Are appearances deceiving? The nature and evolution of the beauty premium[*]

Tatyana Deryugina, University of Illinois[+]

Olga Shurchkov, Wellesley College[++]

PRELIMINARY AND INCOMPLETE. PLEASE DO NOT CITE OR CIRCULATE.

Abstract:

We design a series of laboratory experiments to illuminate channels through which more attractive individuals receive higher wages. Specifically, we are able to distinguish taste-based discrimination from rational statistical discrimination and biased beliefs. Using three realistic worker tasks to increase the external validity of our results, we find that the "beauty premium" is highly task-specific: While more attractive workers receive higher wage bids in a bargaining task, there is no such premium in an analytical task and a data entry task. The premium in the bargaining task is driven both by biased beliefs about worker performance and taste-based discrimination. Unlike previous experiments on the beauty premium in the labor market, we run two rounds of task performance, in some cases allowing the second-round task to change and in some cases keeping it the same. We find that there is substantial learning between rounds, highlighting the importance of longer-run interactions for a deeper understanding of the mechanism behind the beauty premium.

[*] We are grateful to Kristin Butcher, Phillip Levine, Julian Reif, Casey Rothschild, Kartini Shastri, Akila Weerapana and the participants of Wellesley Economics Department work-in-progress seminar for valuable comments and discussion. In addition, we would like to acknowledge the generous research support we received from the Harvard Decision Science Laboratory to organize and run our experiment. Shurchkov would like to thank the Harvard Kennedy School's Women and Public Policy Program fellowship for providing the resources that made this research possible. Lizi Chen, Divya Gopinath, Mehrnoush Shahhosseini, and Sizhe Zhang provided excellent research assistance. All remaining errors are our own.

[+] Department of Finance, University of Illinois at Urbana-Champaign, 1206 South Sixth St., Champaign, IL, USA.
e-mail: deryugin@illinois.edu.

[++] Department of Economics, Wellesley College, 106 Central St., Wellesley, MA, USA.
e-mail: olga.shurchkov@wellesley.edu.

I. Introduction

Labor market discrimination based on characteristics such as gender, age, race, and national origin is illegal. Appearance-based discrimination, while not currently unlawful, has been the subject of several lawsuits in recent years.[1] But are there efficiency reasons to discriminate based on looks or is there simply a bias against less attractive people?

A number of studies have found that people who are rated as being more attractive are paid more, even when the situation does not appear to warrant it (Hamermesh and Biddle, 1994; Hamermesh and Biddle, 1998). This has been termed the "beauty premium".[2] One potential explanation for its existence is that appearance may in fact be positively correlated with certain skills that are important for job performance but are not observed directly and thus cannot be controlled for in an empirical study (for example, the ability to be persuasive). Another explanation may be employers holding biased beliefs about the skills of more attractive people. Finally, employers may have unbiased beliefs about performance but have a preference for hiring more attractive people, which is known in the literature as "taste-based discrimination". The policy implications of these explanations are very different. But because most studies are observational rather than experimental, the existing evidence does not allow for these three channels to be separately identified.

We perform a series of computer-based laboratory experiments that allow us to distinguish these three causes. The seminal laboratory study on the beauty premium (Mobius and Rosenblat, 2006) focuses on a task that has no appearance-based performance difference. We, on the other hand, allow for the possibility that appearance may be predictive of performance, either directly or through a correlation with some other, unobserved, characteristic. The three tasks we choose are also more realistic than those previously studied, thereby increasing the external validity of our results.

In the experiment, we randomly assign "workers" one of three tasks: A bargaining task, a data entry task, and an analytical task. In the bargaining task, workers see pictures of their bargaining opponent, so we may expect to find positive returns to attractiveness (beauty premium). In the simple data entry task, on the other hand, we do not expect to see positive returns. Finally, in the more difficult analytical task, we may find a beauty penalty. We are careful to keep the rest of the experimental procedure identical across tasks to ensure that we are not picking up on other differences across the treatments.

We test the extent to which employers *correctly* predict the relationship between appearance and performance in various tasks, thus determining what share of the beauty premium is statistical discrimination. In addition to performance predictions, we separately elicit wage bids. Thus, we are able to estimate the part of the beauty premium that is not driven by performance expectations. Finally, because we observe workers' actual performance,

[1] See for example MSNBC "Librarian sues Harvard of discrimination," March 21, 2005.

[2] The importance of beauty has been studied in contexts other than the labor market. See Ravina (2009) for the beauty premium in credit markets; Berri et al. (2011) for the beauty premium in professional sports; Andreoni and Petrie (2008) for the beauty premium in public goods games; Wilson and Eckel (2006) for the beauty premium in trust games; Berggren et al. (2007) for the beauty premium in electoral outcomes.

we also test for a correlation between performance predictions and worker appearance, thus identifying any biased beliefs about the skills of more attractive people.

When we pool data from all three tasks, we find a significant beauty premium in the first round of the experiment, even after we control for worker resume characteristics and date and task fixed effects. The beauty premium is significant only in the sessions where the resume information was shown to the employer first and the employer had to click to see the worker's picture. In the pooled data, controlling for the employer prediction does not fully eliminate the beauty premium, although it slightly decreases the point estimate, suggesting that taste-based discrimination plays a role in the beauty premium.

Varying the tasks while holding constant every other detail of the experiment helps us in several ways. Our study is the first to credibly test whether the beauty premium varies with the type of skills involved in completing the task. Because the tasks we choose are realistic, this is likely to be similar to the variation in the beauty premium by occupation category. Second, varying the tasks helps us corroborate our estimate of the taste-based component of the beauty premium, which should be approximately the same across tasks. On the other hand, biased beliefs about performance may vary with the task. Unlike statistical or taste-based discrimination, discrimination based on biased beliefs is never efficient. Estimating the variation in the bias is thus useful for inferring which occupation categories have the largest deadweight loss from discrimination associated with them.

When we consider the first rounds of the three tasks separately, we find that the beauty premium is only statistically significant in the bargaining task. As before, the positive relationship between beauty and the wage offer appears only when the employer sees the worker's resume information first. Once we control for the employer's performance prediction, the beauty premium in bargaining is only marginally significant. The explanation for this result comes from the positive relationship between beauty and employer's performance prediction in the bargaining task, which does not appear in the other two tasks. The belief that more attractive workers would perform better in bargaining turns out to be incorrect: There is no significant positive relationship between the attractiveness rating and performance in any of the tasks, including bargaining.[3]

It is also not known how the beauty premium for a particular worker evolves over time. It is possible that attractiveness is used as a proxy for ability initially but becomes irrelevant as employers observe workers' actual performance. To test for the existence of this type of learning, we reveal workers' performance in the tasks to all employers. We then repeat the bidding process, performance prediction, and task performance and allow employers to update their bids and expectations. Thus, we can estimate what fraction of the beauty premium disappears once performance measures are available. In some of the experimental sessions, the task stays the same, while in others we randomly choose a different task for the second round.

We find that learning completely eliminates the beauty premium and more attractive workers do not receive higher wage bids in any of the tasks in the second round. In fact, we

[3] This finding is consistent with the recent empirical evidence that shows a negative correlation between attractiveness and performance on SATs and quantitative reasoning tests and no relationship between overall GPA and attractiveness, on average, at the undergraduate level (Deryugina and Shurchkov, 2012).

observe a beauty penalty in the second round bargaining task, although this effect only appears when the first round task was not a bargaining task. Once we control for the performance in round one, however, this "beauty penalty" disappears. We conclude that learning about past performance tends to reduce the effects of beauty on the wage bid as it changes employers' expectations of future worker performance. This evidence is consistent with statistical discrimination in the first round which is subsequently eliminated through learning and updating of beliefs.

