Discrimination: Empirics

Heidi L. Williams
MIT 14.662
Spring 2015

Outline

(1) Regression analysis (Goldberger, 1984; Neal and Johnson, 1996)

(2) Audit studies (Riach and Rich, 2002; Bertrand and Mullainathan, 2004)

(3) Quasi-experiments (Goldin and Rouse, 2000)

(4) Testing models (Charles and Guryan, 2008; Chandra and Staiger, 2010)

1 Regression analysis

Many researchers have taken an "unexplained" or residual gap between male and female wages (or black and white wages) conditional on observables as evidence of discrimination. Although such differences in residual wages are consistent with discrimination, omitted variables are always a concern, and this is also a very indirect way of testing for discrimination. Moreover, as we will discuss, some of the characteristics that researchers have conditioned on (such as education) are themselves a function of human capital investments that may be made in the context of future expectations of labor market discrimination. These types of concerns have motivated researchers to pursue alternative methods of testing for empirical evidence of discrimination.

We will first discuss Goldberger (1984)'s analysis of direct regression and reverse regression - a commonly-used approach at the time - in order to clarify what difficulties can arise with traditional regression approaches. We then focus on three alternative methods that have been the focus of more recent papers: audit studies, quasi-experimental analyses, and empirical tests of equilibrium predictions that emerge from theoretical models of discrimination.

1.1 Goldberger (1984): Direct and reverse regression

1.1.1 Definitions: Direct and reverse regression

Are men paid more than equally productive women? Let y denote an earnings variable, $\mathbf{x} = (x_1, x_2, ..., x_k)'$ denote a vector of productivity-related qualifications, and z denote a gender indicator ($z = 1$ for men and $z = 0$ for women). Suppose that the conditional expectation of earnings given qualifications and gender is:

$$E(y|x,z) = b'\mathbf{x} + az \tag{1}$$

in which the coefficient a is taken to be the discriminatory premium paid to men. Goldberger discusses how this type of model has been estimated in the academic literature (*e.g.* Oaxaca (1973)) as well as in discrimination-related law suits. The usual finding of $a > 0$ is often interpreted as evidence of salary discrimination in favor of men: among men and women with equal qualifications (equal $\mathbf{x}$), men are paid more. However, one concern is omitted productivity-relevant characteristics: if some productivity-related characteristics are unobserved by the econometrician and correlated with gender ($cov(z, \epsilon|x) \neq 0$) then estimates from this direct regression approach may be biased.

An alternative approach is to ask: are men less qualified than equally paid women? Let $q = b'\mathbf{x}$ denote the scalar index of qualifications implied by the direct regression. Now regress q on y and z:

$$E(q|y, z) \quad = \quad c^* y + d^* z \tag{2}$$

where the coefficient d^* is taken to be the excess qualifications required of men for the same salary. A finding of $d^* < 0$ would here be interpreted as evidence of salary discrimination in favor of men: among men and women receiving equal salaries (equal y), men possess lower qualifications. Following Goldberger, to make the direct and reverse regression estimates more comparable, take $a^* = \frac{-d^*}{c^*}$ as a measure of salary discrimination, so that discrimination in favor of men implies $a > 0$ and $a^* > 0$.

1.1.2 Conflicting estimates

You might expect that direct and reverse regression estimates would provide qualitatively similar estimates of discrimination: if men are paid more than equally qualified women, then they should be less qualified than equally paid women, so $a > 0$ should imply that $d^* < 0$. However, that reasoning relies on a deterministic relationship where $y = b'\mathbf{x} + az = q + az$ implies $q = y - az$, which is not likely to be true empirically given that relationships between variables are non-deterministic. In practice, direct and reverse regression approaches often give conflicting results. As a motivating example, Goldberger discusses a 1976 study of annual salaries for 199 male and 153 female faculty members at the University of Illinois which found that "on average males are paid about $2,000 more than females with the same number of publications" while "females publish about two fewer articles per five years than males who receive the same salary" - implying both a and d^* are positive. In general, reverse regression tends to suggest a lower estimate of salary discrimination (in favor of men) than does direct regression, and in fact often suggests reverse discrimination (against men).

The Goldberger paper is very clearly written, but a little short on intuition for the "big picture," so it's worth walking through the general structure of the paper before we start. Goldberger suggests that a common notion at the time was that direct regression was biased but that reverse regression was not. Goldberger first walks through an errors-in-variables model for the single qualification case x. This model clarifies that (under the strong set of assumptions outlined in that model), the reverse regression approach will lead to a downward-biased estimate of the discrimination measure α whereas the direct regression approach will lead to an upward-biased estimate of the discrimination measure α, and the direct and reverse regression estimates will give upper and lower bounds for the true parameter value. In the context of this errors-in-variables model, Goldberger shows that in the special case where salary is a deterministic function of productivity and gender, it is true that the reverse regression is unbiased while the direct regression remains biased. Goldberger then walks through a proxy variable model for the single qualification case x which illustrates that under an alternative set of assumptions, the direct regression estimate of the discrimination measure α is unbiased, whereas the reverse regression is downward biased (and may even be of the wrong sign). Thus, without knowing the underlying data generating process there is no sense in which either the direct regression approach or the reverse regression approach is *a priori* more "correct." Goldberger then discusses the multivariate case $\mathbf{x} = (x_1, x_2, ..., x_k)'$, although the results don't easily generalize to that case; we will focus on discussing the single qualification case here.

I'm not going to work through the mechanics of Goldberger's results in detail in class, but have included detailed derivations (that are hopefully much clearer than those in his paper) in this handout. His insight on this is *very* important, and comes up in a variety of contexts, so it is important for you to be familiar with.

1.1.3 An errors-in-variables model for the single qualification case

Let salary be a stochastic function of productivity and gender, where the structural parameter of interest is α - the discriminatory premium paid to men:

$$y = p + \alpha z + v \tag{3}$$

where y is salary, p is productivity, z is gender, and v is independent of z and distributed with mean 0 and variance σ_v^2. Let productivity p be an exact function of true qualifications x^*: $p = \beta x^*$, where $\beta > 0$. Let the expectation of true qualification vary by gender:

$$x^* = \mu z + u \tag{4}$$

where u is independent of z and v and distributed with mean 0 and variance σ_u^2. Note that this assumes mean differences in true qualification by gender: this is important, because the concern is that if unobserved components of productivity are correlated with gender, in a direct regression the gender variable may pick up the effects of these omitted variables. Goldberger takes $\mu > 0$ to be the empirically relevant case - that is, that the expectation of true qualification is higher for men than women. Finally, assume that measured qualification x is an imperfect indicator of true qualification in a classic errors-in-varables manner:

$$x = x^* + \epsilon \tag{5}$$

where ϵ is independent of z, u, and v and distributed with mean 0 and variance σ_ϵ^2. Note that this mis-measurement is important because otherwise we could perfectly control for true productivity x^* in our earnings regression. As Goldberger notes, imperfect measurement of qualifications was the main argument made in favor of reverse regression, which was meant to correct for measurement error by putting the measured-with-error variable on the left-hand-side rather than the right-hand-side.

