Beautiful or White? Discrimination in Group Formation

Marco Castillo*
Georgia Institute of Technology
marco.castillo@pubpolicy.gatech.edu

Ragan Petrie
Georgia State University
rpetrie@gsu.edu

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

May 2008

Abstract

We explore the importance of appearance in group formation using a series of experiments. Subjects choose partners, and we manipulate the amount of payoff-relevant information on behavior, thereby making it costly to discriminate based on appearance. Beauty predicts desirability, yet it might mask racial preferences. Payoff-relevant information reduces discrimination a great deal, yet discrimination based on appearance remains. Behavior across groups is the same, but unattractive subjects have a one in ten chance of making it to the most preferred group, whereas attractive subjects have a one in three chance. This is most consistent with taste-based, rather than statistical, discrimination.

* The authors would like to thank the Inter-American Development Bank for research funding. We also thank Kevin Ackaramongkolrotn, John List, Andrea Moro, Jorge de la Roca, David Solis, and Néstor Valdivia. This paper has benefited greatly by comments from seminar participants at University of Pittsburgh, the LACEA Meetings in Mexico City, Mexico and 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.

Introduction

One can imagine how sorting by physical appearance might emerge. Absent relevant information on behavior, people may resort to physical cues, such as race, gender or beauty, to make inferences on behavior. Perceptions (or misperceptions) may result in exclusion of certain groups based solely on physical appearance. If enough sorting takes place, interactions among certain segments of the population may never transpire. Therefore, it may be difficult to say how one group interacts with another because no such interactions occur. Theories of statistical discrimination (Arrow, 1972; Phelps, 1972) argue that differential treatment is due to a lack of information. Since relying on poor signals when good ones are available is potentially costly, providing relevant information on behavior should raise the cost of differential treatment based on physical cues and should reduce or eliminate differences. The research on discrimination is extensive and rich, yet to what extent people are willing to incur costs to discriminate in their choice over partners is still an open question. In this research, we look to multiracial settings and ask what we can learn about the extent and nature of discrimination in how groups form.

We use economic experiments to study how physical appearance interacts with preferences over group members. Experiments allow us to present the necessary counterfactual situations to study the nature of discrimination. In the experiments, we present payoff-relevant information to people, thereby making it costly to discriminate in choosing partners. In addition, we create incentives for people to behave counter to stereotypes, and we further increase the cost to discriminate. Despite all of this, we find that people still discriminate. Conditioning on payoff-relevant information, unattractive people have 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. Discrimination is costly. Without it, there would be a 13% absolute change in earnings. We also find that race and beauty are confounded in the minds of subjects. This suggests that a revealed preference for beauty might mask racial discrimination.

We use Peru as our laboratory because of its rich multiracial heritage. 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.[1] 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.[2] Also, in segmented societies, like Peru, interactions across certain groups will likely be limited and hierarchical as populations are sorted into different occupations.[3] 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.[4] Artefactual economic experiments (Harrison and List, 2005) allow us to explore these issues more fully.[5, 6]

Our experimental design addresses many of the issues raised above. First, our experiments test the hypothesis of the existence of taste-based discrimination 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 costly to base one's decision on. We argue that, regardless of the fact that a subject's expectations on the future behavior of others might be correct or not or might be observed by the researcher or not, taste-based discrimination should be *robust* to information on performance. That is, taste-based discrimination does not suggest that people disregard information on appearance once information on performance is revealed. Our experiments can test this. We consider this approach -- manipulating information and

---

[1] 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.
[2] 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."
[3] 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.
[4] Castillo and Petrie (2007) exploit the fact that a recent civil conflict in Peru pitched populations with different backgrounds to show that human right abuses were not random.
[5] Sampling from integrated populations (like colleges) might bias our results toward not finding evidence of discrimination.
[6] 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.

performance 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).

We use a repeated linear public goods game to explore these issues. Repeated public goods experiments represent a natural environment to study group formation because payments in the experiment are a function of both individual and group behavior and mimic many social situations. The more cooperative are other group members, the more money a person makes. In our experiment, subjects are asked to 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 to subjects. Subjects are shown either digital photographs of others in the experiment or information on past performance (or both).

