and

[24] The Cronbach alpha for interrater reliability is 0.94.
[25] The Cronbach alpha for interrater reliability 0.96 for the white dimension, 0.93 for the indigenous dimension, 0.94 for the black dimension, 0.91 for the asian dimension.
[26] Note that we also had "trained" raters, in addition to our cohort raters, rate the photos in terms of racial intensity. These raters were trained to minimize variance in racial perceptions. There was a large amount of agreement between the trained raters and cohort raters. For example, along the indigenous dimension, pairwise correlations ranged from 0.20-0.79, with an average of 0.55, and along the white dimension, pairwise correlations ranged from 0.27-0.78, with an average of 0.57. This indicates to us that race can be measured and defined. It is clearly observable, and people can clearly use it to discriminate. This gives us confidence that the race variables we are using are actually picking up the effects of race and not something else.

indigenous. This is in line with the general population in Peru, where blacks make up 2% of the population and Asians make up 3% of the population. Average intensity is 2.83 for white, 3.91 for indigenous, 1.89 for black, and 1.31 for asian. Because the majority of our participants were primarily a mix of white and indigenous, we concentrate on these two dimensions in our analysis. None of our analysis changes if we add asian and black intensities.

While the rating scale for race ranged from zero to ten and for beauty from one to nine, some raters did not use the full range of the scale. For example, for race, some used intensities up to 10 and some only up to 6. To be able to make comparisons across raters, we standardize each rater's rating by her own mean and standard deviation. This permits us to take an average across all twenty raters' standardized ratings for race and for beauty to get the final average ratings we use to analyze the data.

Also, in order to pick up nonlinear effects of race and beauty in our analysis, we create dummy variables. A person is classified as White if the average standardized rating for that person in the white racial dimension falls in the top tercile of the distribution *and* the rating in the indigenous dimension falls in the bottom tercile of the distribution. A person is classified as Indigenous if he falls in the upper tercile of the indigenous distribution *and* in the lower tercile of the white distribution. Given this definition, 25.6% of participants are classified as White and 22.5% are classified as Indigenous. In the results section, we present results using the dummy variables, however, all results are qualitatively similar using the continuous measure of race and beauty. [27]

For beauty, women are rated as more attractive than men. The average standardized attractiveness measure for women is 0.35 and -0.22 for men. Therefore, we classify participants as attractive or unattractive, conditional on their sex. So, a man is classified as attractive if his average standardized attractiveness rating falls in the upper tercile of the distribution of attractive ratings for men. And, a man is classified as unattractive if his rating falls in the lower tercile of the distribution of ratings for men. The same procedure is used for a woman, conditional on how her rating falls in the distribution of ratings for women.

5. Basic Experimental Results

[27] We also tried a dummy variable with an upper and lower quartile cutoff. Results are qualitatively similar to what is reported in the paper, but the dummy variable with the tercile cutoff explains more of the variance in the regression analysis. The tercile cutoff dummy variable also explains more variance than the continuous measures of race and beauty.

5.1 What Did People Contribute in the Experiment?

As is commonly observed in public goods experiments (see Ledyard, 1995), contributions tend to decline over time. In the second sequence, contributions in all treatments, except Two Types, start around 30% and decline to around 15% in the last round. A similar pattern is also observed in the first sequence of the experiment.

The incentives of the Two Types treatment successfully induce a separation in behavior between high and low types. High types contribute about two and a half times as much to the public good as low types. They contribute about 79% over all rounds, and low types contribute about 30%.[28] We do not observe a round effect, as the incentives to contribute little or a lot are strong and constant across rounds. Clearly, not all participants are selfish, as we see low types contributing non-zero amounts and high types contributing less than their full endowment. There appears to be some altruism or inequality aversion among both low and high types. Nonetheless, there is a split in behavior, and because types were randomly assigned, the correlation between behavior and appearance is very low. This behavior is necessary for us to test for the nature of discrimination.

A basic premise in theories of statistical discrimination is that, in the absence of better information, ethnic or cultural background can be used as a proxy for behavior. For instance, migrants might experience rough market conditions, making them behave (or thought to behave) more selfishly. Or, more affluent participants can afford to be more altruistic or take more risks. Table 2 shows a series of OLS regressions aimed at determining if different people do behave differently. All regressions include group-level fixed effects to control for the fact that different levels of contributions might be observed due to interactions within a particular group. The regressions also include clustered errors at the individual level.[29]

The specification in Table 2 tests whether personal characteristics and experimental treatment variables affect contribution behavior. We regress contributions to the public good in sequence 2 on the following independent variables: sex, age, education, race, beauty, round

[28] Contributions by low types in the Two Types treatment and by those in the remaining treatments are similar. This does not appear to be due to confusion, instead of incentives. In Two Types, the correlation between the rank of average contribution in sequence 2 and in sequence 3 is 0.78. And, an OLS regression of contribution in sequence 3 as a function of average contribution in sequence 2, controlling for personal characteristics & assigned type, yields a significant coefficient of 0.71.

