An Experimental Study of Statistical Discrimination by Employers*

Nick Feltovich
Department of Economics
University of Houston
Houston, TX 77204
nfelt@mail.uh.edu

Chris Papageorgiou
Department of Economics
Louisiana State University
Baton Rouge, LA 70803
cpapa@lsu.edu

February 2003

Abstract

This paper reports findings of an experiment motivated by a dynamic labor market model that considers the problem faced by an employer in making hiring decisions of workers of different types. The question examined here is how quickly employers learn about the ability of a group of workers through observing the performance of representatives of that group in the workplace. If prior opinions are weak, the employer will quickly update any incorrect group–based stereotypes it may have, with information from the workplace. On the other hand, if priors are heavily weighted, incorrect initial perceptions will result in persistent differences in wages. Our experimental findings are twofold. First, subjects' (employers') behavior moves quickly toward optimal choices. Second, strong priors are hard to establish. These results suggest that it would take a long time for employers to form group–based stereotypes, and that such stereotypes should go away quickly in response to signals that contradict these stereotypes.

JEL Classification Numbers: J24, J71, D81.

*We thank John Duffy, Sudipta Sarangi, Dek Terrell, and an anonymous referee for helpful comments and discussions.

1 Introduction

This paper reports results from an experiment that was motivated by the literature on labor market discrimination. Our aim for conducting this experiment was to investigate whether employers' initial perceptions of employee ability on the basis of group characteristics can lead to lower wages that persist for a long time. Under the maintained assumption that the employer learns about the group's ability through Bayesian updating, we examine how quickly he learns.

The idea that inaccurate prior assessment by managers formed on the basis of an employee's group influences wages is one of many theories of labor market discrimination in economics. For economists, discrimination implies that workers in one group earn less than the competitive market rate for their labor, typically due to their gender or ethnic group. The economics literature contains many theories seeking to explain labor market discrimination and empirical work which attempts to test those theories and measure discrimination through observed wages (see, for example, Altonji and Blank (1999), who provide a thorough survey of theoretical and empirical research on wage differences and discrimination, including statistical discrimination, as well as their probable causes).

Economic models of discrimination were initially developed to address the empirical findings of many researchers that wages differ across groups and the widespread belief that wage differences stem in part from discrimination. Existing theoretical models of labor market discrimination fall into two distinct categories that are based on the source of discrimination. First, there are theories based on tastes, such as Becker's (1972). According to such theories, employers have a preference for not hiring workers of a particular group, or fellow employees have a preference for not working with workers of a particular group. Second, there are statistical theories, pioneered by Phelps (1972) and Arrow (1973). Statistical theories of discrimination focus on the idea that, when a prospective employee's true ability is unobservable, the employer may rationally use the employee's ethnic group or gender as a proxy for his ability.

Our study focuses on an extension to the basic statistical discrimination model. The initial model by Phelps (1972) simply argued that high–ability workers in groups with more variability in ability would earn less than high–ability members of other groups. Lundberg and Startz (1983) developed a more complex model to show that even if two groups possessed the same average ability, a higher variance of ability in one group would lead to lower wages for members of that group relative to a low–variance group. Farmer and Terrell (1996) further extended this model to look at the possibility that inaccurate initial assessments of ability could become self–fulfilling prophecies. For example, lower initial assessments of worker ability by employers could diminish that worker's marginal returns to additional training or education, and thus decrease his incentives to obtain skills. His ability would then remain low, reinforcing the employer's assessment.

The empirical literature fails to produce a decisive conclusion on the sources of discrimination or even the extent to which it affects wages in the United States today. While average wages differ across groups, wage differences could simply reflect differences in worker ability. Explanations for differences in ability vary considerably. For example, Herrnstein and Murray's (1994) controversial book *The Bell Curve: Intelligence and Class Structure in American Life* asserts that races simply differ in inherent ability, while Card and Krueger (1992) argue that differences in quality of education for black and white workers explain a significant portion of the wage gap. A critical issue for empirical studies is that worker ability is unobserved. This makes it difficult to break wage differences across groups into one portion that is attributable to differences in ability and a second portion attributable to pure discrimination. Testing theories that explain discrimination or sources of differences in ability is even more difficult.[1]

In this paper we turn to a laboratory experiment as a step toward understanding the impact of an employer's prior opinions formed on the basis of an employee's group on wages. The critical issue is how quickly employers learn about workers' true abilities through observing noisy information about their performance in the workplace. If prior opinions are weak, the employer will quickly update any group–based stereotypes with information from the workplace. However, if initial assessments are heavily weighted, the initial perception may lead to persistent differences in wages.

The remainder of the paper is organized as follows. Section 2 describes a simple model of labor market discrimination in which manager learns about an employee's ability through Bayesian updating. Sections 3 and 4 present the experimental design and procedures employed in this study. Sections 5 and 6 discuss the main results of the experiment and tie them to the existing literature.