Because performance and appearance may be correlated, it is important that our experimental design addresses this issue. If they are, simply providing information on appearance and performance will not be enough to identify which element affects sorting. To observe the importance of appearance on sorting we need a counterfactual situation where behavior contradicts held beliefs. So, one of our treatments randomly assigns incentives so the correlation between performance and appearance is broken, and this allows us to identify discrimination for other than statistical reasons. Our environment is strategic and relevant to understanding how groups form.

Second, the study develops a measure of race based on intensity of a genotype (Torero, Saavedra, Nopo and Escobal, 2004; Nopo, Saavedra and Torero, 2007).[7] Separate questions were asked about how much White, Indigenous, Black, or Asian a person has. This task was performed by subjects not involved in the experiment, but recruited from the same population as the experimental subjects. In addition, another set of raters were recruited and trained to identify genotypical features from the pictures of subjects and ignore any feature related to looks or dress. As discussed later in the paper, the results are robust to either method used. Additionally, another

[7] Similar techniques were used before by Angel and Gronfein(1988) and Anderson, Silver and Abramson (1988).

set of subjects was recruited to rate experimental subjects according to physical attractiveness. This gives us independent measures of race and beauty.

Finally, in an attempt to observe non-student populations that normally do not interact perform the same task, we recruited among the working population between 20 and 35 years of age in Lima, Peru. The sample, while small, is similar to the population at large. Moreover, by restricting ourselves to the working population we diminish the common criticism that student populations 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.[8]

We find that the answer to the issue of racial 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. Once information on performance is provided, most evidence of discrimination is eliminated. However, despite the high cost of ignoring information on behavior, differential treatment remains.

There have been other studies trying to discern the nature of discrimination,[9],[10] 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

---

[8] Harrison and List (2004) discuss these issues extensively.

[9] 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.[9] 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*.

[10] Dickinson and Oaxaca (2006) examine distributional risk as it relates to detecting statistical discrimination.

(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.[11] Our study differs in that we test for discrimination by manipulating the equilibrium at the experimental level.

To our knowledge, our work is the first to present evidence consistent with taste-based discrimination in the experimental literature. The research shows the advantage of experimental methods in tackling difficult identification issues. It also shows the importance of measurement of personal characteristics and sampling in the study of race and beauty in experiments.

The Sample

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 twofold. First, we want to create an environment in which people of various social distances who might not normally interact with one another can. Second, at the same time, we want to have a sample of subjects which is representative of the young working population in metropolitan Lima. To this end, eligible subjects are between 20-35 years of age, live 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 so that subjects would be distributed homogenously across gender and income levels. To get our sample, we worked with two companies specialized in surveys and recruiting to help us

[11] 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.
[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.

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 subjects 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 give 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 out experimental subjects to a sub-sample from the *Encuesta Nacional de Hogares* (ENAHO) 2004. The sub-sample complies with the eligibility criteria for all of our subjects. 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 subjects 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 subjects are slightly more educated. This is most likely a reflection of the requirement in our experiment that subjects know how to use the internet. This comparison gives us confidence that the subjects in our experiment are a good representation of the larger population in metropolitan Lima.

Experimental Design

[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 subjects 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 subjects. 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.

We use a linear public goods game to explore discrimination in group formation. This design was first developed and used by Castillo and Petrie (2006). Each subject 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 subject.[16] Each token placed in the public investment yields a return of α_i to the subject 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 subjects in each experimental session. Subjects are randomly assigned to a five-person group and play 10 rounds with that same group. At the end of each round, subjects learn their payoff, π_i, and the total number of tokens contributed to the public investment by the group, G. Subjects make decisions privately on a computer and do not talk to one another. They do not interact with other subjects in any way other than through decisions on the computer.

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. Subjects 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 subject behavior. No personal or individual contribution information is revealed in the first 20 rounds of the game. We run two 10-round sequences before subjects choose their groups to give subjects experience with playing the game.

In order to create an incentive for people to reveal who they would prefer to be in their group, we create the following procedure. Subjects rank all the other 19 subjects in the session from 1 (most preferred) to 19 (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

[16] There are 100 centimos in 1 sol (the Peruvian currency). At the time of the study, US$1 = 3.2 soles.

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 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, subjects then see a screen with the information corresponding to the subjects in their new group. Subjects 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.[17] For this reason, we opted for the mechanism described above because it is easy to explain to subjects and can be implemented quickly once subjects have submitted their lists of rankings.