[29] The same results hold if run as a random-effects Tobit regression with group-level fixed effects. We report the OLS results for ease of interpretation of the coefficients.

number and assigned type in the Two Types treatment. At this stage in the experiment, groups were randomly assigned, so all independent variables are exogenous. The variables for race and beauty are the dummy variables constructed from the average standardized continuous measures as described in Section 3.4. Model 1 pools the first three treatments together for a more complete picture of behavior. We can do this because, in sequence 2, all participants made decisions under exactly the same conditions in these three treatments. And, this gives us a larger number of observations upon which to draw conclusions about behavior. Model 2 shows behavior only in the Two Types treatment, and Model 3 pools all four treatments together and controls for assigned type.

The regressions in Table 2 show that behavior is essentially not correlated with personal characteristics. On average, contributions decrease by 10% from round 1 to round 10, and there is an effect of men giving 4.5% more in Model 1 and 6.0% more in Model 3. In the Two Types treatment, the effect on men is not significant, but high types contribute 46.2% more than low types.

Importantly, these results show that personal characteristics are of little help in predicting others' behavior. While men contribute more in the first three treatments, the effect is not significant in the Two Types treatment. Race and beauty are not correlated with behavior at all. The weakness of personal characteristics as an explanation of behavior will be useful in interpreting the results on selection of group members.

5.2 How Were People Ranked?

We have seen that personal characteristics explain little, if any, of behavior. But, are personal characteristics used when choosing group members? Table 3 reports how individual rankings are affected by the age, sex, race, beauty and expected rank (defined below) of potential group members.[30] The dependent variable is the rank that a person is given. A person with a rank of 1 is ranked highest and a person with a rank of 19 is ranked lowest. This means that if a coefficient is positive then the variable associated with it tends to lower one's rank. If a coefficient is negative the presence of the covariate tends to improve one's rank.

[30] The results in Table 3 are robust to alternative estimations, including Ordered Logit instead of OLS, OLS with clustered errors instead of fixed effects, random-effects Tobit and OLS with fixed effects for rankers and random effects for the participant being ranked. The results are also robust if we use racial intensities (continuous), instead of dummy variables, and attractiveness intensities (continuous), instead of dummy variables. We report results from the OLS regressions because of ease of interpretation of the coefficients.

Rank is regressed on the following independent variables: age, sex, race, beauty and expected rank of behavior in sequence two. The race and beauty dummy variables are the same as were used in the regression in Table 2. Because of random assignment to session, age, sex, race and beauty are all exogenous variables.

We also want to control for average contribution behavior in the treatments that showed this information when participants ranked others. However, average contributions are not strictly comparable across experimental sessions. In one session, an average contribution of 10 tokens may be the highest contribution but it may lie in the middle of the distribution in another. Therefore, to make sessions comparable, we create a variable called Expected Rank. This variable assigns a rank to each contribution in the distribution with a rank of one going to the highest contribution. This means that if a person had the highest average contribution in sequence two in that session, then the expected rank would be one. The lowest contributor has an expected rank of 19, and any ties are assigned the average rank. The estimated coefficient on this variable should be 1 if information on others' behavior is the only relevant information in creating ranks. Again, because of random assignment to groups in sequence two, expected rank will be exogenous.

Results show that people seem to understand that having high contributors in the group is the best strategy. For instance, expected rank alone explains 67% of the variance of ranks in Contribution Only (not shown in Table 3).[31] In all treatments where information on previous contribution is provided, expected Rank is a strong predictor of rank and explains a large part of the variation.

Interestingly, despite the fact that race and beauty have no bearing on the contribution choices of people in the experiment, they tend to predict the way people are ranked in the Photo Only treatment, as shown in Table 3. While men give slightly more than women, they are ranked on average 1.6 to 1.9 ranks lower. Without controlling for beauty, white participants are ranked 1.7 ranks higher and indigenous participants are ranked 0.95 ranks lower. When beauty is added, Whites are still ranked higher (now 1.3 ranks higher), but indigenous participants are no longer significantly ranked lower. It is unattractive participants that are ranked 2.4 ranks lower. Being

[31] The coefficient on Expected Rank is 0.82 and is significant. The relationship between Expected Rank and ranking is not one to one because not all participants ranked others strictly from highest contribution to lowest contribution.

attractive helps raise one's ranking by one rank, and being unattractive really hurts.[32] Beauty also helps explain another 4% of the variation.

Who is doing the discriminating in the Photo Only treatment? Table 4 shows results conditioning on the sex, race or attractiveness of the one doing the ranking. All groups rank unattractive people lower, and it is significantly lower for all but one group. Women, Whites and attractive people rank Whites significantly higher. All groups, rank men lower, but it is not significant for Whites. In-group effects are very strong for Whites, with a strong preference for Whites and attractive people. Unattractive people really do not want unattractive people in their group.