The direct regression of y on x and z, $E(y|z,x) = bx + az$, gives the following:[1]

$$
\begin{aligned}
b &= \frac{cov(x,y|z)}{var(x|z)} && (6)\\
&= \frac{cov(x^* + \epsilon, p + \alpha z + v|z)}{var(x|z)} && (7)\\
&= \frac{cov(x^*, p|z)}{var(x|z)} && (8)\\
&= \frac{cov(x^*, \beta x^*|z)}{var(x|z)} && (9)\\
&= \frac{\beta var(x^*|z)}{var(x|z)} && (10)\\
&= \beta\pi^* && (11)
\end{aligned}
$$

where $\pi^* = \frac{var(x^*|z)}{var(x|z)} = \frac{\sigma_u^2}{\sigma_u^2+\sigma_\epsilon^2}$. If x is an imperfect measures of x^* (that is, if $\sigma_\epsilon^2 > 0$) then π^* will be less than one, implying that $b = \beta\pi^* < \beta$.

Taking the expectation of both sides of $E(y|z,x) = bx+az$ conditional on z, and recalling that by the law of iterated expectations $E(E(y|z,x)|z) = E(y|z)$, we have that $E(y|z) = bE(x|z)+az$. That implies that for $z = 1$, we have $E(y|z = 1) = bE(x|z = 1) + a$, which implies that $a = E(y|z = 1) - bE(x|z = 1)$. We then have:

$$
\begin{aligned}
a &= E(y|z = 1) - bE(x|z = 1) && (12)\\
&= E(p + \alpha z + v|z = 1) - bE(x^* + \epsilon|z = 1) && (13)\\
&= \alpha + E(\beta x^*|z = 1) - bE(x^*|z = 1) && (14)\\
&= \alpha + \beta E(x^*|z = 1) - bE(x^*|z = 1) && (15)\\
&= \alpha + (\beta - b)E(x^*|z = 1) && (16)\\
&= \alpha + (\beta - \beta\pi^*)E(\mu z + u|z = 1) && (17)\\
&= \alpha + (1 - \pi^*)\beta\mu && (18)
\end{aligned}
$$

[1] If you're rusty on how to derive the $b = \frac{cov(x,y|z)}{var(x|z)}$ expression, recall that with a linear conditional expectation function the best linear predictor is the same as the conditional expectation function (Goldberger (1991), page 54). Go back to write out the least squares problem where you derive the α and β to minimize $E(Y - \alpha - \beta X)^2$. The first order conditions give that $\hat{\alpha} = E(Y) - \hat{\beta}E(X)$ and $\hat{\beta} = \frac{cov(X,Y)}{var(X)}$. You can analogously derive expressions for $\hat{\alpha}$ and $\hat{\beta}$ that are conditional on Z, as in the Goldberger paper.

Given our assumptions that $\mu > 0$ and $\beta > 0$, we know that $(1-\pi^*)\beta\mu > 0$ and therefore that $\alpha < a$. Hence, under the assumptions outlined above the direct regression approach will lead to an upward-biased estimate of the discrimination measure α.

Now consider the reverse regression of x on y and z, $E(x|y,z) = cy + dz$:

$$\begin{aligned}
c &= \frac{cov(x,y|z)}{var(y|z)} && (19)\\
&= \frac{cov(x^* + \epsilon, p + \alpha z + v|z)}{var(p + \alpha z + v|z)} && (20)\\
&= \frac{cov(x^*, p|z)}{var(p + v|z)} && (21)\\
&= \frac{cov(x^*, \beta x^*|z)}{var(\beta x^* + v|z)} && (22)\\
&= \frac{cov(x^*, \beta x^*|z)}{var(\beta x^*|z) + var(v|z)} && (23)\\
&= \frac{\beta var(x^*|z)}{\beta^2 var(x^*|z) + var(v|z)} && (24)\\
&= \frac{\beta\sigma_u^2}{\beta^2\sigma_u^2 + \sigma_v^2} && (25)\\
&= \frac{1}{\beta}\frac{\beta^2\sigma_u^2}{\beta^2\sigma_u^2 + \sigma_v^2} && (26)\\
&= \frac{\pi}{\beta} && (27)
\end{aligned}$$

where $\pi = \frac{\beta^2\sigma_u^2}{\beta^2\sigma_u^2+\sigma_v^2}$. Note that π will be less than one as long as the earnings function is stochastic ($\sigma_v^2 > 0$), and will be greater than zero as long as $\sigma_u^2 > 0$. The gender coefficient is then:

$$\begin{aligned}
d &= E(x|z=1) - cE(y|z=1) && (28)\\
&= E(x^* + \epsilon|z=1) - cE(p + \alpha z + v|z=1) && (29)\\
&= E(x^*|z=1) - c\alpha - c\beta E(x^*|z=1) && (30)\\
&= (1 - c\beta)E(x^*|z=1) - c\alpha && (31)\\
&= (1-\pi)E(\mu z + u|z=1) - c\alpha && (32)\\
&= (1-\pi)\mu - c\alpha && (33)
\end{aligned}$$

The implied discrimination measure $a^* = \frac{-d}{c}$ is then $\alpha - \frac{(1-\pi)\mu}{\frac{\pi}{\beta}} = \alpha - \frac{1-\pi}{\pi}\beta\mu$ which is less than α given our assumptions that $\mu > 0$ and $\beta > 0$. Hence, under the assumptions outlined above, the reverse regression approach will lead to a downward-biased estimate of the discrimination measure α. Taken together, in this model we thus expect the direct and reverse regression estimates to bound the true parameter value above and below, respectively.

Goldberger notes that past authors had asserted that the reverse regression estimate would be unbiased while the direct regression estimate would be biased, which he shows is true in the special case where the earnings function is deterministic ($\sigma_v^2 = 0$). If $\sigma_v^2 > 0$ then $\pi = 1$, implying that $c = \frac{1}{\beta}$, $d = -c\alpha$, and $a^* = \frac{-d^*}{c^*} = \frac{\frac{1}{\beta}\alpha}{\frac{1}{\beta}} = \alpha$. On the other hand, the direct regression estimate will still be biased upwards.

1.1.4 A proxy variable model for the single qualification case

Goldberger next considers an alternative model - a proxy variable model - also for the single qualification case x. His goal is to illustrate that this reasonable alternative model implies different conclusions about biases in direct and reverse regression estimates (and, in particular, that the direct regression estimate would be unbiased in this case while the reverse regression estimate would be biased).