[1] Altonji and Pierret (2001) attempt to measure the ability of statistical discrimination to explain racial differences in wages. One of their findings is that either firms do not statistically discriminate on the basis of race, or there is little correlation between race and productivity among workers.

2 The Model

In order to motivate our experiment, we discuss a model based on Farmer and Terrell (1996) and Lewis and Terrell (2001), who examine a statistical discrimination framework with Bayesian updating of employers' beliefs. A large number of employers hire workers for one period from a large pool of potential employees. The labor market is competitive, so that workers are paid their expected marginal product in each period, $w_{it} = E(y_{it})$. The marginal output of worker i in period t is given by the following production technology:[2]

$$y_{it} = A_i^{\alpha} e^{\varepsilon_{it}}, \text{ where } \varepsilon_{it} \sim N(0, \sigma^2) \text{ i.i.d..} \tag{1}$$

The random variable A reflects the ability of all workers with the same observable characteristics as worker i, while the random variable ε is an individual–specific component that is normally distributed and i.i.d. across workers. The values of A and ε are unobservable to the employer, but A can be gradually learned over time. Note that because of our assumption about the distribution of ε, one can generate the log–normal distribution of wages initially observed by Mincer (1974). Taking logarithms in Equation (1) yields

$$\log y_{it} = \alpha \log A_i + \varepsilon_{it}.$$

Without loss of generality we assume that $\alpha = 1$, so that $y_{it} = A_i e^{\varepsilon_{it}}$ and $\log y_{it} = \log A_i + \varepsilon_{it}$. The normality assumption on ε implies that the distribution of log–output conditional on group log–ability is normal and given by

$$(\log y_{it} | \log A_i) \sim N(\log A_i, \sigma^2)$$

or more explicitly,

$$f(\log y_{it} | \log A_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(\log y_{it} - \log A_i)^2\right].$$

We further assume that the (representative) employer gradually learns about the ability of an employee type by making T sequential observations of employees' output. This assumption is at the heart of our investigation in examining the persistence of the employers' priors about employees'

[2]The models of Farmer and Terrell (1996) and Lewis and Terrell (2001) add an endogenous individual–specific human–capital parameter Z_{it}, which affects the worker's marginal product. Since we are not modeling workers' decisions in this context, and because workers are employed by a firm for only one period, we can set $Z_{it} = 1$ to obtain our simpler model.

abilities that are initially unobservable. This assumption along with the independence property of the assumed error distribution, implies

$$f(\log y_{i1}, ..., \log y_{iT} | \log A_i) = \frac{1}{(2\pi\sigma^2)^{\frac{T}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum_{t=1}^{T}(\log y_{it} - \log A_i)^2\right].$$

The employer's initial beliefs about employees' ability is characterized by the prior distribution function given by

$$(\log A_i) \sim N(\bar{\mu}, \bar{\sigma}^2)$$

where $\bar{\mu}$ is the mean of a normal prior reflecting the "best guess" about employees' group ability, and $\bar{\sigma}^2$ is a measure of certainty of prior beliefs. $\bar{\mu}$ and $\bar{\sigma}^2$ are allowed to vary across groups as employers' priors depend on employees' group. Employers are assumed to use Bayesian updating when forming beliefs about the ability of workers. So, beliefs at time T are calculated as

$$f(\log A_i | \log y_{i1}, ..., \log y_{iT}) = \frac{f(\log y_{i1}, ..., \log y_{iT} | \log A_i) f(\log A_i)}{f(\log y_{i1}, ..., \log y_{iT})}.$$

This in turn implies

$$(\log A_i | \log y_{i1}, ..., \log y_{iT}) \sim N(\mu_T, \sigma_T^2)$$

where

$$\mu_T = \sigma_T^2 \left(\frac{\sum_{t=1}^{T} \log y_{it}}{\sigma^2} + \frac{\bar{\mu}}{\bar{\sigma}^2}\right) \quad \text{and} \quad \sigma_T^2 = \left(\frac{T}{\sigma^2} + \frac{1}{\bar{\sigma}^2}\right)^{-1}.$$

The mean of the updated distribution for a worker's type log–ability, μ_T, is the weighted average of predicted ability based on job performance (given by the term $\sum_{t=1}^{T} \log y_{it}$ weighted by $\frac{\sigma_T^2}{\sigma^2}$) and prior opinion about ability (given by $\bar{\mu}$ weighted by $\frac{\sigma_T^2}{\bar{\sigma}^2}$). The variance of the updated distribution for a worker's type log–ability, σ_T^2, depends on the variance of mean log–ability from observed output and the variance of prior opinion.