It is important to note that race and beauty are highly correlated. Seventy-eight percent of indigenous participants are classified as unattractive, and 78% of white participants are classified as attractive. This is extremely telling, since this is a result of the intersection of two *separate and independent* sets of raters, one for race and one for beauty. This means that it is not that one rater perceives the participant to be both indigenous and unattractive, but one rater perceives the participant to be indigenous and *another* rater perceives him to be unattractive. Combining the two ratings, we see that the majority of indigenous participants are also classified as unattractive. The results in Table 4 show that, for attractive participants, it is mainly their whiteness that boosts their rankings, but for indigenous participants, it is their lack of beauty that lowers their rankings.[33]

While race, beauty and sex affect rankings in Photo Only, rankings in treatments where information on past contribution behavior is available are affected only by behavior and beauty. Looking again at Table 3, in Contribution and Photo and Two Types, a large percent of the variation, between 36-61%, can be explained by expected rank. Beauty is significant, in that unattractive people are ranked lower in Contribution and Photo and attractive people are ranked higher in Two Types.[34] While beauty is still important in these regressions, it is far less so when information on behavior is available. The magnitude of the coefficient on Unattractive in the Contribution and Photo regression is one third the size of that in the Photo Only regression.

[32] This is consistent with Andreoni and Petrie's (2008) finding that people expect attractive people to be more cooperative.
[33] The correlation between beauty and skin color is also found in the U.S. Lighter skin is considered to be more attractive (see Hunter, 2002, and Hill, 2002).
[34] We do not report rankings by subgroups (men/women, white/indigenous, attractive/unattractive) for the Contribution and Photo and Two Types treatments because there is no particular pattern by subgroup.

It is important to recall the purpose of the Two Types treatment. If contribution behavior and personal characteristics are strongly correlated, then the Contribution and Photo treatment will not allow us to cleanly measure which variables, contribution or characteristics, affect rankings. The Two Types treatment allows us to do so by making the correlation between performance and characteristics weak and not predictable. The results from the Two Types treatment show that both beauty and performance affects rankings, even when the cost to discriminating is very high. Indeed, the results from Tables 2 and 3 together suggest that there is some taste-based discrimination in preferences over group members. We explore this further in Section 5.4.

5.3 Most and Least-Preferred Groups

To test the robustness of our results of preferences for group members, we run some further checks by looking at the extremes of the rankings. It might be that the relationship between personal characteristics and ranking is non-linear across the full list of ranking. Participants might pay more attention to the top and the bottom of the list. That is, they may pay attention to who they place in the top four on the list because those people would be in the most-preferred group. Also, they may pay attention to who they place in the bottom four on the list because those people would be in the least-preferred group. To investigate this, we run a logit model to see if the probability of making it to the top 4 or the bottom 4 on the ranking list is affected by the following independent variables: personal characteristics and the expected probability of making it to the top or bottom group. The latter variable is constructed as follows. If the participant's expected rank was strictly less than 5, then the participant was assigned the value of one for the expected probability of making it to the top group. If the participant's expected rank was strictly greater than 15, then the participant was assigned the value of one for the expected probability of making it to the bottom group. Tables 5 and 6 show these results.

Race is not a factor in making it to the top 4, but beauty is. In Photo Only, an attractive person is more likely to be in the top 4, and an unattractive person is not. Men are also less likely to be in the top 4. When past performance is available, people who are higher contributors are more likely to be ranked in the top 4.

In Two Types, unattractive people are less likely to be in the top 4. Since the parameters in a logit regression measure changes in the log of odds ratio, the estimates in Table 5 imply that

if the odds of making to the top are 1 in 4 (the probability of making it to the top group if groups are formed randomly), being unattractive would drop the odds to about 1 in 10 in both Photo Only and Two Types. This effect is very large.

For the bottom 4, beauty is also a significant predictor. An unattractive person is more likely to be in the bottom 4 in Photo Only, as are men. In Contribution and Photo and Two Types, the only significant variable is the person's past contribution.

5.4 Explanation of Two Types Ranking

We designed the Two Types treatment to see how group selection is affected when the costs of discriminating are raised. This, we argue, is a stronger test of taste-based discrimination. Since we do observe people using non-payoff relevant information (beauty) to rank group members in Tables 3 and 5, before we conclude that this is due to taste-based discrimination, it is worth exploring some alternative explanations. It could be that people believe that others will change their behavior in the third sequence. It might be risk aversion, or it could be statistical discrimination. We look at each of these explanations in turn.

First, for third sequence behavior, the rankings in the Two Types treatment are consistent with the belief that attractive people will increase contributions by 10% and unattractive people will decrease contributions by 8%.[35] If we consider the actual mean and variance of average contributions of attractive and unattractive participants (known by participants when they did the ranking), there is no significant difference. So, this means that people would need to believe that behavior by attractive and unattractive participants would change significantly (and in opposite directions) in the third sequence. In actuality, behavior in sequence 3 is strongly predicted by behavior in sequence 2 (correlation of 0.78). For these reasons, this explanation for differential rankings by beauty is more difficult to believe.[36]

Second, could it be due to risk aversion? Risk aversion would predict that, given the same mean, more variable participants should be ranked lower. Attractive participants are less variable in their behavior (although not significantly), and they are ranked higher. So, this is consistent with the prediction. However, the behavior of men is significantly more variable, so they should

[35] Assuming that average behavior in sequence 3 is perfectly predicted by average behavior in sequence 2, we calculate how much contributions would have to change to be consistent with the rankings given. This is done by taking the difference between the contribution of the ranking given to the person and their actual contribution and averaging across all rankings.
[36] We could always find a set of beliefs on behavior that would justify rankings.

be ranked lower. [37] They are not. This leaves us with the conclusion that the differential ranking for attractive people in the Two Types treatment is less likely due to risk aversion.