Our key insight is that having measures of (a) expectations about the worker's performance and (b) the willingness to hire the worker are helpful for identifying different kinds of discrimination. This idea can be applied in other settings and to other types of discrimination. For example, asking car dealers how much money they expect to make on a particular sale *and* observing their actual bargaining behavior can help determine whether offers made to women and minorities are driven by statistical or taste-based discrimination, something an earlier study is not able to do conclusively (Ayres and Siegelman, 1995). Castillo et al. (2012) conduct a field study on gender differences in bargaining outcomes over taxi fares in Peru and find evidence consistent with both statistical and taste-based discrimination.

Our study demonstrates that, even in the laboratory, the beauty premium can be highly variable and depends on the exact labor market setting. This is consistent with previous literature, which posits that more attractive individuals sort into occupations that command a higher beauty premium, all else equal (Hamermesh and Biddle, 1994; Hamermesh and Biddle, 1998; von Bose, 2012).

The rest of the paper is organized as follows. In Section II, we present an overview of our experimental procedures. Section III outlines the framework that allows us to separately identify biased beliefs about performance, statistical discrimination, and taste-based discrimination. Section IV reports and discusses the results, and Section V concludes.

II. Overview of the Experiment

A. The Stylized Labor Market

The labor market portion of the experiment was conducted at the Decision Science Laboratory at Harvard University. Subjects were undergraduate and graduate students from Harvard and other Boston area universities. Each session included four employers and four workers. [4] All subjects started by having their photograph taken and answering survey questions about their basic labor market characteristics (student status, major, GPA, and typing, analytical, and communication skill levels). Employers were then granted access to a website that displayed worker photographs and the corresponding "resumes" based on each worker's answers to the survey questions. In 25 sessions, photos were shown on the front webpage with links to resume information underneath each photo, and in 22 sessions, this order was

[4] The first four subjects to arrive at the laboratory and sign the consent form (our "employers") were immediately taken from the waiting room, photographed, and seated at their stations. The next four subjects to arrive at the laboratory and sign a consent form (our "workers") were photographed and seated afterwards. In order to avoid further face-to-face interactions between the two groups, employers and workers were seated at stations separated by a wall divider.

reversed. Throughout the session, we were able to track clicking patterns of the employers to establish which workers received a disproportionate amount of employer attention.

The remainder of the experiment, programmed using the standard software package, zTree (Fischbacher 2007), consisted of two procedurally identical rounds. Each round started with a prediction stage during which employers submitted estimates for the expected performance of all workers in the subsequent task (E_{ij}, where i indexes employers and j indexes workers), and workers submitted estimates for their own expected performance (E_j). This information was kept secret from all other subjects.

Next, employers submitted wage offers to "hire" workers. The total amount offered to four workers could not exceed a predetermined maximum number of points.[5] The employer with the highest wage offer for a particular worker "hired" that person and had to pay the worker the second highest wage (W_j) offered to that worker. Each employer could be matched to none or all four workers, depending on the wage offers. A worker could be left unmatched if all four employers offered a zero wage to that worker, although this did not happen in practice. The wage amount (if any) was not revealed to the worker until after the task completion stage. The identity of the employer was never revealed to the worker. Employers had full knowledge about the task workers would have to perform.

The task completion stage began after the employer-worker matching was established. The tasks included: A bargaining task, a data entry task, and a data analysis task (see detailed task descriptions below). Table 1 records the number of sessions for each task type and the corresponding number of subjects that participated in a given session.

Each round ended with an information screen. Employers learned about the performance of every worker and their own payoff for the round. Workers learned about their own performance and payoff for the round, including any wage payment. The following equations represent the total within-round payoffs.[6]

Employer i's Payoff:

$$\pi_i = 125 + \frac{1}{3}\sum_{j=1}^{4} P_t Y_j \times Hire_{i,j} - \sum_{j=1}^{4} W_j \times Hire_{i,j} - M_t \sum_{j=1}^{4} |Y_j - E_{i,j}|$$

Worker j's Payoff:

$$\pi_j = 25 + \frac{2}{3} P_t Y_j + W_j \times Hire_{i,j} - M_t |Y_j - E_j|$$

where $i \in \{1,4\}$ is the set of employers, $j \in \{1,4\}$ is the set of workers, and $t \in \{Data\ Entry, Data\ Analysis, Bargaining\}$ is the set of tasks; P_t is the piece rate of 5 points for t = *Data Analysis* and 1 point for the other tasks; M_t is the weight on the deviation of the performance estimate from actual output and equals $\frac{5}{4}$ for t = *Data Analysis* and $\frac{1}{4}$

[5] In 22 earlier sessions, this amount equaled the employer's endowment of 125 points, while in the 25 subsequent sessions this amount was raised to 175 points while the endowment remained at 125 points. The reason for the increase was to allow employers to base their bid on their estimate of the expected performance by the worker, rather than the mechanical constraint imposed by the bid maximum.

[6] The last term in both equations represents a "misprediction penalty" which we include in order to incentivize truth-telling in accordance with previous literature (Mobius and Rosenblat, 2006).

otherwise; and $Hire_{i,j}$ is an indicator function that takes on the value of 1 for all workers j hired by employer i , and 0 otherwise.

Table 1: Treatment Summary

	Data Entry		Data Analysis		Bargaining	
	# Sessions	#Subjects	# Sessions	#Subjects	# Sessions	#Subjects
Round 1	16	128	15	120	16	128
Round 2	15	120	16	128	16	128
Round 1 = 2	7	56	8	64	8	64

Round 1 and Round 2 rows list all sessions, regardless of whether the task was the same or different in the second round. Sessions with the same task in both rounds are listed in row "Round 1=2".

At the end of the session, all subjects filled out a post-experiment questionnaire that asked for detailed demographic information. In addition, employers were asked whether they were acquainted with any of the workers.[7]

A. The Tasks

Data Entry

In the data entry task, workers had 6 minutes to enter as much numerical data as possible from a paper copy into an Excel spreadsheet. The data consisted of various economic statistics for regions in Russia. The spreadsheets had been opened on the workers' computers prior to the start of the experiment with the column and row headings prepared in advance, so that subjects only had to enter numerical values into the individual cells. The data had to be entered exactly as it appeared to get credit. Workers were credited with 1 point per correctly entered item. There was no penalty for an incorrectly entered item.

Data Analysis

In the data analysis task, workers had 12 minutes to answer as many mathematical questions out of 30 as possible. Questions were based on data similar to that used in the data entry task. Workers had 6 minutes for the first 15 questions and 6 minutes for the second 15 questions. Because questions required mathematical calculations, workers could use calculators that were placed on their desks in advance. Workers were credited with 5 points per correctly answered question, and there was no penalty for answering questions incorrectly.