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 $\alpha_i = 2$ centimos, so the price of contributing to the public good is 2. It is in the group's interest for everyone to contribute their full endowment to the public investment, but each individual in the group maximizes his own payoffs by putting all his tokens in the individual investment. 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. Because groups are randomly assigned in the first and second sequences, all subjects 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 subject is any more likely to be 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 subjects are asked to rank others, they see the photos of all other subjects. And, in the Contribution and Photo treatment, subjects see the photo and the

[17] With 20 subjects, each subject would need to rank 3,876 groups.

average amount contributed to the public good in the second 10-round sequence. The average is listed below each subject's photo.

In the Two Types treatment, as in the Contribution and Photo treatment, when subjects are asked to rank others, subjects 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 subjects are randomly assigned a value of 0.5 and half are randomly assigned a value of 5.0. Subjects keep the same value for all 30 rounds of play. All subjects know this information before making decisions. A subject with an $\alpha_i = 5.0$ has a price of contributing to the public good of 0.8. If he is selfish or altruistic, he should invest his entire endowment in the public good. If he is not altruistic or is inequality averse, however, he might not contribute his full endowment, despite the low price of giving.[18] A subject with an $\alpha_i = 0.5$ has a price of contributing to the public good of 8, so investing in the public good is very expensive. We would expect subjects 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. Complete separation is not necessarily expected due to the asymmetry faced by subjects within a group. Because subjects are randomly assigned incentives, however, performance and appearance should not be correlated. The Two Types treatment is important to our ability to identify whether appearance or performance affects sorting.

Each treatment was run twice, and each experimental session had 20 subjects. An experimental session lasted at least two hours. In total, 160 subjects participated in the four treatments. Each session ended with an extensive post-experiment questionnaire. The 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 subjects were randomly assigned to treatments. Since most subjects worked full time, the experiments were conducted on weekend afternoons.

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

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).[19]

Race and Beauty Classifications

We are interested in knowing if people sort into groups based on physical characteristics. While a person's sex is easy to determine, a person's race or beauty is not. We want to develop an independent measure of the race and beauty of a person that reflects the general perception of that 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 subjects in the experiment, to rate the photos of the subjects in terms of race and beauty.[20] A rater only rated the photo in terms of one characteristic, race or beauty, not both.[21]

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 classifying a subject along one dimension of "white" or "mestizo," we evaluate subjects in 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.

To obtain these ratings, we had twenty people (10 women and 10 men), not involved in the public goods experiment, rate each subject along each of these four dimensions. Each dimension was rated from zero to ten, with zero being complete absence of the dimension and 10 being the most intense. Raters were instructed to choose whichever number between zero and 10 best described the person for each of the four racial dimensions. The four numbers did not need to

[19] The minimum wage in Peru is about $1/hour.
[20] This technique has been used in other experimental research on beauty, including Andreoni and Petrie (2007), Eckel and Wilson (2006), and Mobius and Rosenblat (2006). Hamermesh and Biddle (1994) had interviewers rate the interviewees in terms of beauty.
[21] 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.

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. There was a high degree of agreement among raters in terms of attractiveness. Pairwise correlations among raters ranged from 0.13-0.75, with an average of 0.50.[22]

In terms of agreement among raters on race, there was also a high degree of agreement in terms of each racial dimension. 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.[23]

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 rates. 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

[22] The Cronbach alpha for interrater reliability is 0.94.
[23] 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.

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.

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

In lieu of using average intensities of race or beauty, we create dummy variables for our analysis. 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 subjects are classified as White and 22.5% are classified as Indigenous.

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 subjects 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.

Basic Experimental Results

What Did People Contribute in the Experiment?

As is commonly observed in public goods experiments (see Kagel and Roth, 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. Low types in the Two Types treatment contribute about 30% and decline slightly in the last few rounds, and high types contribute roughly the same in early and late rounds. In the Two Types treatment, we do not observe a round effect, as the incentives to contribute little or a lot are strong and constant across rounds.

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 than low types. They contribute about 79% over all rounds, and low types contribute about 30%.[24] Clearly, not all subjects 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 split is uncorrelated with appearance. It is the differentiation in behavior and lack of correlation with appearance that we need to test 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)

[24] 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. 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.

more selfishly. Or, more affluent subjects 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.[25]

Included in the regressions are variables for sex, age, education, race, beauty and assigned type in the Two Types treatment. The variables for race and beauty are dummy variables constructed from the average standardized continuous measures as described in the previous section.