It is interesting to consider what happens to μ_T and σ_T^2 as $T \to \infty$: that is, as the amount of information about employees' abilities becomes large.[3] It is easy to show that $\lim_{T\to\infty} \sigma_T^2 = 0$, which intuitively implies that in the limit there is no uncertainty in employers' belief about employees performance. Deriving $\lim_{T\to\infty} \mu_T$ requires a bit more work. First rewrite μ_T using the definition of σ_T^2 as

$$\mu_T \quad = \quad \left(\frac{T}{\sigma^2} + \frac{1}{\bar{\sigma}^2}\right)^{-1} \left(\frac{\sum_{t=1}^{T} \log y_{it}}{\sigma^2} + \frac{\bar{\mu}}{\bar{\sigma}^2}\right)$$

[3]We thank the referee for suggesting this exercise.

$$= \frac{\frac{\sum_{t=1}^{T} \log y_{it}}{T}}{1 + \frac{1}{T}\frac{\sigma^2}{\bar{\sigma}^2}} + \frac{\frac{\sigma^2 \bar{\sigma}^2 \bar{\mu}}{T}}{\sigma_A^4 + \frac{\sigma^2 \bar{\sigma}^2}{T}}. \tag{2}$$

Since $\sigma_T^2 \to 0$ as $T \to \infty$, the second term in Equation (2) approaches 0 as $T \to \infty$, and the denominator of the first term approaches 1. Hence,

$$\begin{aligned} \lim_{T\to\infty} \mu_T &= \lim_{T\to\infty} \frac{\sum_{t=1}^{T} (\log A_i + \varepsilon_{it})}{T} \\ &= \log A_i, \end{aligned} \tag{3}$$

as $\log \varepsilon_{it} \sim N(0, \sigma^2)$ and therefore $\lim_{T\to\infty} \frac{\sum_{t=1}^{T} \log \varepsilon_{it}}{T} = 0$. The main intuition behind Equation (3) in light of the mean predicted ability's two components (job performance and prior opinion) is as follows: first, prior beliefs about ability become unimportant as the employer has $T \to \infty$ observations of the employee's output (as $T \to \infty$, the second term in Equation (2) approaches 0). Second, mean predicted ability based on job performance, in the limit, converges to log–ability of the group (as $T \to \infty$, the first term in Equation (2) approaches $\log A_i$).

Using Bayesian updating in the specification of employer beliefs has the advantage of implying that if there are systematic differences in ability across worker types, employers gradually learn to show preference for the higher–ability types, and thus, to be willing to pay them higher wages. Unfortunately, the speed at which this happens depends on characteristics of the employer that are unobservable to the researcher—in particular, how heavily they weight their prior probabilities, relative to the information they receive in each period. Bayesian updating is probably the most commonly used (by economic theorists) model for combining old and new probability information. However, its success as a descriptive theory is mixed. The psychology and behavioral economics literatures are replete with examples in which individuals, even when given complete descriptions of a probability–updating problem including both base rates (which are equivalent to employers' priors in our model) and likelihood information (equivalent to the observed new productivity information in our model), underweight base rates relative to the likelihood information, and other examples in which they overweight the base rates. For example, Camerer (1995) provides a thorough survey of experimental studies of individual decision making in economic situations. On the other hand, Bayesian updating has been used successfully by some researchers for describing individual decision making in probabilistic situations (see, for example, Anderson and Holt (1997)).

3 The Decision Problem Used in the Experiment

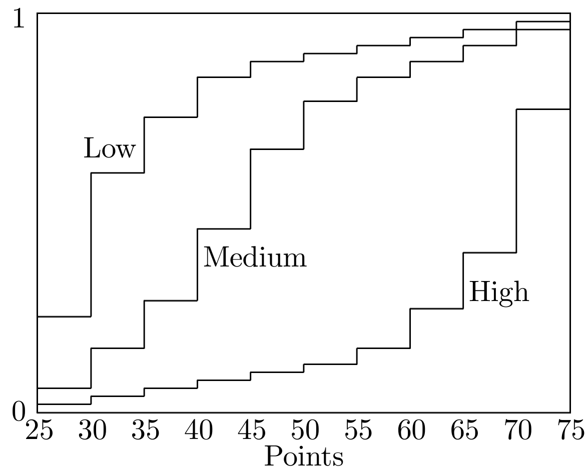
The experiment was designed in a attempt to capture a simplified version of the decision problem faced by an employer in the above model, while avoiding obviously loaded terms. All subjects in the experiment faced the same decision problem, which we now describe. In each of nine rounds, subjects were presented with two buckets, each of which contained 50 cards. Subjects were asked to draw a total of four cards (with replacement) from the two buckets; they could draw all four from a single bucket, two from each bucket, or three from one bucket and one from the other. The buckets are meant to correspond to a group of workers sharing some observable characteristic; the individual cards represent individual workers having that characteristic (that is, workers of a given "type"). Each card had a number printed on it, representing the true marginal productivity of that worker. (Thus the mean of the numbers in a bucket represents the "average ability" of that worker type.) The subject's total revenue (in points) was the sum of the numbers on the four cards drawn. Her total cost was determined by the number of cards drawn from each bucket; it cost 60 points to draw two from each bucket, 70 points to draw three and one, and 100 points to draw entirely from one bucket. The subject's profit in a round was her total revenue minus her total cost.