Finally, could it be statistical discrimination (even though we randomly assigned participants to behavioral incentives)? According to Phelps (1972), statistical discrimination can be thought of as an error-in-variables problem. Participants might judge similar evidence on performance differently if either behavior of some groups is more variable or participants have less informative priors on the behavior of others, for example, as social distance increases. Regarding the latter explanation, if we measure social distance by observable characteristics (beauty, gender, race), there is no evidence that different groups of people doing the ranking are any more likely to rank attractive people higher, so it does not appear to be an issue of social distance.[38] Regarding the former explanation, the lower variability of attractive versus unattractive participants was common knowledge since all participants saw the contributions of all participants and their pictures.

This above argument would suggest that the attenuation bias on the parameter associated with expected rank will also be associated with appearance. Indeed, regressions, shown in the Appendix, with interaction terms for attractiveness and expected rank show that the impact of expected rank is smaller for attractive participants. This means that people are more likely to disregard the same performance from an attractive person than from an unattractive person. This seems to contradict the fact that information on attractive people is more reliable (a similar argument has been made by Aigner and Cain, 1977 and Cornell and Welch, 1996). Because attractive participants are *less* variable in their behavior, we would expect people to pay more attention to their behavior when ranking rather than disregarding it.[39]

This leaves us to conclude that ranking in the Two Types treatment is more likely due to taste-based discrimination. Indeed, one advantage of the Two Types treatment is that, because it makes relying on stereotypes to assess future behavior imprecise, discrimination based on

[37] More variable behavior by men was also found by Andreoni and Vesterlund (2001) and Andreoni and Petrie (2008).
[38] Our measure of social distance by observable characteristics is weak. Attractive people do interact with unattractive people at some point in their lives. The argument we make is that if there is sorting by race (which is also strongly correlated with attractiveness), then unattractive people will likely have less experience and knowledge of the behavior of attractive people (and vice versa). This could produce rankings based on attractiveness.
[39] Phelp's (1972) model suggests that the slope parameter on Expected Rank should be steeper with lower variance. The same argument applies to a model based on beliefs that unattractive people are either confused or cognitively challenged. Bad signals from attractive people must be taken more seriously than those from unattractive people.

stereotypes becomes very costly. So, observed discrimination in this treatment is more consistent with taste-based discrimination.

6. Costs of Discrimination

In this section, we quantify the costs of discriminatory preferences in our group or team formation setting. We ask two questions, how would earnings change if an individual changed his rankings to no discrimination, but everyone else maintains their original rankings? And, how do earnings change if discrimination based on physical characteristics was completely eliminated?

To address these questions, we run bootstrapped estimations to compare alterative scenarios. In each bootstrap, we take the rankings submitted by participants, run the group formation procedure to make groups, and calculate expected payoffs (using average contribution from sequence two) under two different scenarios.[40] The expected payoffs are calculated from the point of view of the person doing the ranking. In the first scenario, we calculate expected payoffs for everyone when we change one randomly-chosen individual's rankings to one where rankings are based only on average contribution (with the highest contributor ranked the highest), but everyone else's submitted rankings are maintained. In the second scenario, we calculate expected payoffs when we change everyone's rankings to be based only on average contribution. The expected payoffs calculated under each scenario are compared to expected payoffs using the submitted rankings.

The results are illuminating. Expected payoffs under the first scenario, when an individual does not discriminate, increase by 1.5% in Photo Only, 0.6% in Contribution and Photo, and 3% in Two Types. In the second scenario, the absolute difference in expected payoffs when discrimination is completely eliminated is 5% in Photo Only, 2% in Contribution and Photo, and 13% in Two Types.[41]

These differences in expected earnings show that people are willing to give up significant amounts of money to not be in groups with unattractive people (or to be in groups with attractive people). The results also indicate that most of the difference in earnings comes from

[40] This procedure was bootstrapped 100,000 times to get estimates.
[41] For differences in expected payoffs when discrimination is completely eliminated, the sum of the difference in expected payoffs is zero, so we look at the absolute value of how much that difference changes.

discrimination by others (the second scenario). The difference in earnings is the largest in the Two Types treatment when the costs to discriminate are the highest.

7. Conclusions

We present a series of experiments aimed at determining the importance of appearance on team or group formation. Participants play a linear public goods game and, in a surprise, are allowed to choose group members for the last rounds of play. We recruited a diverse sample of individuals currently working in the labor market to participate in the experiments.

Our design systematically manipulates the information made available about others when sorting into groups. This allows us to examine what is more relevant to choosing teams or groups, information on past behavior or physical characteristics. The design is unique in that we can identify the effect of personal characteristics different from behavior and do not have to rely on measuring expectations. Indeed, by creating a situation where people who do not normally interact are asked to sort into groups and by providing information on appearance and behavior, we can cleanly test the source of differential treatment in the formation of teams or groups. The data and results we generate would be impossible to see in naturally-occurring observational data on group composition, especially if some groups of people never interact with each other.