Bargaining

In the bargaining task, workers were randomly assigned to be a buyer or a seller of a "widget" and participated in three 90-second periods of a standard double-auction. Workers

[7] Instructions for the experiment and questionnaire contents available upon request.

were randomly re-matched and roles were randomly assigned every bargaining period. Each worker saw a photo of his or her bargaining partner on the decision screen. Every time a transaction was made, the seller's profit equaled the difference between the price and the seller's true cost of the "widget," and the buyer's profit equaled the difference between the buyer's true value and the price of the "widget." Profits were calculated in tokens. If the time ran out before a transaction was made, both the buyer and the seller earned 0 tokens in that bargaining period. Each token was equivalent to one point for the purposes of calculating the total payoff for the round. Buyers' values and sellers' costs were determined randomly from two uniform distributions. In some cases, the buyer's value was below the seller's cost, making agreement impossible.[8]

B. The Rating Procedures

The rating portion of the experiment was conducted at the University of Illinois, Urbana-Champaign (UIUC). Subjects were undergraduate and graduate students from UIUC. Each session consisted of 4-15 subjects (raters) who were instructed to view and evaluate photos on a scale from 1 (extremely homely) to 10 (strikingly handsome or beautiful).[9] Each rater was asked to go through 4 sets of 100 photos,[10] which appeared in random order within each photo set. Due to the large number of photos, each rater evaluated only a subset of photos. The individual rating variable used in subsequent analysis is demeaned by the rater's average across the photos that appeared in a given photo set.[11]

Each rating session lasted between forty minutes and one hour, including the reading of the instructions and payment. Raters were paid a show-up fee of $5 and an additional $7 payment for completing the task of rating all photos and providing demographic information. Table 2 summarizes our rater pool.

---

[8] To avoid the possibility of negative profits, sellers could not agree to an offer that was lower than their cost and buyers could not agree to an offer that was higher than their value.

[9] The scale was expanded from a 1-5 point scale previously used in the literature (Hamermesh and Biddle 1994, 1998) to a 1-10 point scale.

[10] The photos were not limited to the laboratory subjects from the current experiment. Some photosets contained photos that were rated as part of another study (Deryugina and Shurchkov, 2012). However, photos from different studies were not mixed within a given photo set. We capped the number of photos shown to a given rater at 400 in order to avoid boredom and fatigue effects and to keep the rating session at under one hour.

[11] In other words, our regression analysis controls for the rater and photo set fixed effects.

Table 2: Number of raters in each gender and race category

	Male	Female
White	30	39
Asian	32	32
Black	5	8
Hispanic	7	8
Native American	0	0
Pacific Islander	0	3
Total	66	80

Notes: The sum of the number of raters by each race category exceeds the total number of males in column 1 and the total number of females in column 2 because some raters identified with more than one race.

III. Empirical Strategy

A. *Separating taste-based and statistical discrimination*

In this section, we outline a simple framework that will help understand how we can separate statistical discrimination from taste-based discrimination and test whether statistical discrimination, if any, is based on rational or biased beliefs.

Assume that workers are indexed by their productivity, θ_j, and their attractiveness, α_j. An employer gets utility $U(\theta_j, \alpha_j)$ from hiring worker j. We assume that the marginal utility of hiring a more productive worker is positive, that is $\frac{\partial U(\theta_j, \alpha_j)}{\partial \theta_j} > 0$. This is true in our setting, where employers who hire more productive workers make more money. If the employer also gets direct utility from hiring a more attractive worker, then $\frac{\partial U(\theta_j, \alpha_j)}{\partial \alpha_j} > 0$.

Recall that we ask each employer for her beliefs about the worker's productivity, $E_i[\theta_j]$, and for her willingness to hire a particular worker, as measured by the wage bid, W_{ij}.[12] If employers are maximizing utility, then, on average:

(1) The wage bid, W_{ij}, should increase with performance expectations, $E_i[\theta_j]$.
(2) If the employer gets direct utility from hiring more attractive workers ($\frac{\partial U(\theta_j, \alpha_j)}{\partial \alpha_j} > 0$), W_{ij} should increase with the worker's attractiveness, α_j, holding performance expectations constant.

Note that statement (2) is conditional on performance expectations. Simply seeing that employers bid more on more attractive workers does not imply that employers get direct utility

[12] Both are elicited in an incentive-compatible way, as described in the previous section.

from hiring more attractive workers because θ_j and α_j may be correlated.[13] In this case, if an employer cannot observe θ_j perfectly, she may view α_j as informative about θ_j and bid more on attractive workers even if $\frac{\partial U(\theta_j,\alpha_j)}{\partial \alpha_j} = 0$. We refer to this situation as "statistical discrimination": the employer does not value hiring a more attractive worker per se, but uses his appearance as a signal of ability, perhaps incorrectly. But employers will also place a higher value on more attractive workers if $\frac{\partial U(\theta_j,\alpha_j)}{\partial \alpha_j} > 0$. We refer to this case as "taste-based discrimination" because it is driven by direct preferences for attractive workers. Separating these two types of discrimination requires estimating the share of the beauty premium, if any, that results from statistical discrimination and taste-based discrimination.
A necessary condition for statistical discrimination is that employers believe more attractive workers are also more productive.[14] Because we observe each employer's performance expectations, we can test for this directly. If the correlation is significantly different from 0, we can further test if these beliefs are correct on average by testing whether actual performance θ_j and attractiveness α_j are themselves correlated. If there is no residual correlation between attractiveness and performance once the employer's prediction is controlled for, this suggests that employers have correct expectations about the performance of more attractive workers. If employers overestimate (underestimate) the ability of more attractive workers, then the relationship between attractiveness and performance should be negative (positive) once the employer's expectations are controlled for.

To separate taste-based discrimination from statistical discrimination, we use the fact that the performance prediction captures the employer's beliefs about θ_j, while the bid captures the employer's value from θ_j *and* from α_j. Thus, if only statistical discrimination is present (regardless of whether beliefs about performance are correct), then any effect of attractiveness on wage bids should operate only through the performance expectation, $E_i[\theta_j]$. In other words, once we properly control for $E_i[\theta_j]$, the worker's attractiveness, α_j, should have no further explanatory power in the case of pure statistical discrimination. If the effect of α_j on the employer's wage bid is significant after $E_i[\theta_j]$ is controlled for, we conclude that there is taste-based discrimination, where employers bid more on more attractive workers even though they do not expect them to be more productive.

Specifically, we use the following regression specification:

$$\log(1 + W_{ij}) = \beta_1 a_j + \gamma_1 E_i[\theta_j] + X_j' \rho + \delta_T + \varepsilon_{ij} \tag{1}$$

i indexes the employer and j indexes the worker. W_{ij} is the bid of employer i on worker j.[15] X_j' is a vector of worker characteristics consisting of indicators for student status (graduate or

[13] This could either be due to genetics or simply because attractiveness is driven by general ability. For example, someone who is responsible on the job may also take better care of their appearance thus seeming more attractive.

[14] This is the measure of the beauty premium used by Mobius and Rosenblat (2006).

[15] The logarithmic specification corrects for skewness and outliers in our wage bid data.

undergraduate), major, GPA range, self-reported abilities, race, and gender. β_1 captures taste-based discrimination, while γ captures statistical discrimination, both of which may be present. α_T is a set of task fixed effects that we include whenever we combine multiple tasks in a single regression.

Finally, we are also able to examine the evolution of the beauty premium over time. If attractiveness a_j is used as a signal of ability θ_j, it should be less informative in the second round, after a worker's actual performance is revealed. This will not necessarily hold in treatments where the first and second round tasks are different because θ_j may be task-specific. However, in treatments where the first and second round tasks are the same, we expect the statistical discrimination component to be weaker in the second round of the experiment. To test for this, we estimate the relationship between (a) employers' second round bids and (b) workers' attractiveness interacted with an indicator for whether the first and second round tasks are the same.

$$\log(1 + W_{ij2}) = \beta_2 a_j + \beta_3 a_j 1[Same] + \gamma_2 \theta_{j1} + X_j' \rho + \delta_T + \varepsilon_{ij2} \tag{2}$$

where subscripts 1 and 2 correspond to rounds 1 and 2, respectively. $1[Same]$ is an indicator equal to 1 if the tasks in round 1 and round 2 are the same. Because the taste-based component should stay the same, any changes in this relationship can be attributed to a change in the statistical discrimination component.