Since subjects draw four cards in each round, and are not allowed to hold on to cards for future rounds, we are making the implicit assumption that firms hire new workers in every round. Notice also that we do not address wages here, but rather only demand for employees of a given type. Of course, unless labor supply is infinitely elastic, there will be a positive relationship between labor demand and equilibrium wages. Because each bucket contains a nontrivial distribution of cards, subjects (managers) are unable to know ahead of time exactly how productive a given worker will be. But, if the distributions are different across buckets, as they are in the first six rounds, the bucket from which a card is drawn (that is, the type of the worker) contains some information about the worker's expected productivity. The rationale for costs increasing as more cards are drawn from the same bucket is to model diminishing returns in a particular type of worker. An additional consequence of these increasing costs is that when the two buckets have the same distribution of cards, it is *strictly* optimal to draw equally from both buckets.

There were three distributions of cards used—High, Medium, and Low—which could be ordered by stochastic dominance. These distributions are shown in Figure 1.[4] In the first six rounds, one

[4] One departure of our experiment from the model is the non–normality of the distribution of the ε's. Some aspects of the non–normality were necessary, such as the discreteness of the distribution (an infinite number of cards would be

Figure 1: Distributions of cards used in decision problem (cdf's)



bucket contained a High distribution and the other a Low distribution of cards; we will refer to these as Bucket 1 and Bucket 2, respectively. In the last three rounds, both buckets contained Medium distributions (that is, there was no difference in distribution across buckets in these rounds). The two buckets were differently colored (one green and one tan), so subjects could easily tell them apart. The distributions were chosen so that (1) profits were guaranteed to be nonnegative, (2) if the subjects had perfect information about the distributions, the optimal choice for the first six rounds would be to choose all four cards from Bucket 1, and (3) if the subjects had perfect information about the distributions, the optimal choice for the last three rounds would be to choose two cards from each bucket.

The motivation for the round–to–round sequence of distributions we used was to allow the subjects to build up experience of one bucket being noticeably "better" than the other one, and then to see how they respond to a situation in which neither bucket is better on average (though even in this latter case, because of the randomness in the distributions, it may seem to a given subject that one or the other bucket is better). This corresponds to a situation an employer might face, where one type of worker has historically been more productive than another (though there is variability in productivity within a type), but there is no longer a difference between the types.

time–consuming to prepare) and its boundedness (negative profits were avoided, as well as potentially large positive ones). We also thought it desirable to give all three distributions the same support, so that any inferences about which distribution was better could be made only probabilistically. This last desideratum is the reason for the high skewness of the High and Low distributions.

4 Experimental Procedures

The experimental sessions were conducted at Louisiana State University and at the University of Houston. Each subject was seated at a desk and given written instructions and a record sheet on which to record decisions and resulting outcomes.[5] These instructions were then read aloud, and any questions were answered, prior to the first round of play.

The experiment was conducted with pen–and–paper. During a given round of play, each subject decided how many cards to draw out of each bucket, and circled the chosen buckets on her record sheet. The monitor would then go to the subject's desk, and the subject would draw from the appropriate buckets, one at a time. After drawing a card from a bucket, the subject would record the card's number on her record sheet and then replace the card in the bucket before drawing again. After drawing four cards and recording the results in this way, the subject would fill in the entries for total revenue, total cost, and profit, and the monitor would move on to the next subject.

After the third and sixth rounds, it was announced that the cards in the buckets would be replaced, and subjects were able to observe the monitor putting new cards into the buckets. Announcing changes of the distributions is of course a departure from the situation faced by actual employers, who typically would not have this information. However, it was necessary to avoid deception of the subjects, which is generally considered bad methodology by experimental economists. The cards used in rounds 4–6 had the same distributions as those used in rounds 1–3, as mentioned previously; the distributions were changed for rounds 7–9 (see Table 1 and Figure 1).

Table 1: Distributions of cards used in the experiment

Rounds	Bucket 1 cards	Bucket 2 cards
1–3	High	Low
4–6	High	Low
7–9	Medium	Medium

Within a three–round block, it was known by the subjects that the distributions of cards in the buckets did not change, so that the results of the first round in a block would be useful for making decisions in the second round in that same block, and the results of the first two rounds would be

[5] Sample instructions can be found in the Appendix. Additionally, sample record sheets and the raw data from the experiment are available from the authors upon request. Notice that our instructions and record sheets use context–free language. We wanted to avoid any language that would allow subjects to figure out that the subject of this experiment was labor–market discrimination, for fears that demand effects would be strong.

useful for making decisions in the third round in that block. Since the results of each round were recorded on the subjects' record sheets, it was easy for them to do so, if they wanted. After the third round in a block was over, and the cards were physically changed, it should have been much less apparent to subjects that previous results would be useful (though in Rounds 4–6, they would have been, and they might think so in Rounds 7–9 also).