Table 2: Fixed Effect Regression on Percent of Endowment Contributed to the Public Good (Sequence 2)

Variables	Contribution Only, Photo Only, & Contribution and Photo	Two Types	All Treatments
Male	4.52 (0.05)	10.66 (0.13)	6.00 (0.02)
Age (years)	0.12 (0.65)	-0.67 (0.30)	-0.11 (0.69)
Education (years)	0.40 (0.63)	-1.78 (0.26)	-0.40 (0.62)
White	0.09 (0.98)	-6.08 (0.41)	-0.64 (0.84)
Indigenous	-0.44 (0.90)	-3.50 (0.73)	-1.07 (0.76)
Attractive	-1.14 (0.75)	8.20 (0.40)	0.87 (0.80)
Unattractive	-1.56 (0.67)	6.86 (0.38)	1.06 (0.76)
Low Type			-23.46 (0.00)
High Type		46.19 (0.00)	19.91 (0.02)
Round	-1.19 (0.00)	-0.12 (0.75)	-0.92 (0.00)
Constant	25.49 (0.07)	60.73 (0.04)	47.87 (0.00)
Individual Clustered Errors	*Yes*	*Yes*	*Yes*
Group Fixed Effects	*Yes*	*Yes*	*Yes*
R2	0.18	0.55	0.44
N	1200	400	1600
p-values in parentheses			

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 the first regression and 6.0% more in the pooled regression. High types in the Two Types treatment contribute 46.2% more than low types.

[25] The same results hold with random-effects regressions at the individual level and group-level fixed effects.

Table 2 shows that personal characteristics are of little help in predicting others' behavior. While men contribute more in the first three treatments, they do not in Two Types. It is important to note that Table 2 shows that 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 in the following section.

How Were People Ranked?

We have seen that personal characteristics explain little, if any, of behavior. But, are personal characteristics used when choosing groups? Table 3 reports individual ranker's fixed effects regressions of ranking on age, sex, race, beauty and expected rank for each treatment separately.[26] 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 rank 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.

The race and beauty dummy variables are the same as were used in the regression in Table 2. We also want to control for average contribution behavior in the treatments that showed this information when subjects 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 use a variable called Expected Rank. This variable indicates the rank that a person should have if only contributions to the public good are used to rank others. 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 expected coefficient on this variable should be 1 if information on others' behavior is the only relevant information in creating ranks.

---

[26] The results in Table 3 are robust to alternative estimations: OLS with clustered errors instead of fixed effects, random-effects Tobit and OLS with fixed effects for rankers and random effects for the subject being ranked. The results are also robust if we use racial intensities, instead of dummy variables, and attractiveness intensities, instead of dummy variables. The results are also similar if we use racial intensities of trained raters, instead of cohort raters.

All regressions include fixed effects of the person doing the ranking since each individual ranked 19 people.

Table 3: Fixed Effects Regression on Individual Ranking (highest = 1, lowest = 19)						
Variables	**Photo Only**		**Contribution & Photo**		**Two Types**	
Age (years)	-0.00 (0.96)	-0.01 (0.81)	0.04 (0.23)	0.03 (0.32)	-0.02 (0.52)	-0.04 (0.31)
Male	1.98 (0.00)	1.61 (0.00)	0.07 (0.79)	0.08 (0.77)	-0.03 (0.92)	0.23 (0.48)
White	-1.85 (0.00)	-1.35 (0.02)	-0.14 (0.62)	-0.24 (0.48)	-0.03 (0.94)	0.54 (0.29)
Indigenous	1.20 (0.02)	-0.84 (0.20)	0.22 (0.45)	-0.20 (0.60)	0.17 (0.66)	-0.40 (0.37)
Attractive		-0.89 (0.10)		0.15 (0.69)		-0.85 (0.08)
Unattractive		2.35 (0.00)		0.74 (0.03)		0.63 (0.12)
Expected Rank			0.83 (0.00)	0.82 (0.00)	0.66 (0.00)	0.67 (0.00)
Constant	8.83 (0.00)	9.12 (0.00)	0.65 (0.47)	0.70 (0.45)	4.01 (0.00)	4.27 (0.00)
Individual Fixed Effects	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Within-R2	0.06	0.09	0.71	0.71	0.44	0.44
N	760	760	760	760	760	760
p-values in parentheses						
Note that the above results are robust to specifications including other variables such as education, height, high school attended, mother tongue and religion.						