B. Bidding mechanism

We employ a sealed-bid second price auction mechanism to match workers and employers. Conceptually, hiring a worker has a common value and a private value component. The common value component consists of the worker's ability to earn money for the employer, while the private value component may consist of the employer's utility from hiring a worker with particular characteristics, including his or her appearance.

Given the complicated nature of the bidding mechanism, we do not derive an analytical solution for the rational bidding strategy. In fact, our baseline regression specification (1) above makes the strong assumption that the employer's bid for the worker is linear in her performance prediction. Previous experimental literature (for example, Cooper and Fang, 2008) finds that behavior consistently deviates from purely rational bidding strategies. Thus, even if we could solve for the optimal bidding strategy, we could not necessarily assume that subjects followed it.

To address these concerns, we develop tests that control for the performance prediction in a more flexible manner.

$$W_{ij} = \beta a_j + \gamma_1 E_i[\theta_j] + \gamma_2 E_i[\theta_j]^2 + X_j' \theta + \alpha_T + \varepsilon_{ij}$$

Our experimental setting allows us to learn even more about the beauty premium. Specifically, if the *direct* utility from hiring an attractive worker is independent from the level of the worker's skill and the task the worker has to perform, we can make the prediction that the taste-based discrimination component, if any, should be constant across the tasks. The same should be true of the statistical discrimination component, once we adjust for the different payment rates across tasks.

IV. Results

A. Stylized facts and correlations

A1. Laboratory Labor Market Summary Statistics

We start out with a total of 376 subjects split evenly between employers and workers (752 employer-worker pairs – our unit of observation). However, a few subjects drop out from our sample. First, we exclude two pilot sessions held on December 07, 2011 from the main analysis. Second, we drop a subject who self-identified as not being a student (an employer). Third, we drop a subject who appeared in our pool twice, keeping the first observation (an employer) and dropping the second observation (a worker). Fourth, we drop employers who did not use the worker resume information (that is, who did not click on the worker's photo or resume). Including these observations does not significantly change the results. Finally, we also drop two subjects (workers) who chose to withdraw from our study after the experiment. The final dataset consists of 174 employers and 177 workers for a total of 685 employer-worker pairs.

Table 3 provides the descriptive statistics. In Round 1, the average wage bid ranges from 24 to 33. Male employers bid slightly higher than females in the bargaining task but not in the other two tasks. However, there are no significant gender differences in employers' performance predictions in any of the tasks, suggesting that males are bidding higher in the bargaining task for reasons other than different beliefs about worker performance.

In Round 1, male workers predict that they will perform better than female workers in both the analytical and bargaining tasks. However, these predictions turn out to be incorrect in the data analysis task, where both males and females perform equally well. In the bargaining task, males outperform females by about 24 points. Females outperform males in the data entry task.

In Round 2, the pattern of worker performance predictions and actual performance is reversed. There is no significant difference between male and female performance in data entry or bargaining, but there is a significant difference between males and females in data analysis, with males outperforming females by about 2.5 questions. However, we see no significant differences in bidding behavior or performance prediction for male and female employers.

Table 3: Averages by task, gender, and round

	Data Entry		Data Analysis		Bargaining	
	Male	Female	Male	Female	Male	Female
Round 1:						
Employer Wage Bid	26.3	27.9	24.0	25.3	32.6*	25.1*
Employer Prediction	67.3	68.6	10.7	11.5	62.8	63.2
Worker Prediction	78.4	86.4	11.2**	9.3**	68.8**	60.2**
Worker Performance	74.1**	88.3**	7.2	7.1	79.9**	55.5**
# of Employer-Worker Pairs	67	181	129	80	103	125
Round 2:						
Employer Wage Bid	28.0	32.9	23.9	22.1	32.8	26.7
Employer Prediction	79.7	82.3	9.0	9.3	69.1	64.1
Worker Prediction	106.5*	93.4*	10.2	9.8	69.5	66.4
Worker Performance	86.0	84.9	11.1**	8.5**	69.2	64.0
# of Employer-Worker Pairs	86	139	140	96	73	151

Notes: Significance levels based on t-tests of differences between male and female subjects: * 5 percent, ** 1 percent. Mann-Whitney U test that compares distributions produces similar results, with the exception of worker prediction in data entry in round 2 which no longer shows a significant gender difference.

B. Evidence of statistical discrimination in the first round

We begin our analysis by focusing on the first round. Because our theoretical framework predicts that employer expectations should play a role in the relationship between attractiveness and wage offers, we first ask whether employers believe that more attractive workers are also more productive in the various tasks.

Result 1: Employers expect more attractive workers to be more productive in the bargaining task, but not in other tasks. These beliefs turn out to be incorrect.

Table 4 estimates the relationship between an employer's performance expectations for a given worker and that worker's attractiveness. In all specifications, we regress the employer's prediction of worker performance in round 1 on the beauty rating, controlling for date fixed effects, worker resume characteristics, and clustering standard errors by employer. When we pool the data across all three tasks (Specifications 1 and 2), we find that employers do not expect more attractive workers to perform better, on average. The effect of beauty on employer's performance prediction is not statistically significant whether or not we control for photo order effects[16]. The rest of the specifications break up the analysis by task. As we anticipated, employers do not expect more attractive workers to have a performance advantage in the data entry task (Specifications 3 and 4). We also do not find a significant relationship between the worker's attractiveness and expected performance in the data analysis task in the pooled data (Specifications 5). In fact, when we control for worker characteristics and break the sample up by worker information order in Specification 6, we find a negative effect of beauty on the wage bid in the data analysis task, although it appears only when the photo of the worker is shown first. A robust positive relationship between beauty and expected performance emerges in bargaining (Specifications 7 and 8). When the resume appears on the employers' information screen first, a one standard deviation increase in attractiveness is associated with a 4.8 point increase in predicted performance. When the photo appears on the employers' information screen first, a one standard deviation increase in attractiveness is associated with a 2.8 point increase in predicted performance, although the effect is only marginally significant.[17] This finding is consistent with the nature of the task being "beauty-related," since workers can see their opponent's photo during bargaining.

Table 4: Relationship between an employer's performance expectations and a worker's attractiveness in round 1

Outcome variable:	Employer's prediction of worker performance in round 1							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	All Tasks		Data Entry		Data Analysis		Bargaining	
Attractiveness of worker	0.871		-2.492		-0.423		3.430**	
	(0.704)		(2.397)		(0.337)		(1.080)	
Attractiveness of worker & Resume shown first		1.262		-3.454		0.379		4.480*
		(0.984)		(4.266)		(0.416)		(2.053)
Attractiveness of worker & Photo shown first		0.572		-1.841		-1.160*		2.759^{+}
		(1.052)		(2.308)		(0.513)		(1.534)
Worker characteristics	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	685	685	248	248	209	209	228	228
R-squared	0.74	0.74	0.45	0.46	0.26	0.27	0.27	0.27

Notes: Round 1 data only. All regressions include date fixed effects, indicators for student status (graduate or undergraduate), major, GPA range, self-reported abilities, race, and gender. In cases where we pool the data across multiple tasks, we also include a task fixed effect. Robust standard errors clustered by employer in parentheses. Significance levels: $^{+}$10 percent, * 5 percent, ** 1 percent

[16] Worker resume characteristics in this and other regressions include indicators for student status (graduate or undergraduate), major, self-reported abilities (typing, analytical, and communications), race, and gender. Photo order effects include an indicator for the sessions where the photo appeared first on the employer's information screen, as well as an interaction term between this indicator and the worker beauty rating.