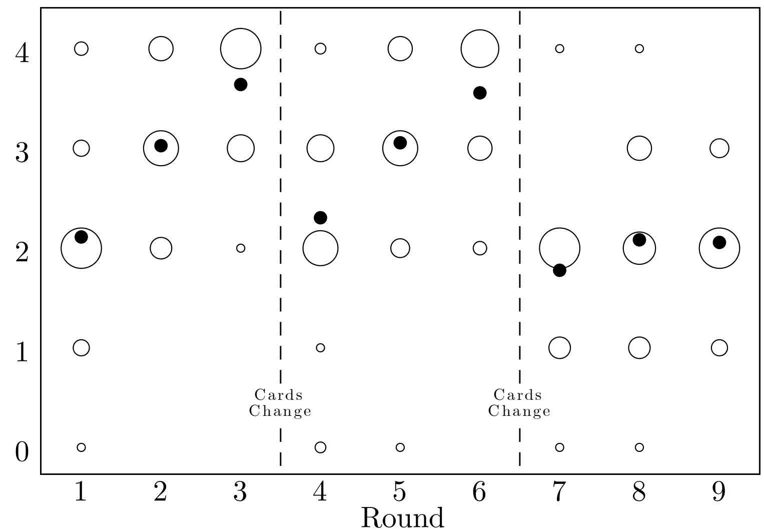
After the ninth round was over, the session ended. Subjects were paid a \$2.00 showup fee; in addition, one round was randomly chosen, and subjects were paid their profit in that round, at the exchange rate of 5 cents per point. Subjects earned an average of about \$9.00 for participating in the experiment.

Our hypotheses were as follows. First, within a block of three rounds (Rounds 1–3, 4–6, and 7–9), if subject behavior is originally different from optimal behavior, it will tend to move in the direction of optimal behavior: choosing entirely from Bucket 1 in the first six rounds, and choosing equally from both buckets in the last three rounds. Second, and more interesting, we expected that the results of the first 6 rounds will lead to subjects' having inaccurate (though reasonable, given their results up to that point) priors for the last 3 rounds, and will affect their behavior accordingly. In particular, the early experience of Bucket 1 being better will lead subjects to choose Bucket 1 more often than Bucket 2, even when it is no longer better.

5 Results

A total of 36 subjects participated in the experiment. Figure 2 shows some features of the experimental data. Shown in this figure are the number of subjects choosing Bucket 1 (the bucket with the higher distribution of cards, when the buckets had different distributions) in each round. These distributions are represented by the open circles; the area of a circle is proportional to the number of subjects making that choice in that round. Also shown is the average frequency of Bucket 1 choices by all subjects in each round (closed circles; see also Table 2). If subjects actually knew the distributions in the buckets, they should optimally choose to draw all four cards from Bucket 1 in the first six rounds, and two cards from each bucket in the last three rounds. In fact, they didn't know the distributions, but the data are consistent with their learning these distributions. In Round 1, most subjects draw equal numbers of cards from each bucket, but the number drawn from Bucket 1 increases from Round 1 to Round 3, where the modal choice is all four cards from

Figure 2: Bucket 1 choices—distributions (open circles) and mean choices (closed circles)



Bucket 1. In Round 4, when new cards are put into the buckets, the modal choice falls back to 2 cards from each bucket. Again, the number drawn from Bucket 1 increases from Round 4 to Round 6, until the modal choice in Round 6 is all four cards from Bucket 1. In Round 7, when new cards are again put into the buckets (this time with identical distributions in the two buckets), the modal choice again falls back to 2. It remains at 2 for the remaining rounds.

The increase in Bucket 1 choices from Round 1 to Round 3 is statistically significant (Page test for ordered alternatives, $p \approx 0.001$), as is the increase from Round 4 to Round 6 (Page test, $p \approx 0.002$). (See Siegel and Castellan (1988) for descriptions of the nonparametric statistical tests used in this paper.) There is no significant increase over the last three rounds according to the Page test ($p \approx 0.509$), and even according to the weaker Friedman two–way analysis of variance test, there is no significant difference in the level of Bucket 1 choices over these rounds ($p \approx 0.18$).

In every round, the modal choice is also the median choice, and looking at the means doesn't affect these qualitative conclusions substantially. In essence, when Bucket 1 contains a higher distribution of cards than Bucket 2, subjects learn to choose Bucket 1 more and more often; when the buckets contain the same distribution of cards, subjects continue to choose both buckets roughly equally, on average. In other words, consistent with our first hypothesis, subjects' choices moved toward optimal play (though not actually reaching it).