Suppose again that salary y is a stochastic function of productivity p and gender z: $y = p+\alpha z+v$. But instead of $p = \beta x^*$, assume that productivity is a stochastic function of measured qualifications: $p = \beta x + \epsilon$, where we assume $\beta > 0$. You can think of this case as x being a proxy variable for p, in the sense that it is an imperfect correlate. Also let the expectation of measured qualification vary by gender: $x = \mu z + u$, where we assume $\mu > 0$.

Consider the direct regression of y on x and z, $E(y|x, z) = bx + az$:

$$
\begin{aligned}
b &= \frac{cov(x, y|z)}{var(x|z)} &(34)\\
&= \frac{cov(x, p + \alpha z + v|z)}{var(x|z)} &(35)\\
&= \frac{cov(x, p|z)}{var(x|z)} &(36)\\
&= \frac{cov(x, \beta x + \epsilon|z)}{var(x|z)} &(37)\\
&= \frac{\beta var(x|z)}{var(x|z)} &(38)\\
&= \beta &(39)
\end{aligned}
$$

and

$$
\begin{aligned}
a &= E(y|z = 1) - bE(x|z = 1) &(40)\\
&= E(p + \alpha z + v|z = 1) - bE(\mu z + u|z = 1) &(41)\\
&= \alpha + E(\beta x + \epsilon|z = 1) - b\mu &(42)\\
&= \alpha + \beta E(\mu z + u|z = 1) - b\mu &(43)\\
&= \alpha + \beta\mu - \beta\mu &(44)\\
&= \alpha &(45)
\end{aligned}
$$

These expressions imply that the direct regression estimate of α is unbiased in this model.

Re-write $y = p + \alpha z + v = \beta x + \alpha z + (\epsilon + v) = \beta x + \alpha z + t$ where $t = \epsilon + v$, t is mean zero with variance σ_t^2, and t is independent of z and x. Now consider the reverse regression of x on y and z, $E(x|y,z) = cy + dz$:

$$\begin{aligned}
c &= \frac{cov(x,y|z)}{var(y|z)} && (46)\\
&= \frac{cov(x, \beta x + \alpha z + t|z)}{var(\beta x + \alpha z + t|z)} && (47)\\
&= \frac{\beta var(x|z)}{\beta^2 var(x|z) + var(t|z)} && (48)\\
&= \frac{\beta var(\mu z + u|z)}{\beta^2 var(\mu z + u|z) + var(t|z)} && (49)\\
&= \frac{\beta var(u|z)}{\beta^2 var(u|z) + var(t|z)} && (50)\\
&= \frac{\beta \sigma_u^2}{\beta^2 \sigma_u^2 + \sigma_t^2} && (51)\\
&= \frac{1}{\beta} \frac{\beta^2 \sigma_u^2}{\beta^2 \sigma_u^2 + \sigma_t^2} && (52)\\
&= \frac{\pi}{\beta} && (53)
\end{aligned}$$

where $\pi = \frac{\beta^2 \sigma_u^2}{\beta^2 \sigma_u^2 + \sigma_t^2}$. The gender coefficient is then:

$$\begin{aligned}
d &= E(x|z=1) - cE(y|z=1) && (54)\\
&= E(x|z=1) - cE(\beta x + \alpha z + t|z=1) && (55)\\
&= E(x|z=1) - c\alpha - c\beta E(x|z=1) && (56)\\
&= (1 - c\beta)E(x|z=1) - c\alpha && (57)\\
&= (1 - c\beta)E(\mu z + u|z=1) - c\alpha && (58)\\
&= (1 - \pi)\mu - c\alpha && (59)
\end{aligned}$$

The implied discrimination measure $a^* = \frac{-d}{c}$ is then $\frac{\frac{\pi}{\beta}\alpha - (1-\pi)\mu}{\frac{\pi}{\beta}} = \alpha - \frac{1-\pi}{\pi}\beta\mu$. As in the errors-in-variables model, this estimate is less than α as long as $\sigma_t^2 > 0$, given our assumptions that $\mu > 0$ and $\beta > 0$.

Thus, in the proxy variable case the direct regression is unbiased, but the reverse regression is downward biased. Note that this bias persists even if the salary function is deterministic: even with $\sigma_v^2 = 0$, $\sigma_t^2 = \sigma_\epsilon^2 + \sigma_v^2 > 0$ and thus $\pi < 1$. Note also that the bias may be large enough that the reverse regression estimate a^* may even be of the wrong sign.

1.2 Neal and Johnson (1996): Pre-market factors

In general, both the direct and reverse regression approaches are somewhat "out of style" given that it is difficult to construct data with adequate control variables for all productivity-relevant characteristics. However, one example of a relatively recent regression approach paper that remains very influential is Neal and Johnson (1996).

The basic question the Neal and Johnson paper aims to answer is: how much of the black-white earnings gap is explained by differences in skills acquired prior to labor market entry? They use the National Longitudinal Survey of Youth (NLSY) data to examine black-white wage gaps among workers in their late twenties as a function of their AFQT score at age 18 or younger.

Table 1. Their main estimates are presented in Table 1. Adding linear and quadratic variables for AFQT in Columns (3) and (6) explains around three quarters of the racial wage gap for young men, and nearly all of the racial wage gap for young women.

TABLE 1

LOG WAGE REGRESSIONS BY SEX

	MEN (N = 1,593)			WOMEN (N = 1,446)		
	(1)	(2)	(3)	(4)	(5)	(6)
Black	−.244	−.196	−.072	−.185	−.155	.035
	(.026)	(.025)	(.027)	(.029)	(.027)	(.031)
Hispanic	−.113	−.045	.005	−.028	.057	.145
	(.030)	(.029)	(.030)	(.033)	(.031)	(.032)
Age	.048	.046	.040	.010	.009	.023
	(.014)	(.013)	(.013)	(.015)	(.014)	(.015)
AFQT	...	...	.172	...	...	.228
			(.012)			(.015)
AFQT^2	...	...	−.013	...	...	.013
			(.011)			(.013)
High grade by 1991	...	.061	...	...	.088	...
		(.005)			(.005)	
R^2	.059	.155	.168	.029	.191	.165

NOTE.—The dependent variable is the log of hourly wages. The wage observations come from 1990 and 1991. All wages are measured in 1991 dollars. If a person works in both years, the wage is measured as the average of the two wage observations. Wage observations below $1.00 per hour or above $75 are eliminated from the data. The sample consists of the NLSY cross-section sample plus the supplemental samples of blacks and Hispanics. Respondents who did not take the ASVAB test are eliminated from the sample. Further, 163 respondents are eliminated because the records document a problem with their test. All respondents were born after 1961. Standard errors are in parentheses.