[17] When we include employer fixed effects in the specifications, we find a significant positive correlation between attractiveness and employer prediction of worker performance in the pooled data. The effect is significant whether or not the worker photo appears first on the information screen. The task-specific results are qualitatively similar. For space reasons, we do not show these specifications. Results are available upon request.

Except for the data analysis task, these beliefs turn out to be incorrect, on average. Table 5 documents that there is no systematic positive relationship between attractiveness and performance in our experiment. In all specifications, we regress performance in round 1 on the beauty rating, controlling for date fixed effects and clustering standard errors by worker. Specifications 1 and 2 pool the data across tasks and include task fixed effects. Whether or not we control for worker characteristics, there is no significant relationship between the beauty rating and performance. Specifications 3 and 4 restrict the sample to include only sessions where the task in round 1 was data entry. The effect of beauty is not statistically significant in Specification 3 and becomes significantly negative when we control for subject characteristics in Specification 4. Specifications 5 – 6 and 7 – 8 show that there is no significant correlation between beauty and performance in data analysis and bargaining, respectively.

Table 5: Relationship between a worker's attractiveness and her performance in round 1

Outcome variable:	Worker performance							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	All Tasks		Data Entry		Data Analysis		Bargaining	
Attractiveness of worker	-0.237	-0.515	-6.654	-15.655*	0.168	-0.007	0.87	-0.844
	(3.193)	(3.231)	(6.535)	(7.304)	(0.331)	(0.328)	(6.318)	(6.424)
Worker characteristics	No	Yes	No	Yes	No	Yes	No	Yes
Observations	685	685	248	248	209	209	228	228
R-squared	0.55	0.60	0.30	0.57	0.35	0.61	0.15	0.43

Notes: Round 1 data only. All regressions include date fixed effects, indicators for student status (graduate or undergraduate), major, GPA range, self-reported abilities, race, and gender. In cases where we pool the data across multiple tasks, we also include a task fixed effect. Robust standard errors clustered by worker in parentheses. Significance levels: +10 percent, * 5 percent, ** 1 percent

Next, we turn to our first main result, which estimates the relationship between the first round wage bid of a potential employer and the attractiveness of the worker, conditional on the task and worker characteristics.

Result 2: Overall, there is a beauty premium in the bargaining task, but not in other tasks.

Support for Result 2 comes from Table 6, which shows an overall positive correlation between the natural logarithm of the wage bid in round 1 and worker attractiveness. All specifications in Table 6 include worker resume characteristics and date fixed effects with standard errors clustered by employer. Specifications 1 and 2 pool the data across tasks and include task fixed effects. Furthermore, Specification 2 includes an additional control for whether the photo or the resume appeared on the front page of the worker information website (the photo first dummy), as well as an interaction between photo first and worker beauty rating. The positive and statistically significant coefficient on the beauty rating indicates that more attractive workers in the sessions with worker resume shown first receive higher wage offers, on average. More attractive workers whose photos are shown to employers first also receive higher wage offers, but the magnitude of the effect is smaller.

When we further decompose our analysis by task, we observe that the beauty premium does not appear in all settings. In particular, we do not find any statistically significant effect of beauty on wage bids in either the data entry or the data analysis task (Specifications 3 and 5). The inclusion of the photo first dummy and its interaction with the beauty rating in Specifications 4 and 6 does not change the results. The beauty premium is statistically significant only in the bargaining task (Specification 7). When we interact worker information order with attractiveness, we observe that the beauty premium is only statistically significant when worker resume information was the first to appear on the employer's information screen.[18] The fact that the beauty premium shows up precisely in the task that we perceived as "beauty related" but not in the tasks that we perceived as "beauty unrelated" suggests that employer expectations play a role in explaining the existence of the overall beauty premium.

Table 6: Relationship between an employer's bid in round 1 and a worker's attractiveness

Outcome variable:	Natural Logarithm of employer's wage bid in round 1							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	All Tasks		Data Entry		Data Analysis		Bargaining	
Attractiveness of worker	0.165*		-0.003		0.154		0.343*	
	(0.074)		(0.139)		(0.157)		(0.137)	
Attractiveness of worker &		0.272*		-0.069		0.388		0.550*
Resume shown first		(0.110)		(0.219)		(0.239)		(0.225)
Attractiveness of worker &		0.083		0.041		-0.061		0.21
Photo shown first		(0.094)		(0.173)		(0.202)		(0.167)
Worker characteristics	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	685	685	248	248	209	209	228	228
R-squared	0.16	0.16	0.21	0.21	0.26	0.27	0.20	0.21

Notes: Round 1 data only. All regressions include date fixed effects, indicators for student status (graduate or undergraduate), major, GPA range, self-reported abilities, race, and gender. In cases where we pool the data across multiple tasks, we also include a task fixed effect. Robust standard errors clustered by employer in parentheses. Significance levels: $^{+}$10 percent, * 5 percent, ** 1 percent

The positive correlation between wage bids and attractiveness does not tell us whether the premium is driven by statistical or taste-based discrimination. In order to separate these two channels, we explicitly include the employer's belief about performance as one of the determinants of the wage bid in round 1.

[18] The results are robust to including employer fixed effects in the regression specifications. The beauty premium in bargaining remains statistically significant without interacting the beauty rating with worker information order. Since we lose power due to the inclusion of employer fixed effects, the beauty premium disappears when we interact beauty with photo order, although the magnitude of the coefficient remains unchanged. For space reasons, we do not show these specifications. Results are available upon request.

Result 3: The beauty premium is at least partly driven by employer expectations.

Table 7 estimates the effect of the beauty rating on wage bids in round 1, controlling for the employer prediction of worker performance. [19] In specifications 1-4, the effect of the employer's prediction on log wage bid is positive and statistically significant at the 1 percent level. Overall, the inclusion of the prediction does not eliminate the beauty premium completely, although the magnitude of the effect decreases as a result (Specification 1). In data entry and in data analysis, beauty remains an unimportant determinant of the wage bid (Specifications 2 and 3). In the bargaining task, the inclusion of beliefs decreases the magnitude of the beauty premium, although in remains marginally statistically significant (Specification 4). We conclude that, at least in the bargaining task, beauty affects employer expectations of worker performance, and these expectations partly explain why more attractive workers receive higher wage offers.