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).[27] Expected Rank remains a strong predictor of rank, and explains a large part of the variation, in all treatments where information on previous contribution is provided.

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. Men give slightly more than women but are ranked on average 1.6 to 2.0 ranks lower.

[27] 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 subjects ranked others strictly from highest contribution to lowest contribution.

Without controlling for beauty, white subjects are ranked 1.8 ranks higher and indigenous subjects are ranked 1.2 ranks lower. When beauty is added, Whites are still ranked higher, but indigenous subjects are not ranked lower. It is unattractive subjects that are ranked 2.4 ranks lower. Being attractive does not seem to help in the rankings, but being unattractive really hurts. Beauty also helps explain another 3% of the variation.

Who is doing the discriminating in the Photo Only treatment? Table 4 shows results conditioning on the sex or race of the one doing the ranking. All groups rank unattractive people lower. Women and Whites rank Whites significantly higher. All groups, but Whites, rank men lower.

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

Variables	Men	Women	Whites	Indigenous
Age (years)	0.01	-0.05	-0.12	0.05
	(0.86)	(0.59)	(0.29)	(0.69)
Male	1.54	1.74	1.37	2.57
	(0.00)	(0.02)	(0.15)	(0.01)
White	-1.10	-1.78	-3.40	-0.26
	(0.12)	(0.06)	(0.01)	(0.84)
Indigenous	-1.30	-0.03	0.07	0.51
	(0.10)	(0.98)	(0.96)	(0.71)
Attractive	-1.05	-0.57	-2.18	0.27
	(0.12)	(0.55)	(0.08)	(0.80)
Unattractive	2.83	1.60	2.67	3.09
	(0.00)	(0.09)	(0.03)	(0.02)
Constant	8.47	10.00	12.36	5.75
	(0.00)	(0.00)	(0.00)	(0.00)
Individual Fixed Effects	Yes	yes	yes	Yes
Within-R2	0.10	0.08	0.25	0.12
N	494	266	133	171

p-values in parentheses

It is important to note that race and beauty are highly correlated. Seventy-eight percent of indigenous subjects are classified as unattractive, and 78% of white subjects 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. In essence, it is not that one rater perceives the subject to be both indigenous and unattractive, but one rater perceives the subject to be indigenous and another rater perceives him to be unattractive. Combining the two ratings, we see that the majority of indigenous subjects are also classified as unattractive. For attractive subjects, it is mainly their whiteness that boosts their rankings, but for indigenous subjects, it seems that it is their lack of beauty that affects their rankings.[28]

While race, beauty and sex affect rankings in Photo Only, rankings in treatments where information on past performance is available are affected only by performance 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. Because the beauty ratings are gender specific, this means that both attractive men and attractive women are ranked higher in Two Types.

It is important to recall the purpose of the Two Types treatment. If contribution behavior and personal characteristics are 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 breaking the correlation between performance and characteristics. The results from the Two Types treatment give us confidence that both beauty and performance affects rankings, even though, by the experimental design of Two Types, beauty is orthogonal to performance. Indeed, the results from Tables 2 and 3 together suggest that there is some stereotyping or taste-based discrimination. We explore this further later in the paper.

Most and Least-Preferred Groups

It might be that the relationship between personal characteristics and ranking is non-linear across the full list of ranking. Subjects 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

[28] 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).

bottom four on the list because those people would be in the least-preferred group. To investigate this, we run a logit model, with individual fixed effects, to see if the probability of making it to the top 4 or the bottom 4 on the ranking list is affected by personal characteristics and the expected probability of making it to the top or bottom group. The latter variable is constructed as follows. If the subject's expected rank was strictly less than 5, then the subject was assigned the value of one for the expected probability of making it to the top group. If the subject's expected rank was strictly greater than 15, then the subject was assigned the value of one for the expected probability of making it to the bottom group. Tables 5 and 6 show these results.