After documenting these estimates, Neal and Johnson present very thoughtful interpretations of several issues:

1. Is the AFQT racially biased? In 1991, the National Academy of Sciences (NAS) completed an exhaustive study with the Department of Defense of the validity of the AFQT, with special emphasis on racial fairness of the test; this NAS review found no evidence that AFQT score systematically under-predicted performance of blacks.

2. Do blacks underinvest in skill because the return is lower? Models of statistical discrimination such as Lundberg and Startz (1983) predict that the payoff to skill is lower for blacks than for whites, raising the possibility that differences in skill at the time of labor market entry arise because black youths anticipate that the returns from acquiring skills will be low. Although very intuitive, this prediction is difficult to test. Neal and Johnson present the results of a test they acknowledge to be imperfect, which is to ask whether the returns to AFQT scores in terms of wages differ by race. This regression is analogous to Table 1 but adds interactions of Black with the AFQT variables. They can't reject that the returns to skill are the same for blacks and whites. Although they conclude that "the law of one price roughly holds for skills as measured by AFQT," this test is problematic because AFQT score is an endogenous investment - ideally you would use an instrument here.

TABLE 2

TESTING FOR RACIAL DIFFERENCES IN THE RETURN TO AFQT: MEN

	All Races (N = 1,593) (1)	White (N = 825) (2)	Black (N = 466) (3)	Hispanic (N = 302) (4)
Black	−.107 (.033)	...	...	...
Hispanic	.003 (.029)	...	...	...
Age	.038 (.013)	.052 (.017)	.047 (.025)	−.014 (.035)
AFQT	.172 (.015)	.183 (.017)	.208 (.031)	.124 (.031)
AFQT^2	−.023 (.013)	−.018 (.015)	.031 (.025)	−.066 (.031)
Black × AFQT	.037 (.031)	...	...	...
Black × AFQT^2	.056 (.028)	...	...	...
R^2	.170	.155	.129	.074

NOTE.—The "all races" sample includes all men from the sample described in table 1. All respondents were born after 1961. Standard errors are in parentheses.

TABLE 3

TESTING FOR RACIAL DIFFERENCES IN THE RETURN TO AFQT: WOMEN

	All Races (N = 1,446) (1)	White (N = 726) (2)	Black (N = 428) (3)	Hispanic (N = 292) (4)
Black	.079 (.037)	...	...	...
Hispanic	.137 (.034)	...	...	...
Age	.023 (.015)	.017 (.022)	.015 (.024)	.055 (.030)
AFQT	.212 (.019)	.189 (.030)	.223 (.029)	.202 (.030)
AFQT^2	.031 (.016)	.059 (.025)	−.039 (.030)	−.025 (.029)
Black × AFQT	−.011 (.038)	...	...	...
Black × AFQT^2	−.071 (.037)	...	...	...
R^2	.168	.137	.166	.154

NOTE.—The "all races" sample includes all women from the sample described in table 1. All respondents were born after 1961. Standard errors are in parentheses.

3. What about labor market dropouts? Comparing wages among workers may be misleading if selection into employment differs for blacks and whites. Neal and Johnson present estimates from two approaches for dealing with this selection.

    - Median regressions. Under the assumption that all non-participants have wage offers below the median offer made to workers with comparable skills, the median wage gap is identified. The estimates in Table 4 suggest that less - but still much - of the black-white median wage gap is explained by AFQT.
    - Smith and Welch (1986) method. Smith and Welch (1986) observed that the mean of the wage offer distribution is a weighted average of the mean wage offers for participants and non-participants. They use this insight to derive a selection factor to get the ratio of the means of the unconditional (wage offer) distributions.

Table 5. Neal and Johnson then turn to examine determinants of AFQT scores. The raw gaps in Table 5 (men) and Table 6 (women) are large: the mean black score is one standard deviation below the mean white score for men. Using the covariates in the NLSY data, they show that controlling for family background, family size, and school quality significantly reduce these gaps. Although sizable gaps remain, this analysis suggests that "pre-market" factors may explain much of the AFQT score gap.

TABLE 5

DETERMINANTS OF AFQT: MEN

	Full Sample (N = 1,873) (1)	(2)	(3)	Valid Response to School Survey (N = 954) (4)
Black	−1.03	−.70	−.57	−.42
	(.05)	(.05)	(.05)	(.07)
Hispanic	−.70	−.31	−.22	−.02
	(.06)	(.05)	(.05)	(.08)
Mother high school graduate	...	.36	.26	.18
		(.04)	(.04)	(.06)
Mother college graduate	...	.21	.16	.09
		(.08)	(.08)	(.11)
Father high school graduate	...	.32	.25	.22
		(.05)	(.05)	(.06)
Father college graduate	...	.32	.30	.31
		(.07)	(.07)	(.09)
Mother professional	...	.20	.17	.08
		(.07)	(.07)	(.10)
Father professional	...	.26	.23	.21
		(.06)	(.06)	(.08)
Number of siblings	...	...	−.05	−.05
			(.01)	(.01)
No reading materials	...	...	−.19	−.31
			(.06)	(.09)
Numerous reading materials	...	...	.25	.27
			(.04)	(.06)
Student/teacher ratio	...	...	...	−.017
				(.006)
Disadvantaged student ratio	...	...	...	−.002
				(.001)
Dropout rate	...	...	...	−.004
				(.001)
Teacher turnover rate	...	...	...	−.005
				(.003)
R^2	.219	.382	.415	.392

NOTE.—The dependent variable is the age-adjusted AFQT score. In all specifications, the sample excludes respondents with invalid AFQT scores. In specification 4, the sample also excludes respondents with invalid responses to the school survey items employed in col. 4. Specifications 3 and 4 also include dummies for whether or not the respondent has knowledge of the educational background of his or her mother or father. Specification 4 also includes a private school dummy. The estimated coefficient is positive but not statistically significant. All background information comes from the 1979 wave of the NLSY. The dummy variables for reading materials are constructed from information about magazines, newspapers, and library cards in the home. "Numerous" means all of the above. "No" means none of the above. All respondents were born after 1961. Standard errors are in parentheses.