Our results show that participant behavior is not correlated with personal characteristics, be it race or beauty. However, people do use the personal characteristics of others when given the opportunity to choose group members. Our results are consistent with taste-based discrimination. Interestingly, we find a great reduction in discrimination once information on others' behavior is provided. This is good news, since it shows that discrimination can be greatly diminished if relevant information is made available.

Participants tend to prefer groups of women and white-looking people and dislike groups with unattractive people. Unattractive participants only have a one in ten chance of being chosen in the most-preferred group, compared to attractive participants who have a one in three chance. While discrimination is markedly reduced by revealing information on others' behavior, there is still evidence that beauty and race are important factors, even when information is revealed and past behavior is made to be essentially uncorrelated with appearance. Intriguingly, while women and white-looking people are preferred in the absence of information, they are no more likely to make it to the top ranks when information is revealed. The effect of beauty, however, seems to

be constant. Being unattractive and looking indigenous are highly correlated, which suggests that the discrimination that indigenous people face could also be attributable to the fact that they are regarded as unattractive.

Without information on others' behavior, not everyone uses others' characteristics in ranking in the same way. This suggests some form of stereotyping in the absence of information. While there is agreement across sex and race that women are more desirable partners and unattractive people are less desirable partners, the effect of race on rankings is explained by the behavior of women and white participants.

Nonetheless, even in our robustness treatment, when people are provided information on past performance and the costs to discriminate are raised, discrimination remains. The differential treatment by personal characteristics in this treatment is difficult to reconcile with theories of statistical discrimination. Also, discrimination is very costly, especially in this treatment. In the absence of any discrimination, the absolute change in earnings would be 13%.

One may wonder about the external validity of our results.[42] We argue that it tells us something about behavior outside the experiment for two reasons. First, our sample is a good representation of the young, working population in urban Lima, Peru. Second, we would expect that in a laboratory environment, where decisions are tracked by the experimentalist, interactions within groups are short and there is no direct contact among group members, the amount of discrimination that we pick up would be lower than in a natural environment where people can hide their actions. Therefore, we view our results as representing a lower bound on the amount of discrimination in society at large. In future work, we will look at discrimination in this broader context.

Our research shows that understanding racial discrimination requires not only distinguishing its nature but also overcoming the problems caused by measurement and sorting. Indeed, our results are suggestive that racial discrimination might be masked as a beauty premium.[43] The fact that people sort into professions and teams requires that measurement of unequal treatment be done in tasks that are comparable and with a representative population. Experimental methods can be used to tackle these difficult identification problems. Our design

[42] Levitt and List (2006) raise into question the ability to extrapolate results from laboratory experiments to decisions in the real world. If there is a bias towards participants trying to appear fair, we should expect less evidence of discrimination. Falk and Heckman (2009) argue that there is much to be learned from lab experiments.

[43] We cannot test this with our data because we do not observe the necessary counterfactuals of a large number of "attractive" indigenous people or "unattractive" white people.

keeps the task constant, measures personal characteristics, and creates the necessary counterfactuals to identify the nature and extent of discrimination. We show that race, as well as beauty, is a discernible characteristic, and we find a large degree of agreement among our raters of a person's race and beauty.

Finally, there are some important policy implications from our work. People seem to have preconceptions of the behavior of others that create a barrier to access. That is, if people are excluded based on their appearance, those being excluded are denied the opportunity of showing what they are capable of doing. Given that once information is revealed most discrimination goes away, it seems that it would be recommendable to create opportunities for people to interact and to have independent, third-party measures of performance. For instance, professional accreditation might lower barriers to entry to those otherwise disadvantaged by eliminating the stereotypes associated with lower tier institutions or schools.

References

Aigner, Dennis and Glen Cain. 1977. Statistical theories of discrimination in labor markets. *Industrial and Labor Relations Review* 30 (2): 175-187.

Alchain, Armen and Harold Demstz. 1972. Production, information costs, and economic organization. *American Economic Review* 62(5): 777-795.

Altonji, Joseph and Rebecca Blank. 1999. Race and Gender in the Labor Market. In Ashenfelter, Orley and David Card. *Handbook of Labor Economics* vol 3: 3143-3259.

Altonji, Joseph and Charles Pierret. 2001. Employer Learning and Statistical Discrimination. *Quarterly Journal of Economics* 116(1): 313-350.

Andreoni, James, and Ragan Petrie. 2008. Beauty, gender and stereotypes: evidence from laboratory experiments. *Journal of Economic Psychology* 29: 73-93.

Andreoni, James, and Lise Vesterlund. 2001. Which is the fair sex? Gender differences in altruism. *Quarterly Journal of Economics* 116(1): 293-312.

Anwar, Shamena, and Hanming Fang. 2006. An alternative test of racial prejudice in motor vehicle searches: theory and evidence. *American Economic Review* 96: 127-51.