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

Variables	Photo Only	Contribution & Photo	Two Types
Age (years)	-0.00 (0.82)	-0.03 (0.50)	0.04 (0.13)
Male	-0.54 (0.00)	-0.18 (0.63)	-0.53 (0.10)
White	0.35 (0.13)	-0.21 (0.60)	0.03 (0.94)
Indigenous	0.37 (0.28)	-0.47 (0.32)	-0.13 (0.66)
Attractive	0.38 (0.09)	0.49 (0.28)	-0.20 (0.56)
Unattractive	-0.94 (0.00)	-0.23 (0.59)	-0.79 (0.01)
Expected to be in Group (based on Expected Rank)		4.04 (0.00)	3.32 (0.00)
Individual Fixed Effects	*Yes*	*Yes*	*Yes*
Log-Likelihood	-308.30	-133.13	-217.94
N	760	760	760
p-values in parentheses			

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

Variables	Photo Only	Contribution & Photo	Two Types
Age (years)	-0.02 (0.35)	0.06 (0.18)	-0.03 (0.30)
Male	0.40 (0.05)	0.31 (0.53)	-0.20 (0.37)
White	0.03 (0.92)	0.24 (0.70)	0.05 (0.89)
Indigenous	-0.33 (0.22)	-0.39 (0.48)	0.02 (0.94)
Attractive	-0.24 (0.39)	0.04 (0.95)	-0.28 (0.42)
Unattractive	0.99 (0.00)	0.85 (0.13)	0.40 (0.16)
Expected to be in Group (based on Expected Rank)		4.71 (0.00)	2.75 (0.00)
Individual Fixed Effects	*Yes*	*Yes*	*Yes*
Log-Likelihood	-312.74	-95.54	-224.04
N	760	760	760

p-values in parentheses

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. The only significant variable in Contribution and Photo and Two Types is the person's past contribution.

Money Foregone by Differential Rankings

We would like to know how earnings would change if an individual did not discriminate or if there was no discrimination based on physical characteristics. To do so, we bootstrap expected payoffs given submitted rankings and compare these to expected payoffs in two different scenarios.[29] First, we look at expected payoffs if an individual ranked others only according to average contribution, but everyone else maintained their submitted rankings. Second, we look at expected payoffs if everyone ranked others only according to average contribution.[30]

The difference in expected payoffs when an individual does not discriminate is 1.5% in Photo Only, 0.6% in Contribution and Photo, and 3% in Two Types. The absolute difference in expected payoffs when there is no discrimination is 5% in Photo Only, 2% in Contribution and Photo, and 13% in Two Types.[31]

[29] Expected payoffs were bootstrapped 100,000 times.
[30] This is the same idea as a Oaxaca (1973) decomposition of wages.
[31] For the earnings differential when there is no discrimination, the sum of the difference in payoffs is zero, so we look at the absolute value of how much that difference changes at both extremes.

These differentials in earnings show that people are willing to give up significant payoffs 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 others' discrimination. The difference in earnings is the largest in the Two Types treatment.

Explanations for Differential Ranking in Two Types

There are several alternative explanations for the differential rankings we observe in the Two Types treatments. 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.

The rankings in the Two Types treatment are consistent with people believing that attractive people will increase contributions by 10% and unattractive people will decrease contributions by 8%.[32] If we consider the actual mean and variance of average contributions of attractive and unattractive subjects, there is no significant difference. So, this means that people would need to believe that behavior would change significantly (and in opposite directions) in the third sequence. Also, as previously reported, behavior in sequence 3 is strongly predicted by behavior in sequence 2 (correlation of 0.78). This explanation for differential rankings is more difficult to believe.[33]

Is it due to risk aversion? Risk aversion would predict that, given the same mean, more variable subjects should be ranked lower. Attractive subjects are less variable (although not significantly), and they are ranked higher. However, the behavior of men is significantly more variable, so they should be ranked lower. [34] 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.

[32] 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.
[33] We could always find a set of beliefs on behavior that would justify rankings.
[34] More variable behavior by men was also found by Andreoni and Vesterlund (2001) and Andreoni and Petrie (2007).

Could it be statistical discrimination? According to Phelps (1972), statistical discrimination can be thought of as an error-in-variables problem. Subjects might judge similar evidence on performance differently if either behavior of some groups is more variable or subjects have less informative priors on others behavior as social distance increase. Regarding the latter explanation, 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. Regarding the former explanation, the lower variability of attractive versus unattractive subjects was common knowledge since all subjects saw the contributions of all subjects and their pictures.

This 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 subjects. 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. Because attractive subjects are *less* variable in their behavior, we would expect people to pay more attention to their behavior when ranking rather than disregarding it.[35] This leaves us to conclude that ranking in the Two Types treatment is more likely do to taste-based discrimination. Indeed, the Two Types treatment is successful in distinguishing the sources of discrimination precisely because it forces *all* subjects to act counter to stereotype.