Table 7: Separating statistical from taste-based discrimination

Outcome variable:	Natural logarithm of employer's wage bid in round 1							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	All Tasks	Data Entry	Data Analysis	Bargaining	All Tasks	Data Entry	Data Analysis	Bargaining
Attractiveness of worker & Resume shown first	0.253*	-0.021	-0.04	0.486*	0.238*	-0.038	0.301	0.432^{+}
	(0.105)	(0.198)	(0.195)	(0.219)	(0.105)	(0.193)	(0.228)	(0.220)
Attractiveness of worker & Photo shown first	0.074	0.067	-0.013	0.171	0.068	0.049	-0.018	0.125
	(0.093)	(0.161)	(0.205)	(0.162)	(0.093)	(0.162)	(0.229)	(0.151)
Employer's performance prediction	0.015**	0.014**	0.042*	0.014*	-0.014	-0.026	-0.049	-0.014
	(0.004)	(0.004)	(0.018)	(0.006)	(0.024)	(0.045)	(0.382)	(0.043)
Nonlinear employer prediction	No	No	No	No	Yes	Yes	Yes	Yes
Worker characteristics	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	685	248	209	228	685	248	209	228
R-squared	0.20	0.25	0.29	0.24	0.20	0.25	0.29	0.32

Notes: Round 1 data only. All regressions include indicators for student status (graduate or undergraduate), major, GPA range, self-reported abilities, race, and gender. In cases where we pool the data across multiple tasks, we also include a task fixed effect. Robust standard errors clustered by employer in parentheses. Significance levels: $^{+}$10 percent, * 5 percent, ** 1 percent

The evidence in Specifications 1-4 of Table 7 suggests that there is residual discrimination in favor of more attractive workers in the bargaining task beyond what is explained by expected performance. On the other hand, we do not find any residual preference for more attractive workers in the other two tasks. This is surprising, because, in theory, we would expect to find the same residual component regardless of task, after controlling for the performance prediction.

One possible explanation for the differences in the residual discrimination component is that Specifications 1-4 make the strong assumption that the employer's bid for the worker is linear in her performance prediction (see Section IIIB for detailed discussion of the bidding strategies). In order to test this hypothesis, Specifications 5-8 control for the performance

[19] The rest of the controls are the same as in Table 3, including the photo order controls.

prediction in a more flexible manner by introducing nonlinear employer prediction terms (squared and cubed) into the regression. In the pooled data, the beauty premium remains significant at 5 percent confidence level. However, in bargaining, the effect of beauty, though still positive, becomes only marginally statistically significant.

C. *Does learning about performance affect the beauty premium in the second round?*

So far, the evidence suggests that employers use appearance as a signal about ability, at least in the tasks that might be perceived as favoring more attractive workers. However, we have also shown that the employers' beliefs are incorrect. It is therefore important to ask what happens to the relationship between wage bids and beauty after relevant information about actual previous performance is revealed to the employers. To that end, we estimate the effect of the beauty rating on wage offers in the second round, conditional on first-round performance information. We hypothesize that, because we don't observe a significant relationship between attractiveness and performance, we should observe a reduction in the beauty premium in the second round relative to the first, which would indicate learning. Furthermore, this reduction should be more pronounced in sessions where the tasks were the same in the two rounds, because the performance information from the first round is more informative in this setting.

Result 4: The beauty premium disappears in the second round, once past performance information is taken into account.

Support for Result 4 comes from Table 8. In all specifications we regress log wage bid in the second round on the beauty rating, controlling for whether the task was the same in both rounds and whether the photo or the resume was shown first in that session. We also include a full set of worker resume characteristics and date fixed effects with standard errors clustered by employer.

Specification 1 shows that there is a marginally significant beauty premium when we pool the data across tasks in the second round, but only when the resume was shown first and the task was the same in the second round. Once we control for worker performance in the first round in specification 2 and for the employer's second round prediction in specification 3, the beauty premium disappears in the pooled sample. We do not find a significant effect of the beauty rating in either data entry or data analysis, with the exception of the marginal positive effect of beauty when the task in the second round is data entry in specification 6. In contrast to round one, where we find a strong beauty premium in the bargaining task, we actually observe a negative effect of attractiveness on the wage bid (a beauty penalty) in the second round of the bargaining task when the first round task was not bargaining. However, the effect disappears once we control for performance in the first round and for the employer prediction in the second round.

Table 8: Determinants of employer's wage bid in round 2

Outcome variable:	Natural Logarithm of employer's wage bid in round 2											
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	All Tasks			Data Entry			Data Analysis			Bargaining		
Attractiveness of worker & same task in round 2 & resume shown first	0.266+	0.107	0.056	0.262	0.294	0.206	0.029	-0.356	-0.396	-0.08	0.08	0.092
	(0.160)	(0.167)	(0.164)	(0.456)	(0.451)	(0.381)	(0.558)	(0.490)	(0.459)	(0.186)	(0.179)	(0.184)
Attractiveness of worker & same task in round 2 & photo shown first	0.159	0.178	0.105	0.314	0.239	-0.083	0.07	-0.042	0.059	0.276	0.183	0.159
	(0.138)	(0.135)	(0.126)	(0.457)	(0.451)	(0.397)	(0.212)	(0.207)	(0.195)	(0.233)	(0.231)	(0.202)
Attractiveness of worker & different task in round 2 & resume shown first	0.035	0.101	0.093	0.31	0.433	0.520+	0.392	0.109	0.026	-0.547+	-0.235	-0.425
	(0.138)	(0.134)	(0.128)	(0.344)	(0.319)	(0.278)	(0.314)	(0.223)	(0.198)	(0.281)	(0.300)	(0.308)
Attractiveness of worker & different task in round 2 & photo shown first	0.044	0.055	0.01	0.178	0.16	0.152	0.213	0.313	0.189	-1.155*	-0.446	-0.649
	(0.152)	(0.151)	(0.147)	(0.333)	(0.337)	(0.328)	(0.210)	(0.216)	(0.178)	(0.457)	(0.461)	(0.420)
Worker's performance in round 1 & same task in round 2		0.016**	0.006+		0.01	-0.009		0.373**	0.281**		0.016**	0.005
		(0.003)	(0.003)		(0.007)	(0.006)		(0.073)	(0.067)		(0.005)	(0.006)
Worker's performance in round 1 & different task in round 2		0.007**	0.003		0.007	0.001		0.014**	0.010**		0.016*	0.003
		(0.002)	(0.003)		(0.006)	(0.006)		(0.004)	(0.003)		(0.008)	(0.008)
Employer's performance prediction in Round 2			0.025**			0.038**			0.133**			0.022**
			(0.004)			(0.008)			(0.032)			(0.006)
Worker characteristics	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	685	685	685	225	225	225	236	236	236	224	224	224
R-squared	0.10	0.18	0.24	0.21	0.22	0.33	0.25	0.38	0.46	0.23	0.28	0.35

Notes: Round 2 data only. All regressions include a dummy variable for the task in second round being the same as in the first round, date fixed effects, indicators for student status (graduate or undergraduate), major, GPA range, self-reported abilities, race, and gender. In cases where we pool the data across multiple tasks, we also include a task fixed effect. Robust standard errors clustered by worker and session in parentheses. Significance levels: +10 percent, * 5 percent, ** 1 percent

Information about a worker's performance in the first round has a significant impact on the wage bid when we pool the data across tasks (Specification 2), as well as in data analysis (Specification 8) and in bargaining (Specification 11). The effect is larger and more statistically significant when the task in the first round was the same as the task in the second round. Performance in the first round does not seem to have a positive effect on the wage bid in data entry (Specification 5). Once we control for the employer's performance prediction and pool across tasks (Specification 3), past information is no longer significant when task in round 2 is different and only marginally significant when the task in round 2 is the same. This suggests that employers fully incorporate first round performance into their subsequent predictions. In data entry, round 1 performance remains unimportant in predicting the wage bid in round 2 (Specification 6). In data analysis (Specification 9), the inclusion of performance prediction in round 2 does not eliminate the significance of past performance information, although the coefficients decrease in magnitude. Finally, once we control for the employer's performance prediction in Specification 12, round one performance information is no longer significant, which implies that learning affects wage offers in bargaining through its effect on employers' beliefs about performance.