Arrow, Kenneth. 1972. Models of job discrimination. In *Racial Discrimination in Economic Life*, ed. Anthony H. Pascal. Lexington, MA: Lexington Books.

Bardsley, Nicholas. 2008. Dictator game giving: altruism or artefact? A note. *Experimental Economics* 11(2): 122-133.

Becker, Gary. 1975. *The economics of discrimination.* 2nd ed. Chicago, IL: University of Chicago Press.

Bertrand, Marianne, and Sendhil Mullainathan. 2004. Are emily and greg more employable than lakisha and jamal: a field experiment on labor market discrimination. *American Economic Review* 94: 991-1013.

Biddle, Jeff and Daniel Hammermesh. 1998. Beauty, productivity, and discrimination: lawyers' looks and lucre. *Journal of Labor Economics* 26(1): 172-201.

Blau, Francine and Marianne Ferber. 1992. *The economics of women, men, and work*. 2nd ed. Englewood Cliffs, NJ: Prentice-Hall Press.

Bogomolnaia, Anna, and Matthew Jackson. 2002. The stability of hedonic coalition structures. *Games and Economic Behavior* 38(2): 201-230.

Cain, Glen. 1986. The economic analysis of labor market discrimination: a survey. In Ashenfelter, Orley and Richard Layard. *Handbook of Labor Economics* vol 1:693-785.

Castillo, Marco, and Ragan Petrie. 2010. Discrimination in the lab: does information trump appearance? *Games and Economic Behavior* 68(1): 50-59.

Cornell, Bradford and Ivo Welch. 1996. Culture, information, and screening discrimination. *The Journal of Political Economy* 104(3): 542-571.

Croson, Rachel. 2000. Thinking like a game theorist: factors affecting the frequency of equilibrium play. *Journal of Economic Behavior and Organization* 41: 299-314.

Darity, William, Jason Dietrich and Darrick Hamilton. 2005. Bleach in the rainbow: latin ethnicity and preference for whiteness. *Transforming Anthropology* 13(2): 103-10.

Dickinson, David, and Ron Oaxaca. 2009. Statistical discrimination in labor markets: an experimental analysis. *Southern Economic Journal* 76(1): 16-31.

Eckel, Catherine, and Rick Wilson. 2006. Judging a book by its cover: beauty and expectations in the trust game. *Political Research Quarterly* 59(2): 189-202.

Falk, Armin and James Heckman. 2009. Lab experiments are a major source of knowledge in the social sciences. *Science* 326(5952): 535-538.

Fershtman, Chiam, and Uri Gneezy. 2001. Discrimination in a segmented society: an experimental approach. *Quarterly Journal of Economics* 116(1): 351-77.

Goldsmith, Arthur, Darrick Hamilton, and William Darity. 2006. Shades of discrimination: skin tone and wages. *American Economic Review, Papers and Proceedings* 96(2): 242-245.

Gyimah-Brempong, Kwabena, and Gregory Price. 2006. Crime and punishment: and skin hue too? *American Economic Review Papers and Proceedings* 96 (2): 246-250.

Hammermesh, Daniel, and Jeffrey Biddle. 1994. Beauty and the labor market. *American Economic Review* 84(5): 1174-94.

Harrison, Glenn and John List. 2004. Field experiments. *Journal of Economic Literature*, XLII (December): 1013-1059.

Heckman, James. 1998. Detecting discrimination. *Journal of Economic Perspectives* 12: 101-16.

Hill, Mark. 2002. Skin color and the perception of attractiveness among african americans: does gender make a difference? *Social Psychology Quarterly* 65(1):77-91.

Holmstrom, Bengt. 1982. Moral hazard in teams. *The Bell Journal of Economics* 13(2): 324-340.

Hunter, Margaret. 2002. If you're light you're alright: light skin color as social capital for women of color. *Gender and Society* 16(2): 175-193.

Kahn, Lawrence. 1991. Discrimination in professional sports: a survey of the literature. *Industrial and Labor Relations Review* 44: 395-418.

Knowles, James, Nicoli Persico, and Petra Todd. 2001. Racial bias in motor vehicle searches: theory and evidence. *Journal of Political Economy* 109: 203-229.

Langlois, J., Klakanis, L., Rubenstein, A., Larson, A., Hallam, M., & Smoot, M. 2000. Maxims or myths of beauty? A meta-analysis and theoretical review. *Psychological Bulletin* 126(3): 390–423.

Ledyard, John. 1995. Public goods: A survey of experimental research. In Kagel, John, and Alvin Roth. *Handbook of Experimental Economics*. Princeton, NJ: Princeton University Press.

Levitt, Steven. 2004. Testing theories of discrimination: evidence from Weakest Link. *Journal of Law and Economics* 47: 431-52.

Levitt, Steven, and John List. 2007. What do laboratory experiments measuring social preferences tell us about the real world. *Journal of Economic Perspectives* 21(2): 153-174.

List, John. 2004. The nature and extent of discrimination in the marketplace: evidence from the field. *Quarterly Journal of Economics* 119(1): 49-89.

List, John. 2006. Friend or foe? A natural experiment of the prisoner's dilemma. *Review of Economics and Statistics* 88(3): 463-471.