Conclusions & Policy Implications

We present a series of experiments aimed at determining the importance of appearance on group formation. Subjects play a linear public goods game and, in a surprise, are allowed to choose group members for the last rounds of play. Our experiments systematically manipulate the information made available about others when sorting into groups. This allows us to examine what is more relevant to group formation, information on past behavior or physical

[35] 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.

characteristics. Our design is unique in that we can identify the effect of personal characteristics from behavior and do not have to rely on measuring expectations. We recruited a diverse sample of individuals currently working in the labor market to participate in the experiments.

Our experiments show that subject 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 evidence of stereotyping or taste-based discrimination. Interestingly, our experiments also show a great reduction in discrimination or stereotyping 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.

Subjects tend to prefer groups of women and white-looking people and dislike groups with unattractive people. Unattractive subjects have a one in ten chance of being chosen in the most-preferred group, and attractive subjects have a one in three chance. While evidence of 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 performance is 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. Note that being unattractive and looking indigenous is highly correlated, so some effects of beauty may also be picking up effects of race.

Without information on others' behavior, not everyone uses others' characteristics in ranking in the same way. This suggests some form of stereotyping. 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, discrimination remains even when it is very costly to do so. The absolute change in earnings is 13% if there was no discrimination. Discrimination in this last treatment is difficult to reconcile with theories of statistical discrimination.

One may wonder about the external validity of our results.[36] 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 and interactions across groups are short, that the amount of discrimination that we pick up would be lower than in a natural environment where people can hide their actions. Therefore, we argue that our results represent a lower bound on the amount of discrimination in the 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 measuring race and sorting. Indeed, our results suggest that racial discrimination might be masked as a beauty premium.[37] The fact that people sort into professions requires that measurement of unequal treatment be done in tasks that are comparable and that subjects represent the population at large. Experimental methods can be used to tackle these difficult identification problems. Our design 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 agreements among people.

There are some important policy implications. 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 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.

---

[36] Levitt and List (2006) raise into question the ability to extrapolate results from laboratory experiments to decisions in the real world. If subjects are trying to appear fair, we should expect less evidence of discrimination.

[37] 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.

References

Anderson, B.A., Silver, B.D. & Abramson, P.R, "The effects of the race of the interviewer on race-related attitudes of black respondents in SRC/CPS national election studies. *Public Opinion Quarterly*, 52, 289-324, 1988.

Andreoni, J. and R. Petrie, "Beauty, Gender and Stereotypes: Evidence from Laboratory Experiments," forthcoming, *Journal of Economic Psychology*, 2007.

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

Angel, R. and W. Gronfein, "The Use of Subjective Information in Statistical Models, *American Sociological Review*, 53(3), 464-473, 1988.

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, Kenneth, "Models of Job Discrimination", in *Racial Discrimination in Economic Life*, Anthony H. Pascal, ed., Lexington Books, Lexington, MA, 83-102, 1972.

Bardsley, Nicholas, "Dictator Game Giving: Altruism or Artefact? A Note." forthcoming, *Experimental Economics*, 2007.

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.

Blau, F. and M. Ferber, *The Economics of Women, Men, and Work*, 2nd ed., Englewood Cliffs, NJ: Prentice-Hall, 1992.

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

Castillo, M. and R. Petrie, "Discrimination in the Warplace: Evidence from a Civil War in Peru," Working Paper, Georgia Institute of Technology, 2007.

Castillo, M. and R. Petrie, "Discrimination in the Lab: Experiments Exploring the Impact of Performance and Appearance on Sorting," Working Paper, Georgia State University, 2006.

Croson, R., "Thinking Like a Game Theorist: Factors Affecting the Frequency of Equilibrium Play," *Journal of Economic Behavior and Organization*, Vol 41, 299-314, 2000.

Darity, W., J. Dietrich and D. Hamilton, “Bleach in the Rainbow: Latin Ethnicity and Preference for Whiteness” *Transforming Anthropology*, 13(2), 103-10, 2005.

Dickinson, D. and R. Oaxaca, “Statistical Discrimination in Labor Markets: An Experimental Analysis,” Working Paper, Appalachian State University, 2006.

Eckel, C. and R. Wilson, “Judging a Book by its Cover: Beauty and Expectations in the Trust Game,” *Political Research Quarterly*, 59(2), 189-202, 2006.