Table 2: Desciptive Statistics

Round	Median	Mean	Std. Error
1	2	2.11	0.82
2	3	3.03	0.70
3	4	3.64	0.54
4	2	2.31	0.86
5	3	3.06	0.83
6	4	3.56	0.65
7	2	1.78	0.64
8	2	2.08	0.84
9	2	2.06	0.53

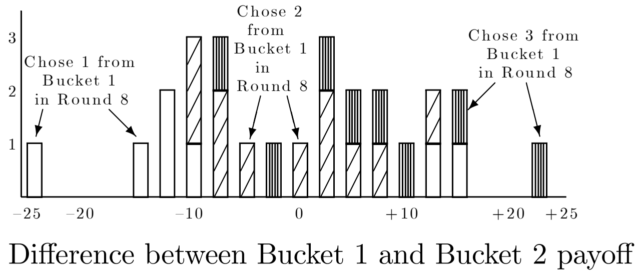
It is harder to detect evidence consistent with our second hypothesis. The number of Bucket 1 choices in Round 4 is slightly higher than in Round 1, which could be taken as evidence that the results of Rounds 1–3 are carrying over into subjects' beliefs in Round 4. However, we don't see the same effect in Round 7; there are actually fewer Bucket 1 choices then (though neither the change from Round 1 to Round 4 nor that from Round 4 to Round 7 is significant).

One aspect of the data that is consistent with early–round results influencing later–round play can be seen in Round 8. Even though the average choice over the last three rounds remains roughly constant at 2 draws from each bucket, the shape of the distribution varies noticeably over these rounds. Notice from Table 2 that the standard deviation of Bucket 1 choices increases sharply from Round 7 to Round 8, then decreases again in Round 9 (this change is also apparent in Figure 2). That is, in Round 8, the average of two Bucket 1 choices conceals a substantial number of choices of more than two or fewer than two draws from Bucket 1. It is possible that the difference in distributions between buckets in the first six rounds leads subjects initially (in Round 7) to expect that there will be a difference in the last three rounds also, even if they don't know which bucket has the higher distribution. If so, subjects' choices in Round 8 should be highly dependent on the results of their own draws in Round 7. This is indeed the case. Subjects earning higher average payoffs from their Bucket 1 draws than their Bucket 2 draws in Round 7 chose Bucket 1 in Round 8 roughly 55% of the time, while those subjects earning higher payoffs from their Bucket 2 draws chose Bucket 1 only about 44% of the time. The difference between these conditional relative frequencies is significant (robust rank–order test, $p \approx 0.075$).

If we look only at subjects drawing two cards from each bucket in Round 7, this result becomes

Figure 3: Number of Round–8 choices from Bucket 1, conditional on Round–7 payoffs

(Subjects choosing 2 cards from each bucket in Round 7)



even more stark (see Figure 3). Of these subjects, those choosing in Round 8 to draw three cards from Bucket 1 had earned on average 6.56 more points from Bucket 1 than Bucket 2 in Round 7. Those continuing in Round 8 to draw two cards from each bucket had earned on average 0.9 *fewer* points from Bucket 1 than Bucket 2 in Round 7, and those drawing only one card from Bucket 1 in Round 8 had earned 6.8 fewer points from Bucket 1 than Bucket 2 in Round 7. The distribution of Round–7 payoff differentials for those choosing one card from Bucket 1 in Round 8 are significantly different from both those choosing two cards (robust rank–order test, $p < .10$) and those choosing three cards (robust rank–order test, $p < .05$).

We also find a difference if, instead of looking at the number of draws from each bucket in Round 8, we look at the *change* in the number of draws from Bucket 1 from Round 7 to Round 8. This will naturally be correlated with the number of draws from Bucket 1 in Round 8, but may be a better measure of learning, since it treats an increase from, say, one to two Bucket 1 choices the same as an increase from two to three. We find that those subjects earning higher payoffs from their Bucket 1 draws increased the number of their draws from Bucket 1 by an average of roughly 0.53 draws, while those earning higher payoffs from their Bucket 2 draws *decreased* the number of their draws from Bucket 1 by an average of roughly 0.15 draws. This difference in behavior is also significant (robust rank–order test, $p \approx 0.023$).

If these changes from Round 7 to Round 8 are indeed due to subjects' learning to expect one bucket to contain a higher distribution of cards than the other, then we would expect similar changes from Round 4 to Round 5. This seems to be the case. Only one subject earned less from Bucket 1 than Bucket 2 in Round 4; this person chose *entirely* from Bucket 2 in Round 5. Of the 31 subjects earning more from Bucket 1 than Bucket 2 in Round 4 (four others only chose from one bucket, and thus couldn't compare their payoffs across buckets), eight had an average payoff from Bucket 1 that was between 10 and 20 higher than that of Bucket 2; these increased their Bucket–1 choices an average of 0.500 in Round 5. The eight subjects who had a Round–4 average payoff from Bucket 1 that was between 20 and 30 higher, increased their Bucket–1 choices an average of 0.625, and the fifteen whose average payoff from Bucket 1 was more than 30 higher, increased their Bucket–1 choices an average of 1.000 in Round 5.