List, John. 2007. On the interpretation of giving in dictator games. *Journal of Political Economy* 115(3): 482-493.

Manski, Charles. 2004. Measuring expectations. *Econometrica* 72(5): 1329-76.

Mobius, Markus, and Tanya Rosenblatt. 2006. Why beauty matters. *American Economic Review* 96(1): 222-235.

Moreno, Andres, Hugo Nopo, Jaime Saavedra, and Maximo Torero. 2004. Gender and racial discrimination in hiring: a pseudo audit study for three selected occupations in metropolitan lima. IZA Working Paper No. 979.

Ñopo, Hugo, Jaime Saavedra, and Maximo Torero. 2007. Ethnicity and earnings in a mixed-race labor market. *Economic Development and Cultural Change* 55(4): 709-734.

Oaxaca, Ronald. 1973. Male-female wage differentials in urban labor markets. *International Economic Review* 14(3): 693-709.

Palfrey, Thomas, and Jeffrey Prisbrey. 1997. Anomalous behavior in public goods experiments: how much and why? *American Economic Review* 87(5): 829-846.

Phelps, Edward. 1972. The statistical theory of racism and sexism. *American Economic Review* 62: 659-661.

Pope, Devin and Justin Sydnor. 2011. What's in a picture? Evidence of discrimination from Prosper.com. *Journal of Human Resources* 46(1): 53-92.

Ravina, Enrichetta. 2008. Love & loans: the effect of beauty and personal characteristics in credit markets. Working Paper.

Riach, P., and J. Rich. 2002. Field experiments of discrimination in the market place. *Economic Journal* 112: F480–F518.

Table 1: Experimental Treatments

		Photo Shown	
		Yes	No
Information Given	Yes	Contribution and Photo (α_i = 2 centimos) Two Types ($\alpha_i \in$ {0.5, 5.0} centimos)	Contribution Only (α_i = 2 centimos)
	No	Photo Only (α_i = 2 centimos)	n/a

Table 2: Percent of Endowment Contributed to the Public Good (Sequence 2)
OLS Regression

VARIABLES	(1) Contribution Only, Photo Only, & Contribution and Photo Treatments Combined	(2) Two Types Treatment Only	(3) All Treatments
Male	4.52*	10.66	6.00**
	(2.32)	(6.88)	(2.46)
Age (years)	0.12	-0.67	-0.11
	(0.26)	(0.64)	(0.27)
Education (years)	0.40	-1.78	-0.39
	(0.82)	(1.55)	(0.78)
White	0.09	-6.08	-0.64
	(3.40)	(7.26)	(3.11)
Indigenous	-0.44	-3.50	-1.07
	(3.64)	(9.94)	(3.45)
Attractive	-1.14	8.20	0.87
	(3.58)	(9.54)	(3.44)
Unattractive	-1.56	6.86	1.06
	(3.63)	(7.75)	(3.46)
Low Type			-23.46***
			(8.87)
High Type		46.19***	19.91**
		(6.24)	(8.48)
Round	-1.19***	-0.12	-0.93***
	(0.25)	(0.39)	(0.21)
Constant	25.49*	60.73**	47.87***
	(14.11)	(27.96)	(13.63)
Individual Cluster Errors	*Yes*	*Yes*	*Yes*
Group Fixed Effects	*Yes*	*Yes*	*Yes*
Observations	1200	400	1600
R-squared	0.18	0.55	0.44

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.10. Results in the table also hold if specified as a random-effects Tobit with group fixed effects, instead of OLS.

Table 3: Individual Ranking (highest = 1, lowest = 19)
OLS Regression

VARIABLES	(1) Photo Only	(2) Photo Only	(3) Contribution and Photo	(4) Contribution and Photo	(5) Two Types	(6) Two Types
Age (years)	-0.00	-0.01	0.03	0.02	-0.02	-0.04
	(0.05)	(0.05)	(0.03)	(0.03)	(0.03)	(0.04)
Male	1.90***	1.58***	0.04	0.03	-0.03	0.23
	(0.41)	(0.41)	(0.25)	(0.25)	(0.31)	(0.32)
White	-1.72***	-1.31**	-0.13	-0.26	-0.03	0.55
	(0.53)	(0.54)	(0.27)	(0.32)	(0.40)	(0.50)
Indigenous	0.95**	-0.92	0.20	-0.21	0.17	-0.39
	(0.48)	(0.59)	(0.28)	(0.34)	(0.37)	(0.43)
Attractive		-0.91*		0.23		-0.84*
		(0.53)		(0.37)		(0.47)
Unattractive		2.39***		0.72**		0.63
		(0.53)		(0.33)		(0.40)
Expected Rank			0.84***	0.83***	0.66***	0.67***
			(0.02)	(0.02)	(0.03)	(0.03)
Constant	8.87***	9.13***	0.91	0.94	4.01***	4.25***
	(1.38)	(1.46)	(0.81)	(0.88)	(0.99)	(1.01)
Observations	760	760	760	760	760	760
R-squared	0.05	0.09	0.71	0.71	0.44	0.44

Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.10. Results in the table also hold if race and attractiveness are defined as continuous (rather than dummy) variables or as dummy variables using a quartile cutoff (instead of a tercile cutoff). All specifications also hold if estimated as an ordered logit, instead of OLS.