Escobal, J., J. Saavedra, M. Torero, “Los Activos de los Pobres en el Peru,” Group of Analysis for Development, working paper no. 26, 1998.

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

Goldsmith, A., D. Hamilton and W. Darity, “Shades of Discrimination: Skin Tone and Wages” *American Economic Review, Papers and Proceedings*, 96(2), 242-245, 2006.

Gyimah-Brempong, K. and G. Price, “Crime and Punishment: And Skin Hue Too?,” *American Economic Review Papers and Proceedings*, 96 (2), 246-250, 2006.

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

Harrison, G. and J. List. "Field Experiments," *Journal of Economic Literature*, XLII (December): 1013-1059, 2004.

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

Hill, M.E., “Skin Color and the Perception of Attractiveness among African Americans: Does Gender Make a Difference?,” *Social Psychology Quarterly*, 65(1),77-91, 2002.

Hunter, M.L., “If You're Light You're Alright: Light Skin Color as Social Capital for Women of Color,” *Gender and Society*, 16(2), 175-193, 2002.

Kagel, J. and A. Roth, *Handbook of Experimental Economics*, Princeton University Press, Princeton, NJ, 1995.

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, 203-229, 2001.

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

Levitt, Steven D. and John A. List, “What do Laboratory Experiments Measuring Social Preferences Tell Us About the Real World,” *Journal of Economic Perspectives*, forthcoming, 2007.

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

List, John A., “Friend or Foe? A Natural Experiment of the Prisoner’s Dilemma,” *Review of Economics and Statistics*, 88(3): 463-471, 2006.

List, J. “On the Interpretation of Giving in Dictator Games, *Journal of Political Economy*, 115(3), 482-493, 2007.

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

Mobius, M. and T. Rosenblatt, “Why Beauty Matters,” *American Economic Review*, 96(1), 222-235, 2006.

Moreno, A., H. Nopo, J. Saavedra and M. Torero, “Gender and Racial Discrimination in Hiring: A Pseudo Audit Study for Three Selected Occupations in Metropolitan Lima,” IZA Working Paper No. 979, 2004.

Ñopo, H., J. Saavedra and T., Maximo. “Ethnicity and Earnings in a Mixed-Race Labor Market,” *Economic Development and Cultural Change*, 55(4), 709-734, 2007.

Oaxaca, Ronald. “Male-female Wage Differentials in Urban Labor Markets.” *International Economic Review*, 14(3), 693-709, 1973.

Palfrey, T. and J. Prisbey, Anomalous Behavior in Public Goods Experiments: How Much and Why?,” *American Economic Review*, 87(5), 829-846, 1997.

Peruvian Truth and Reconciliation Commission, “Final Report,” available at http://www.cverdad.org.pe/ingles/ifinal/conclusiones.php, 2003.

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

Riach, P. and J. Rich, “Field Experiments of Discrimination in the Market Place,” *Economic Journal,* CXII, F480–F518, 2002.

Torero, M., J. Saavedra, H. Nopo and J. Escobal. “An Invisible Wall? The Economics of Social Exclusion in Peru,” *Social Inclusion and Economic Development in Latin America*. Mayra Buvinic, Jaqueline Mazza and Ruthane Deutsch (Eds.) IDB and Johns Hopkins University Press, 2004.

Walker, C., *Smoldering Ashes: Cuzco and the Creation of Republican Peru, 1780-1840*, Durham: Duke University Press, 1999.

Appendix

Fixed-Effects Regression on Ranking (1=Highest, 19=Lowest)		
Variables	**Contribution & Photo**	**Two Types**
Age (years)	0.03 (0.30)	-0.02 (0.69)
Male	0.03 (0.90)	0.10 (0.76)
White	-0.19 (0.59)	0.16 (0.78)
Indigenous	-0.14 (0.70)	-0.32 (0.52)
Attractive	0.58 (0.47)	0.83 (0.39)
Unattractive	0.99 (0.23)	1.21 (0.18)
Expected Rank	0.85 (0.00)	0.72 (0.00)
Expected Rank*Attractive	-0.05 (0.52)	-0.16 (0.04)
Expected Rank*Unattractive	-0.03 (0.70)	-0.06 (0.47)
Constant	0.34 (0.76)	3.20 (0.01)
Individual Fixed Effects	*Yes*	*Yes*
Within-R2	0.71	0.45
N	760	760
p-values in parentheses		