If we focus on subjects who had drawn two cards from each bucket in Round 4, we see a similar result. The one subject to draw zero cards from Bucket 1 in Round 5 had earned 10 *fewer* points from Bucket 1 than Bucket 2 in Round 4. Those continuing in Round 5 to draw two cards from each bucket had earned on average 20 *more* points from Bucket 1 than Bucket 2 in Round 4, those drawing three cards from Bucket 1 in Round 5 had earned 26.875 more points from Bucket 1 than Bucket 2, and those drawing all four cards from Bucket 1 in Round 5 had earned 40 more points from Bucket 1 than Bucket 2. Because of small subsample sizes (especially the one person choosing zero from Bucket 1 in Round 5), not all differences are significant, but the distribution of payoff differences is higher for the subjects drawing 4 cards from Bucket 1 than for those drawing 2 or 3 cards from Bucket 1 (robust rank–order test, $p = .05$ and $p < .05$, respectively) and higher for those choosing three cards than for those choosing zero (Fisher exact test, $p \approx .076$).

However, it is not only the contrast between the +0.53 difference and the –0.15 difference in Rounds 7–8 that is of interest here. The absolute magnitudes of the two numbers are also worth considering. Subjects are likely to increase their draws from Bucket 1 more—in response to favorable information from that bucket (relative to the other bucket)—than they are to increase their draws from Bucket 2 in the opposite case. This may also be a vestige from the first six rounds; after so much experience of Bucket 1 always being better, it may take less new information to convince subjects that Bucket 1 is again better, than is needed to convince them that Bucket 2 is better.

6 Discussion

Our results can be summarized as follows. In the first six rounds, Bucket 1 contains a higher distribution of payoffs than Bucket 2. In Rounds 1–3, and again in Rounds 4–6, subjects learn quickly to choose Bucket 1 most of the time. In the last three rounds, the two buckets contain the same distribution of payoffs. In these three rounds, subjects choose the two buckets roughly equally. The implications of these results for our labor market model are that when workers' observable characteristics are informative, though possibly noisy, signals of their ability, employers learn this, so that the market demand for higher–ability workers increases and the demand for lower–ability workers decreases. The difference in demands should lead to a difference in wages (though as already mentioned, our experiment looks only at demands, not wages). When workers' observable characteristics are unrelated (on average) to their ability, market demands for the two types of worker stay roughly equal, so there should be no resulting wage difference. More precisely, any observed wage difference will be due to other factors.

The results of this experiment provide only weak evidence in favor of our main hypothesis. We were interested in showing that experience in an environment where one bucket yielded higher expected payoffs than the other, would carry over into an environment in which both buckets yielded the same expected payoffs. In particular, it was expected that subjects who learned to choose all, or nearly all, cards from Bucket 1 would continue to do so, even when Bucket 1 was no better on average than Bucket 2. This did not happen in the experiment; once the identical distributions were introduced into the buckets, the average behavior of subjects was roughly two draws from each bucket.

The closest we found to an effect from the first six rounds carrying over into the last three rounds was only a second–order effect. For the most part, subjects' choices in the eighth round were highly dependent on their seventh–round results. Those who obtained higher payoffs from Bucket 1 in Round 7 were more likely to increase their choices from Bucket 1, and to choose more than 2 cards from Bucket 1, in Round 8 (as already mentioned, these two effects are highly correlated). On the other hand, those subjects who obtained higher payoffs from Bucket 2 in Round 7 were more likely to decrease their choices from Bucket 1, and to choose fewer than 2 cards from Bucket 1, in Round 8. In addition, the former increases were larger (in absolute terms) than the latter decreases. This provides some small evidence that the results from earlier rounds were affecting behavior; while the experience of Bucket 1 being better didn't result in more Bucket 1 choices initially (in Round 7), subjects may have had a higher propensity to "believe" favorable information from Bucket 1, so that relatively high payoffs from Bucket 1 seem to cause a greater change in future Bucket 1 choices than relatively low payoffs from Bucket 1.

While the experimental results provide little support for the hypothesis that wage differentials are due to persistent incorrect prior beliefs by employers, it should be emphasized that our experiment was a severe test of the statistical discrimination model. First, real employers would have had much more time to form priors than the six periods subjects had in the experiment. Second, employers would not receive signals that the environment had changed, as our subjects did after the third and sixth rounds. Third, the change from different average productivities to same average productivities between the buckets happened abruptly; in reality, average productivities (and hence their differences) change gradually. The first two of these differences between our experiment and employers' reality probably led to weaker priors in Rounds 4 and 7 of the experiment, and the third difference probably led to faster updating from Rounds 7 to 9.[6] Therefore, we don't expect that this experiment will settle the question of the causes of wage differentials across worker types; rather, we hope that it contributes some understanding toward this issue, and that it will stimulate further research.