TABLE 6

DETERMINANTS OF AFQT: WOMEN

	Full Sample (N = 1,791)			Valid Response to School Survey (N = 926)
	(1)	(2)	(3)	(4)
Black	−.99	−.72	−.62	−.59
	(.04)	(.04)	(.04)	(.06)
Hispanic	−.77	−.45	−.37	−.30
	(.05)	(.05)	(.05)	(.07)
Mother high school graduate	...	.29	.20	.20
		(.04)	(.04)	(.06)
Mother college graduate	...	.33	.32	.24
		(.08)	(.08)	(.11)
Father high school graduate	...	.24	.18	.12
		(.04)	(.04)	(.06)
Father college graduate	...	.32	.29	.31
		(.07)	(.07)	(.09)
Mother professional	...	.15	.09	.16
		(.07)	(.07)	(.09)
Father professional	...	.15	.13	.07
		(.05)	(.05)	(.07)
Number of siblings	...	...	−.027	−.026
			(.007)	(.010)
No reading materials	...	...	−.29	−.21
			(.06)	(.08)
Numerous reading materials	...	...	.23	.23
			(.04)	(.05)
Student/teacher ratio	...	...	...	−.0043
				(.0025)
Disadvantaged student ratio	...	...	...	−.002
				(.001)
Dropout rate	...	...	...	−.003
				(.001)
Teacher turnover rate	...	...	...	−.003
				(.003)
R^2	.244	.390	.419	.431

NOTE.—See table 5.

Neal and Johnson interpret the black-white gap in AFQT scores as reflecting differences in acquired skills. They undertake several analyses that cast doubt on the alternative (extremely controversial) Herrnstein and Murray (1994) argument that AFQT is a measure of inherent ability.[2] First, the estimated racial gaps in scores are larger for older cohorts - suggesting skill investments are important. Second, estimating IV regressions that predict AFQT as a function of quarter of birth provides evidence that schooling increases AFQT scores - again suggesting skill investments are important.

The Neal and Johnson (1996) paper was very influential, primarily in suggesting that a focus solely on market discrimination is likely misplaced, and that some attention should be focused on understanding the sources of the large observed skill gaps between blacks and whites.

[2]For a more extensive critique of Herrnstein and Murray (1994), see the very thoughtful discussion and analysis in Heckman (1995).

2 Audit studies

2.1 Overview of audit studies

A *long* literature (at least four decades old) has tested for evidence of discrimination in labor, housing, and product markets by conducting 'audit' field experiments (largely but not exclusively focused on discrimination based on race and sex). Riach and Rich (2002) provide an excellent synthesis of this literature, and discuss two types of audit experiments that have been used:

1. Audit tester studies. Here, 'real' people (actors) are assigned to matched pairs and sent out to *e.g.* apply for job postings. The goal in constructing these matched pairs is hold all characteristics constant across the testers except for the characteristic of interest - say, race or gender. The matched pairs are trained in how to respond to questions, for example. Tests for discrimination typically track interview call-backs and job offers.

2. Audit resume studies. Here, matched pairs of job applications or resumes are sent out to advertised vacancies in order to test for discrimination at the initial stage of selection for interview.

The general conclusion drawn by Riach and Rich from four decades of these studies is that "...they have demonstrated pervasive and enduring discrimination against non-whites and women." That said, the audit tester studies in particular have been heavily criticized - perhaps most famously by Heckman and Siegelman (1992). Heckman and Siegelman questioned the effectiveness with which pairs can be matched, and also questioned the extent to which the testers may be unconsciously biased in favor of documenting evidence of discrimination (given that these experiments are obviously not double-blind). Moreover, the sample sizes in audit tester studies are often quite small. Despite all of these problems, the results of audit tester studies are often quite compelling. In addition, the use of audit resume studies as an alternative approach arguably overcomes many of the limitations of audit tester studies: resume studies can more easily be scaled up to larger samples, and because the 'testers' are now pieces of paper rather than individuals, it is much easier to try to hold other factors constant.

2.2 Bertrand and Mullainathan (2004)

One recent well-known audit resume study is Bertrand and Mullainathan (2004). Bertrand and Mullainathan sent over 5,000 resumes to help-wanted ads in Boston and Chicago, randomizing otherwise equivalent resumes to have African-American or White sounding names (such as Emily Walsh or Greg Baker relative to Lakisha Washington or Jamal Jones), and measuring interview callbacks for each sent resume. By experimentally varying credentials, they are also able to examine how credentials affect racial differences in callbacks.

Table 1. Table 1 documents large racial differences in callback rates, on the order of a 50 percent gap. Resumes with white names have a 9.65 percent chance of receiving a callback, whereas equivalent resumes with African-American sounding names have a 6.45 percent change of being called back. This difference in callbacks is 3.20 percentage points, or around 50 of the mean callback rate for African Americans.

TABLE 1—MEAN CALLBACK RATES BY RACIAL SOUNDINGNESS OF NAMES

	Percent callback for White names	Percent callback for African-American names	Ratio	Percent difference (*p*-value)
Sample:				
All sent resumes	9.65	6.45	1.50	3.20
	[2,435]	[2,435]		(0.0000)
Chicago	8.06	5.40	1.49	2.66
	[1,352]	[1,352]		(0.0057)
Boston	11.63	7.76	1.50	4.05
	[1,083]	[1,083]		(0.0023)
Females	9.89	6.63	1.49	3.26
	[1,860]	[1,886]		(0.0003)
Females in administrative jobs	10.46	6.55	1.60	3.91
	[1,358]	[1,359]		(0.0003)
Females in sales jobs	8.37	6.83	1.22	1.54
	[502]	[527]		(0.3523)
Males	8.87	5.83	1.52	3.04
	[575]	[549]		(0.0513)

Notes: The table reports, for the entire sample and different subsamples of sent resumes, the callback rates for applicants with a White-sounding name (column 1) an an African-American-sounding name (column 2), as well as the ratio (column 3) and difference (column 4) of these callback rates. In brackets in each cell is the number of resumes sent in that cell. Column 4 also reports the *p*-value for a test of proportion testing the null hypothesis that the callback rates are equal across racial groups.

Table 4. Table 4 documents that race changes the returns to having a better resume: having a higher-quality resume has a smaller effect for African-Americans relative to Whites.

TABLE 4—AVERAGE CALLBACK RATES BY RACIAL SOUNDINGNESS OF NAMES AND RESUME QUALITY

	Panel A: Subjective Measure of Quality (Percent Callback)			
	Low	High	Ratio	Difference (p-value)
White names	8.50	10.79	1.27	2.29
	[1,212]	[1,223]		(0.0557)
African-American names	6.19	6.70	1.08	0.51
	[1,212]	[1,223]		(0.6084)
	Panel B: Predicted Measure of Quality (Percent Callback)			
	Low	High	Ratio	Difference (p- value)
White names	7.18	13.60	1.89	6.42
	[822]	[816]		(0.0000)
African-American names	5.37	8.60	1.60	3.23
	[819]	[814]		(0.0104)

Notes: Panel A reports the mean callback percents for applicant with a White name (row 1) and African-American name (row 2) depending on whether the resume was subjectively qualified as a lower quality or higher quality. In brackets is the number of resumes sent for each race/quality group. The last column reports the p-value of a test of proportion testing the null hypothesis that the callback rates are equal across quality groups within each racial group. For Panel B, we use a third of the sample to estimate a probit regression of the callback dummy on the set of resume characteristics as displayed in Table 3. We further control for a sex dummy, a city dummy, six occupation dummies, and a vector of dummy variables for job requirements as listed in the employment ad (see Section III, subsection D, for details). We then use the estimated coefficients on the set of resume characteristics to estimate a predicted callback for the remaining resumes (two-thirds of the sample). We call "high-quality" resumes the resumes that rank above the median predicted callback and "low-quality" resumes the resumes that rank below the median predicted callback. In brackets is the number of resumes sent for each race/quality group. The last column reports the p-value of a test of proportion testing the null hypothesis that the callback percents are equal across quality groups within each racial group.