We conclude that learning about past performance tends to reduce the effects of beauty on the wage bid as it changes employers' expectations of future worker performance. This evidence is consistent with statistical discrimination in the first round which is subsequently eliminated through learning and updating of beliefs.

D. Robustness tests

D1. Bidding Behavior

One explanation for the effects we are finding might result from strategic bidding by the employers when the maximum bid is binding. The inclusion of the date fixed effects accounts for the possibility that changing the maximum total bid from 125 to 175 in the later sessions may have resulted in higher subsequent wage bids on average. However, it does not account for the strategic implications of the maximum bid rule binding. For example, consider a subject for whom the sum of the optimal bids based on rational expectations of performance exceeded the limit. For such an employer, increasing the bid on one worker necessarily decreases the bid on some other worker. On the other hand, an employer for whom the sum of the optimal bids does not exceed the limit would not have to allocate wage bids in this "zero-sum" way. Thus, the strategic implications of the bidding environment would be different for these two employers. In order to explore the effects of the bidding strategies on the beauty premium, Table 9 decomposes the effects according to whether the maximum bid was binding for a given employer (the sum of bids on the workers added up to 125 or 175, depending on the session).

Table 9: Effects of strategic bidding on the existence of the beauty premium

Outcome variable:	Employer's wage bid							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	All Tasks	Data Entry	Data Analysis	Bargaining	All Tasks	Data Entry	Data Analysis	Bargaining
Attractiveness of worker & maximum	0.269*	-0.074	0.204	0.623*	0.258*	-0.023	0.179	0.588*
bid does not bind & resume shown first	(0.115)	(0.221)	(0.229)	(0.240)	(0.109)	(0.196)	(0.226)	(0.229)
Attractiveness of worker & maximum	0.094	0.096	(0.233)	0.347*	0.094	0.125	(0.194)	0.308*
bid does not bind & photo shown first	(0.094)	(0.178)	(0.214)	(0.148)	(0.090)	(0.168)	(0.214)	(0.139)
Attractiveness of worker & maximum	0.274	0.124	1.472*	0.229	0.238	0.157	1.445*	0.187
bid binds & resume shown first	(0.227)	(0.600)	(0.670)	(0.329)	(0.230)	(0.692)	(0.689)	(0.333)
Attractiveness of worker & maximum	0.026	(0.989)	0.635	(0.279)	(0.037)	(1.018)	0.599	(0.361)
bid binds & photo shown first	(0.315)	(0.670)	(0.762)	(0.392)	(0.317)	(0.610)	(0.743)	(0.386)
Employer's performance prediction					0.015**	0.014**	0.044*	0.013*
					(0.004)	(0.005)	(0.019)	(0.006)
Worker characteristics	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	685	248	209	228	685	248	209	228
R-squared	0.17	0.22	0.31	0.24	0.20	0.26	0.33	0.27

Notes: Round 1 data only. All regressions include a dummy for total maximum bid binding, date fixed effects, indicators for student status (graduate or undergraduate), major, GPA range, self-reported abilities, race, and gender. In cases where we pool the data across multiple tasks, we also include a task fixed effect. Robust standard errors clustered by employer and session in parentheses.
Significance levels: $^{+}$10 percent, * 5 percent, ** 1 percent

The positive beauty premium in the pooled data seems to exist only when the total bid maximum does not bind (Specifications 1 and 5). Once again, we confirm that the beauty premium is not present in the data entry task (Specifications 2 and 6). We find a significant positive relationship between beauty and wage bid in data analysis when the maximum bid binds and resume is shown first (Specifications 3 and 7). Finally, the beauty premium in bargaining is only statistically significant when the bid maximum does not bind (Specifications 4 and 8). Recall that so far we had found a marginally significant beauty premium in the bargaining task when resume was shown first (a coefficient of 0.550 without controlling for the prediction in Table 6 and 0.432 once we control for the prediction in Table 7). The results reported in Table 9 suggest that employers respond to the maximum bid limit by reducing their bids on attractive workers in bargaining, but not in the other tasks.

D2. Second round behavior

Table 10 investigates the relationship between attractiveness and the wage bid in round 2, unconditional on learning by observing performance in round 1 or through the different or same nature of the round-two task.

Table 10: Determinants of employer's wage bid in round 2

Outcome variable:	Natural Logarithm of employer's wage bid in round 2							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	All Tasks		Data Entry		Data Analysis		Bargaining	
Attractiveness of worker &	0.136	0.088	0.348	0.321	0.333	0.124	-0.048	-0.059
Resume shown first	(0.107)	(0.102)	(0.287)	(0.262)	(0.294)	(0.247)	(0.162)	(0.174)
Attractiveness of worker &	0.112	0.042	0.246	0.05	0.137	0.176	0.099	0.076
Photo shown first	(0.104)	(0.096)	(0.275)	(0.257)	(0.158)	(0.137)	(0.217)	(0.184)
Employer prediction in round 2		0.029**		0.030**		0.160**		0.025**
		(0.004)		(0.007)		(0.031)		(0.005)
Worker characteristics	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	685	685	225	225	236	236	224	224
R-squared	0.09	0.22	0.20	0.31	0.25	0.38	0.19	0.33

Notes: Round 2 data only. All regressions include date fixed effects, indicators for student status (graduate or undergraduate), major, GPA range, self-reported abilities, race, and gender. In cases where we pool the data across multiple tasks, we also include a task fixed effect. Robust standard errors clustered by employer in parentheses. Significance levels: $^{+}$10 percent, * 5 percent, ** 1 percent

Unlike the round-one findings in Tables 6 and 7, the analogous analysis reported in Table 10 shows no beauty premium in any task, whether or not we control for the employer's prediction for worker performance in the second round. This is consistent with our learning results discussed in the previous section.

We also consider the relationship between the employer prediction of second round performance and worker attractiveness. The results are shown in Table 11. Employer predictions of worker performance are positively correlated with the beauty rating in the pooled data and in data entry, but not in data analysis or bargaining. Surprisingly, attractiveness and performance predictions are only correlated in the sessions where the worker photo was shown first. The absence of a significant relationship for the bargaining task is again consistent with the previous learning results.

Table 11: Relationship between an employer's performance expectations and a worker's attractiveness in round 2

Outcome variable:	Employer's prediction of worker performance in round 2							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	All Tasks		Data Entry		Data Analysis		Bargaining	
Attractiveness of worker &	1.699	1.282	0.905	5.388+	1.305	0.523	0.425	-0.089
Resume shown first	(1.353)	(1.193)	(3.038)	(2.745)	(0.783)	(0.730)	(2.273)	(1.899)
Attractiveness of worker &	2.441*	2.774**	6.417*	5.499*	-0.242	-0.098	0.916	1.874
Photo shown first	(1.084)	(0.910)	(2.566)	(2.338)	(0.374)	(0.351)	(2.114)	(2.111)
Performance in round 1		0.281**		0.293**		0.036**		0.408**
		(0.031)		(0.068)		(0.008)		(0.067)
Worker characteristics	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	685	685	225	225	236	236	224	224
R-squared	0.74	0.81	0.50	0.58	0.27	0.36	0.46	0.64

Notes: Round 2 data only. All regressions include date fixed effects, indicators for student status (graduate or undergraduate), major, GPA range, self-reported abilities, race, and gender. In cases where we pool the data across multiple tasks, we also include a task fixed effect. Robust standard errors clustered by employer in parentheses. Significance levels: +10 percent, * 5 percent, ** 1 percent

Finally, we look at the relationship between a worker's attractiveness and performance in round 2. The results are shown in Table 12. Once worker characteristics are controlled for, performance in the second round is uncorrelated with beauty in any of the tasks.