[6] The referee has pointed out that having more similar distributions in Rounds 1–6 would have lessened this third difference, though probably at the cost of making it more difficult for subjects to learn in these rounds that Bucket 1 contains higher average payoffs.

References

Altonji, J.G. and R.M. Blank (1999), "Race and Gender in the Labor Market," in *Handbook of Labor Economics, Volume 3C*, O. Ashenfelter and D. Card, eds., North–Holland, Amsterdam, 3143–3259.

Altonji, J.G. and C.R. Pierret (2001), "Employer Learning and Statistical Discrimination," *Quarterly Journal of Economics*, 116, 313–350.

Anderson, L. and C.A. Holt (1997), "Information Cascades in the Laboratory," *American Economic Review*, 87, 847–862.

Arrow, K. (1973), "The Theory of Discrimination," in *Discrimination in Labor Markets*, O. Ashenfelter and A. Rees, eds., Princeton University Press, Princeton, NJ, 3–33.

Becker, G. (1972), *The Economics of Discrimination*, The University of Chicago Press, Chicago.

Camerer, C. (1995), "Individual Decision Making," in *Handbook of Experimental Economics*, J. Kagel and A.E. Roth. eds, Princeton University Press, Princeton, NJ, 587–703.

Card, D. and A. Krueger (1992), "School Quality and Black–White Relative Earnings: A Direct Assessment," *Quarterly Journal of Economics*, 107, 151–200.

Farmer, A. and D. Terrell (1996), "Discrimination, Bayesian Updating of Employer Beliefs, and Human Capital Accumulation," *Economic Inquiry*, 34, 204–219.

Herrnstein, R.J. and C. Murray (1994), *The Bell Curve: Intelligence and Class Structure in American Life*, The Free Press. New York.

Lewis, D. and D. Terrell (2001), "Experience, Tenure, and the Perceptions of Employers," *Southern Economic Journal*, 67, 578–597.

Lundberg, S.J. and R. Startz (1983), "Private Discrimination and Social Intervention in Competitive Labor Markets?" *American Economic Review*, 73, 340–347.

Mincer, J. (1974), *Schooling, Experience and Earnings*, Columbia University Press, New York.

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

Siegel, S. and N. J. Castellan, Jr. (1988), *Nonparametric Statistics for the Behavioral Sciences*, McGraw–Hill, New York.

Appendix

General Instructions

You are about to participate in an experiment in the economics of decision–making. If you follow these instructions carefully and make good decisions, you might earn a considerable amount of money. If you have a question at any time, please feel free to ask the experimenter.

The Decision Task

This experimental session consists of a number of rounds. In each round, the experimenter will carry two buckets, one tan and one green. The buckets contain cards with numbers printed on them. You will be asked to draw a total of four (4) cards from these two buckets. You may choose to draw all four cards from one bucket, or you may choose to draw cards from both buckets, as long as the total number of cards you draw is exactly four. Your total revenue, measured in "points," is the sum of the numbers printed on the four cards. Your total cost depends on how you choose to draw cards from the buckets: Your profit in a round is your total revenue minus your total

Your Choice	Total Cost
Two cards drawn from each bucket	60 points
Exactly three cards drawn from one bucket	70 points
All four cards drawn from one bucket	100 points

cost. Different cards have different amounts printed on them, so your profit will be based (to some extent) on luck. However, the amounts on the cards have been chosen so that you are guaranteed to earn either zero or positive profit.

Some Information About the Cards

In each bucket will be fifty (50) cards, each with a number printed on it. Different cards within each bucket will generally have different point values. The minimum number of points on a card is 25, and the maximum is 75. The distribution of point values in one bucket may be different from that in the other bucket. In each bucket, the same set of cards will be used in each round unless you are told otherwise. Therefore, the cards you draw in early rounds will give you some information about what cards you might draw in later rounds, unless you are told that the cards have been changed. In each round, all players will choose between the same buckets with the same cards in them.

Record Keeping

You have been given a record sheet with spaces to write your choices and the resulting outcomes. In each round, circle the letters in the "Draws" columns corresponding to the color of each bucket you choose from, and below each choice, write in the number of points earned. After you have chosen all four of your cards for the round, fill in the last four columns.

Payments

You will each receive \$2.00 for participating in and completing the experiment. In addition, one round will be chosen at random from the rounds that have been played, and you will earn 5 cents for each point of profit you received in that round (100 points=\$5.00). Your earnings will be paid to you in cash at the end of the experimental session.