One concern is whether Bertrand and Mullainathan's experimental treatment manipulates perceptions of social class above and beyond perceptions of race. Although not definitive in ruling out this possibility, using birth certificate data on mother's education for the first names used in their sample they find little relationship between social background and name-specific callback rates.

In Section V.C, Bertrand and Mullainathan relate the findings of their experiment to the predictions of taste-based and statistical models of discrimination, and argue that neither set of models is able to fully explain the pattern of results they find. One critique of the Bertrand and Mullainathan results is that their randomization essentially assumes a model in which workers randomly search for firms, as opposed to using a more directed search strategy that targets certain employers that are thought to be less discriminatory.

In a related study, Fryer and Levitt (2004) use birth certificate data to investigate the relationship between Black names and a wide range of life outcomes, controlling for background characteristics. They find no compelling evidence of a relationship. Although seemingly in conflict with audit study results like those of Bertrand and Mullainathan, there are at least three interpretations of the data that can reconcile these two sets of results:

1. Black names are used as signals of race by discriminatory employers at the resume stage, but are unimportant once an interview reveals the candidates race

2. Black names provide a useful signal to employers about labor market productivity after controlling for information on the resume

3. Black names themselves have a causal impact on job callbacks and unemployment duration that Fryer and Levitt are unable to detect due to e.g. measurement error in their data

3 Quasi-experiments

Perhaps the best-known quasi-experimental study of discrimination is Goldin and Rouse (2000). For many years, symphony orchestras in the US conducted non-blind auditions, but over time some orchestras changed to use screens during solo auditions to hide the identity of the performers. Historically, many viewed women as unsuitable for orchestras: Goldin and Rouse quote one conductor as saying "I just don't think women should be in an orchestra." This raises the question of whether a blind audition could eliminate the possibility of discrimination against women and increase the number of women in orchestras.

Figures 1 and 3. Goldin and Rouse collected audition records from major symphony orchestras dating back as far as 1940. Figure 1 shows the remarkable change in the gender composition of these orchestras over time. Particularly because turnover is quite low (documented in Figure 2, not shown here), the proportion of new players who were female must have been quite high; they document this directly in Figure 3.

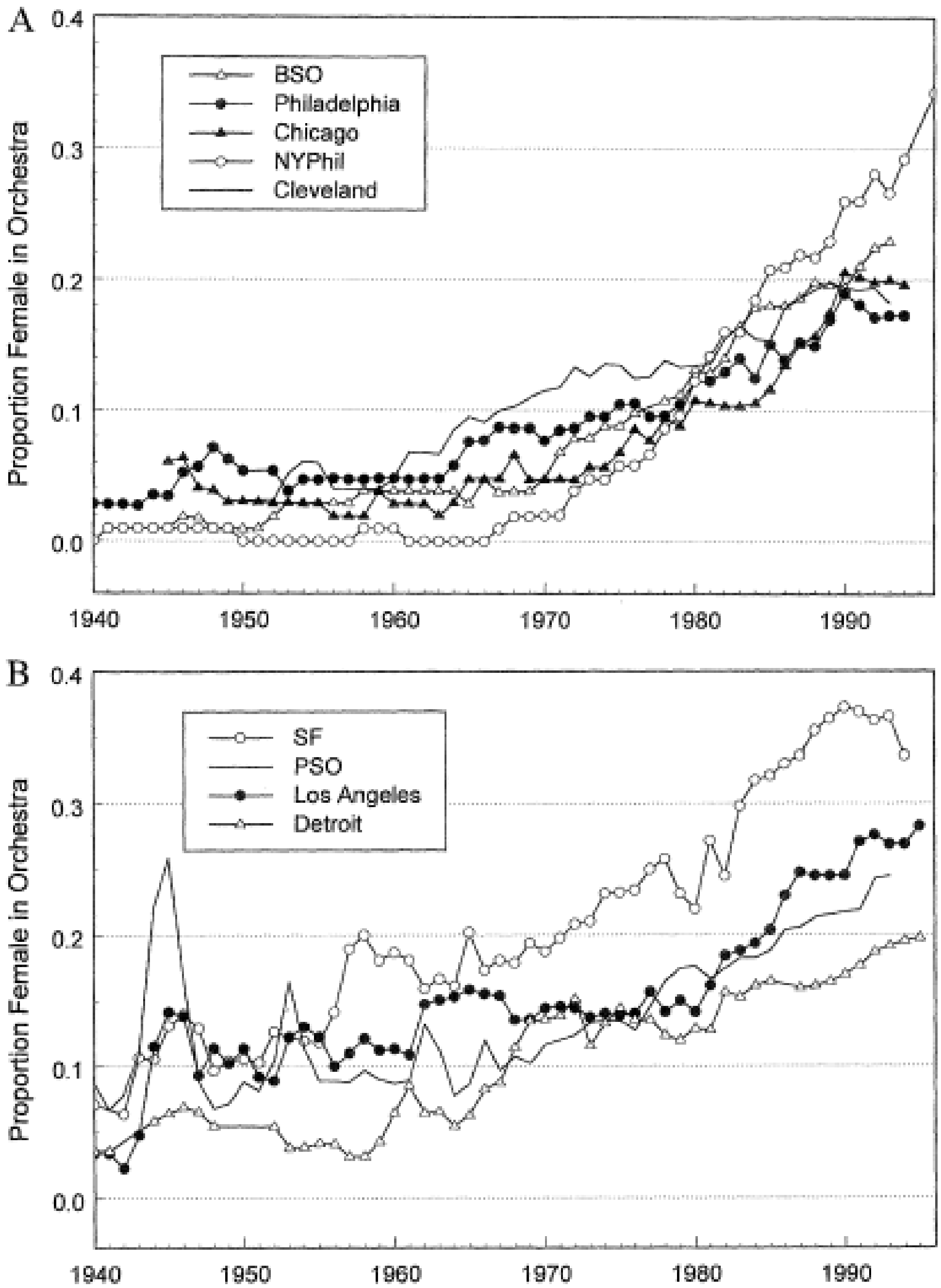

FIGURE 1. PROPORTION FEMALE IN NINE ORCHESTRAS, 1940 TO 1990'S
A: THE "BIG FIVE"; B: FOUR OTHERS

Source: Roster sample. See text.