Table 12: Relationship between a worker's attractiveness and her performance in round 2

Outcome variable:	Worker performance in round 2							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	All Tasks		Data Entry		Data Analysis		Bargaining	
Attractiveness of worker	2.105	0.618	-3.158	7.568	-0.131	-0.315	14.009*	5.113
	(2.788)	(2.831)	(6.938)	(5.916)	(0.686)	(0.731)	(6.225)	(6.154)
Worker characteristics	No	Yes	No	Yes	No	Yes	No	Yes
Observations	685	685	225	225	236	236	224	224
R-squared	0.55	0.64	0.09	0.56	0.36	0.64	0.35	0.63

Notes: Round 2 data only. All regressions include date fixed effects, indicators for student status (graduate or undergraduate), major, GPA range, self-reported abilities, race, and gender. In cases where we pool the data across multiple tasks, we also include a task fixed effect. Robust standard errors clustered by worker in parentheses. Significance levels: +10 percent, * 5 percent, ** 1 percent

V. Conclusion

In this paper, we develop and execute a new method for determining the precise channel through which attractiveness leads to higher worker wages. Our key insight is that having two measures, one that elicits expected worker performance and one that elicits employer willingness to pay, is both necessary and sufficient for separating statistical discrimination from taste-based discrimination. In addition, statistical discrimination can be further decomposed into biased beliefs and rational statistical discrimination given a measure of actual performance data.

We run a laboratory experiment designed to elicit the three measures described above as well as to identify the stability of the beauty premium across different settings. While carefully controlling the overall experimental environment, we vary the tasks that workers must perform. Our results indicate that the beauty premium is highly context dependent: While we find strong evidence of a beauty premium in the bargaining task, there is no beauty premium in the data entry or data analysis task. The beauty premium is composed of both statistical and taste-based discrimination, and the statistical component of the beauty premium can be explained by biased beliefs about the performance of more attractive workers rather than rational statistical discrimination.

We also examine whether the beauty premium disappears after employers get a better signal of worker performance. We find a strong learning effect: The beauty premium in the bargaining task disappears after worker performance is revealed, even in cases where the bargaining task is preceded by data entry or data analysis. This suggests that, in our setting, employers use attractiveness as an imperfect signal of ability. The taste-based component of the beauty premium, on the other hand, remains.

Eliciting performance beliefs and wages in the field can raise the cost of a study. Moreover, researchers must be careful to elicit beliefs in a neutral manner to ensure that they are revealed honestly. Despite these challenges, we believe that proceeding along these lines is an important next step in the study of discrimination.

Because of the experimental nature of our study, we are hesitant to draw strong policy implications from our results. Thus, the following discussion is conditional on our results generalizing to real labor markets. Because we do not find persistent biased beliefs in favor of more attractive people, the welfare losses from allowing beauty-based pay differentials are likely to be small. On the other hand, outlawing discrimination based on looks may be warranted if policymakers do not want to allow for wage differentials that are driven by tastes.

The above discussion is subject to a caveat that we do not observe attractiveness-based performance differentials across the different tasks. If the spectrum of real-world tasks exhibits such differentials, eliminating beauty-based pay differentials may lower the quality of matching between workers and jobs, leading to welfare losses. Testing for the existence of performance differentials across a number of jobs is an important step for future research.

References

Andreoni J. and R. Petrie. 2008. “Beauty, gender and stereotypes: Evidence from laboratory experiments.” *Journal of Economic Psychology* 29: 73–93.

Ayres, I. and P. Siegelman. 1995. “Race and Gender Discrimination in Bargaining for a New Car.” *American Economic Review*, 83(3): 304-321.

Berggren N., H. Jordahl, and P. Poutvaara. 2007. "The Looks of a Winner: Beauty, Gender, and Electoral Success," CESifo Working Paper Series 2002, CESifo Group Munich.

Berri, D. J., R. Simmons, J. Van Gilder, and L. O'Neill. 2011. “What does it mean to find the face of the franchise? Physical attractiveness and the evaluation of athletic performance.” *Economic Letters*, 111: 200-202.

von Bose, Caroline. 2012. “Child Stars vs. Ugly Ducklings: Does adolescent attractiveness contribute to the beauty premium?” Working paper.

Castillo, M., R. Petrie, M. Torero, and L. Vesterlund. 2012. “Gender Differences in Bargaining Outcomes: A Field Experiment on Discrimination.” NBER Working paper No. 18093.

Cooper, D. J. and H. Fang. 2008. “Understanding Overbidding in Second Price Auctions: An Experimental Study.” *The Economic Journal*, 118:1572-1595.

Deryugina, T. and O. Shurchkov. 2012. “Is There a Beauty Premium in Undergraduate Education?” Working paper.

Fischbacher, U. 2007. “z-Tree: Zurich toolbox for ready-made economic experiments.” *Experimental Economics*, 10(2), 171-178.

Hamermesh, D. and J. Biddle. 1994. “Beauty and the Labor Market.” *American Economics Review,* 84(5): 1174-1194.

Hamermesh, D. and J. Biddle. 1998. “Beauty, Productivity, and Discrimination: Lawyers’ Looks and Lucre.” *Journal of Labor Economics,* 16(1): 172-201.

Hamermesh, D. 2011. “Beauty Pays: Why Attractive People Are More Successful.” Princeton University Press, Princeton, NJ, USA.

Mobius, Markus M., and T. S. Rosenblat 2006. "Why Beauty Matters." *American Economic Review*, 96(1): 222–235.

Ravina, E. 2009. “Love & Loans: The Effect of Beauty and Personal Characteristics in Credit Markets”, under revision.

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

Appendix A. Rating procedure and instructions

All pictures were rated by at least 25 female and 25 male raters. Raters were students at a college in a different state and were pre-screened to ensure that they were not familiar with students from the college of interest.

Raters were shown pictures of each student and asked to rate her physical appearance on a 1-10 point scale. Five of the numbers had descriptions describing the level of attractiveness corresponding to that number (see experimental instructions on the next page). Raters were instructed to choose the numbers without descriptions if they felt the student's appearance fell in between the two descriptions.

Each rater was shown four sets of about 100 photos. The order of the photos within each set was randomized for each rater. In early stages of the experiment, we compared the mean and standard deviation of ratings across different sets to see if having subjects rate 400 pictures led to fatigue. There was no significant difference in either the mean or standard deviation of ratings for earlier and later sets, which led us to conclude that 400 pictures was not an excessive amount. We did not use data from three raters who chose 1's 40% or more percent of the time. The "1" option was the closest to the "Next" button. Thus, these subjects were most likely trying to get through the experiment as quickly as possible.

Instructions for the experiment

You are about to participate in an experiment involving the perception of appearance. Once the experiment begins, you will see a photograph of an individual along with the following prompt:

> Rate this person's physical appearance using the following scale:
>
> 10 strikingly handsome or beautiful
> 9
> 8 good-looking (above average for age and sex)
> 7
> 6 average looks for age and sex
> 5
> 4 quite plain (below average for age and sex)
> 3
> 2 homely
> 1

Choose the number that best corresponds to your evaluation. Choose the numbers without descriptive text (1, 3, 5, 7, and 9) if you feel the person's appearance falls in between the descriptions found in the adjacent numbers.

After you have chosen a number, click "Next". You will then see another photograph and be asked to repeat the procedure. Continue selecting the number you feel best reflects your assessment of the individual's appearance until you are told to stop.