Table 4: Individual Ranking (highest = 1, lowest = 19), Photo Only Treatment OLS Regression

VARIABLES	(1) Men	(2) Women	(3) Whites	(4) Indigenous	(5) Attractive	(6) Unattractive
Age (years)	0.01	-0.05	-0.12	0.05	-0.11	0.11
	(0.06)	(0.08)	(0.11)	(0.12)	(0.10)	(0.08)
Male	1.54***	1.68**	1.29	2.53***	2.21***	1.62**
	(0.50)	(0.72)	(0.92)	(0.93)	(0.83)	(0.68)
White	-1.08	-1.71*	-3.27***	-0.17	-1.99*	-0.93
	(0.67)	(0.91)	(1.25)	(1.22)	(1.11)	(0.89)
Indigenous	-1.32*	-0.26	-0.29	0.34	-1.36	-0.73
	(0.73)	(1.01)	(1.35)	(1.28)	(1.15)	(1.01)
Attractive	-1.05	-0.61	-2.25*	0.24	-1.51	-0.01
	(0.65)	(0.92)	(1.21)	(1.06)	(1.11)	(0.86)
Unattractive	2.84***	1.72*	2.82**	3.20**	1.68	3.75***
	(0.67)	(0.89)	(1.14)	(1.26)	(1.04)	(0.91)
Constant	8.47***	10.08***	12.46***	5.72*	11.86***	5.01**
	(1.82)	(2.47)	(3.12)	(3.41)	(2.93)	(2.42)
Observations	494	266	133	171	190	266
R-squared	0.10	0.08	0.25	0.12	0.11	0.13

Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.10. Results in the table also hold if race and attractiveness are defined as continuous (rather than dummy) variables or as dummy variables using a quartile cutoff (instead of a tercile cutoff). All specifications also hold if estimated as an ordered logit, instead of OLS.

Table 5: Probability of Making it to the Top 4
Logit Regression

VARIABLES	(1) Photo Only	(2) Contribution & Photo	(3) Two Types
Age (years)	-0.01	-0.01	0.05*
	(0.02)	(0.04)	(0.03)
Male	-0.55***	-0.03	-0.54*
	(0.19)	(0.37)	(0.33)
White	0.35	-0.20	0.07
	(0.24)	(0.42)	(0.37)
Indigenous	0.44	-0.42	-0.11
	(0.33)	(0.49)	(0.32)
Attractive	0.42*	0.40	-0.16
	(0.23)	(0.44)	(0.36)
Unattractive	-1.01***	-0.32	-0.76**
	(0.31)	(0.45)	(0.30)
Expected to be in Group		4.71***	3.62***
(based on Expected Rank)		(0.33)	(0.35)
Constant	-0.84	-2.79***	-3.24***
	(0.69)	(1.03)	(0.83)
Observations	760	760	760
Log Likelihood	-367.87	-174.10	-269.93

Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.10

Table 6: Probability of Making it to the Bottom 4
Logit Regression

VARIABLES	(1) Photo Only	(2) Contribution & Photo	(3) Two Types
Age (years)	-0.02	0.07	-0.03
	(0.03)	(0.05)	(0.03)
Male	0.42**	0.23	-0.21
	(0.21)	(0.48)	(0.23)
White	0.01	0.28	0.07
	(0.28)	(0.64)	(0.41)
Indigenous	-0.30	-0.40	0.05
	(0.26)	(0.56)	(0.32)
Attractive	-0.24	0.10	-0.29
	(0.29)	(0.69)	(0.35)
Unattractive	1.02***	0.84	0.43
	(0.24)	(0.57)	(0.29)
Expected to be in Group		5.75***	3.03***
(based on Expected Rank)		(0.41)	(0.24)
Constant	-1.25*	-5.77***	-1.54**
	(0.71)	(1.46)	(0.79)
Observations	760	760	760
Log Likelihood	-372.49	-128.48	-276.13

Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.10

Appendix

Table A1: Individual Ranking (highest = 1, lowest = 19)
OLS Regression

VARIABLES	(1) Contribution and Photo	(2) Two Types
Age (years)	0.02	-0.01
	(0.03)	(0.04)
Male	-0.00	0.11
	(0.27)	(0.33)
White	-0.22	0.19
	(0.34)	(0.53)
Indigenous	-0.14	-0.26
	(0.36)	(0.46)
Attractive	0.67	0.82
	(0.77)	(0.93)
Unattractive	1.03	1.29
	(0.79)	(0.83)
Expected Rank	0.86***	0.72***
	(0.06)	(0.04)
Expected Rank*Attractive	-0.05	-0.15**
	(0.07)	(0.07)
Expected Rank*Unattractive	-0.03	-0.06
	(0.07)	(0.07)
Constant	0.57	3.09**
	(1.05)	(1.23)
Observations	760	760
R-squared	0.71	0.45

Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.10