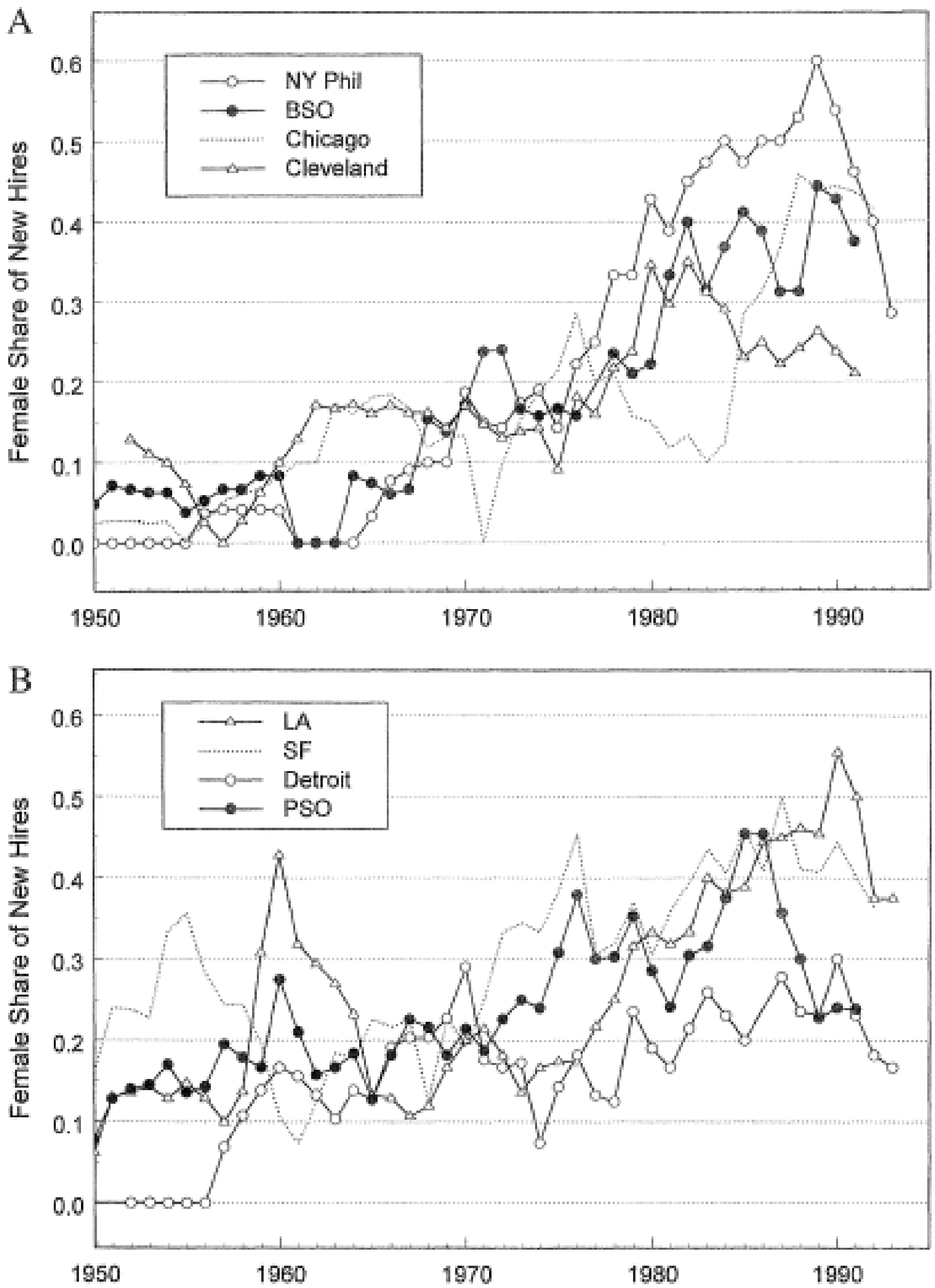

FIGURE 3. FEMALE SHARE OF NEW HIRES IN EIGHT ORCHESTRAS, 1950 TO 1990's
A: FOUR OF THE "BIG FIVE"; B: FOUR OTHERS

Source: Roster sample. See text.

Table 2. Goldin and Rouse examine the effects of adoption of blind auditions in a differences-in-differences framework where they use individual fixed effects to observe the same people in blind and non-blind auditions. One question they tackle is whether symphony-level adoption of blind auditions is endogenous. Table 2 estimates a probit model of screen adoption by year, conditional on not having previously adopted a screen. The proportion female in an orchestra has a positive coefficient, but it is small and not statistically significant. The tenure variable suggests that the stability of personnel increases the likelihood of screen adoption.

TABLE 2—ESTIMATED PROBIT MODELS FOR THE USE OF A SCREEN

	Preliminaries blind		Finals blind
	(1)	(2)	(3)
(Proportion female)$_{t-1}$	2.744	3.120	0.490
	(3.265)	(3.271)	(1.163)
	[0.006]	[0.004]	[0.011]
(Proportion of orchestra personnel with <6 years tenure)$_{t-1}$	−26.46	−28.13	−9.467
	(7.314)	(8.459)	(2.787)
	[−0.058]	[−0.039]	[−0.207]
"Big Five" orchestra		0.367	
		(0.452)	
		[0.001]	
pseudo R^2	0.178	0.193	0.050
Number of observations	294	294	434

Notes: The dependent variable is 1 if the orchestra adopts a screen, 0 otherwise. Huber standard errors (with orchestra random effects) are in parentheses. All specifications include a constant. Changes in probabilities are in brackets. "Proportion female" refers to the entire orchestra. "Tenure" refers to years of employment in the current orchestra. "Big Five" includes Boston, Chicago, Cleveland, New York Philharmonic, and Philadelphia. The data begin in 1947 and an orchestra exits the exercise once it adopts the screen. The unit of observation is an orchestra-year.
Source: Eleven-orchestra roster sample. See text.

Table 7. The authors present results from a number of specifications; Table 7 estimates models for the three orchestras that changed policies over their time period, allowing them to include orchestra and year fixed effects in their specification. These models are less precise than some of their other specifications which include orchestras who did not change policies over their time period, but communicate the same spirit of results. Two interesting results emerge from this table. First, column (3) estimates this model without individual fixed effects, and finds evidence that on average women do worse with blind auditions. However, this could be explained by a compositional bias if the presence of blind auditions induces more women to audition, and if the marginal female candidates are weaker than the average female candidate that auditioned when auditions were non-blind. Second, columns (1) and (2) estimates this model with individual fixed effects, which control for this type of compositional change. Here, there is evidence that women benefitted from blind auditions.

TABLE 7—LINEAR PROBABILITY ESTIMATES OF THE LIKELIHOOD OF BEING ADVANCED: WITH INDIVIDUAL AND ORCHESTRA FIXED EFFECTS

	Include individual fixed effects		Exclude individual fixed effects
	(1)	(2)	(3)
Blind	0.404	0.399	0.103
	(0.027)	(0.027)	(0.018)
Female × Blind	0.044	0.041	−0.069
	(0.039)	(0.039)	(0.022)
Female			−0.005
			(0.019)
p-value of H_0: Blind + (Female × Blind) = 0	0.000	0.000	0.090
Individual fixed effects?	Yes	Yes	No
Orchestra fixed effects?	No	Yes	Yes
Year fixed effects?	Yes	Yes	Yes
Other covariates?	Yes	Yes	Yes
R^2	0.615	0.615	0.048
Number of observations	8,159	8,159	8,159

Notes: The unit of observation is a person-round. The dependent variable is 1 if the person is advanced to the next round and 0 if not. Standard errors are in parentheses. All specifications include an interaction for the sex being missing and a blind audition; "Other covariates" include automatic placement, years since last audition, number of auditions attended, size of the audition round, proportion female in audition round, whether a principal or substitute position, and a dummy indicating whether years since last audition is missing. These regressions include only the orchestras that changed their audition policy during our sample years and for which we observe individuals auditioning for the audition round both before and after the policy change. These regressions include 4,836 separate persons and are identified off of 1,776 person-rounds comprised of individuals who auditioned both before and after the policy change for a particular orchestra.
Source: Eight-orchestra audition sample (three orchestras of which are used; see Notes). See text.

The headline estimate is that blind auditions increase the relative probability that women advance from the preliminary round by 50%, and have a larger effect on the relative probability that women are hired in the final round; the one puzzling estimate is that blind auditions appear to lower the relative probability that women advance from a semi-final round. Taken as a whole, the Goldin-Rouse results suggest that blind auditions reduced discrimination against women and can explain a large share of the time-series increase in the share female of orchestras since 1970.

Although fascinating, the Goldin-Rouse results are not able to distinguish between taste-based or statistical discrimination in non-blind additions. They are also not able to examine the question of whether the same individuals perform differently with versus without a screen. Finally, if you were writing this today it would be nice to show an event study graph around the time of screen adoption.

4 Testing models

As discussed above, many of the papers documenting evidence of discrimination - such as Bertrand and Mullainathan (2004) and Goldin and Rouse (2000) - are unable to provide evidence on whether taste-based or statistical discrimination is a mechanism behind their results. Several recent papers have either explicitly tested one model (such as Charles and Guryan (2008)) or tested between models (such as List (2004) and Chandra and Staiger (2010)).

4.1 Testing Becker: Charles and Guryan (2008)

The Charles and Guryan (2008) paper gathers data and tests the key testable predictions of the Becker taste-based discrimination model:

1. The prejudice of the marginal employer matters more than the prejudice of the average employer for the relative wage difference

2. The number (or fraction) of blacks in the workforce is positively related to racial wage gaps, holding prejudice constant

3. Prejudice in the right tail of the employer prejudice distribution should not matter for racial differences whereas higher prejudice in the left tail of the prejudice distribution should affect racial wage gaps

4. The mechanism that generates these patterns is the tendency of the market to segregate blacks from the most prejudiced whites

Table 3. A key innovation in this paper is to combine 'standard' measures of the residual wage gap from the CPS with 'direct' measures of prejudice from multiple waves of the General Social Survey. Their key results are presented in Table 3. First, they find that prejudice of the 'marginal' white (defined as the white at the p^{th} distribution of prejudice, where p equals the fraction of workers in a state that is black) is much more strongly predictive of racial wage gaps than is the average prejudice. Second, they find that the 10^{th} percentile of the prejudice distribution (the least prejudiced decile) is highly predictive of the wage gap, whereas the 50^{th} and 90^{th} percentiles are not.

TABLE 3
ESTIMATED RELATIONSHIP BETWEEN RACIAL PREJUDICE OF WHITES IN A LABOR MARKET AND BLACK-WHITE RELATIVE WAGES
Dependent Variable: Residual Black-White Wage Gap in Market

Measure of Prejudice among All Whites	(1)	(2)	(3)	(4)	(5)	(6)
Average	−.036		.097	.050		
	(.030)		(.029)	(.033)		
Marginal		−.213	−.328	−.202		
		(.040)	(.050)	(.068)		
10th percentile					−.212	−.292
					(.180)	(.125)
Median					−.006	.007
					(.062)	(.043)
90th percentile					.016	.016
					(.029)	(.020)
Fraction black				−.157		−.304
				(.062)		(.045)
State	45	45	45	45	45	45
R^2	.03	.40	.52	.59	.05	.56

NOTE.—The table reports coefficients (standard errors) from OLS regressions of residual state-level black-white wage gaps on various measures of prejudice among all whites (the mean of the black-white wage gap across states is −0.123, and the standard deviation is .044). Residual black-white wage gaps are estimated using 1977–2002 May/MORG CPS data and control for education, a quadratic in experience, race-specific year effects, and state effects. Data from 1973–76 are dropped because the CPS reports states in groups in those years. States are dropped if they are not sampled in the GSS in the years necessary to measure the marginal index of prejudice. The "marginal" is the pth percentile of the prejudice distribution of the relevant population of whites, where p is the fraction of the population that is black. See the text for details.

Charles and Guryan present a series of robustness checks. Although not definitive, the results of this paper are (perhaps surprisingly) quite supportive of the Becker model.

4.2 Health care: Chandra and Staiger (2010)

A gigantic literature in both medicine and social science has documented evidence of disparities in health care treatment and health outcomes. Chandra and Staiger (2010) attempt to build on this literature by testing whether the data (from one context) looks more consistent with taste-based discrimination or more consistent with statistical discrimination.

Under taste-based prejudice, providers (consciously or unconsciously) use a higher benefit threshold for providing care to minority patients (for example, recommending a treatment to non-minority patients if it prolongs their life by at least three months, but only treating minority patients if it prolongs their life by at least five months). This form of prejudice implies that returns to the marginal minority patient receiving treatment will be higher than the returns to the marginal non-minority patient receiving treatment. Alternatively, under statistical discrimination, it may be that membership in a minority group predicts lower benefit from treatment because of *e.g.* a statistical association with follow-up care, and that physicians take this correlation into account when allocating treatments. Both models result in minority groups receiving less treatment, but the statistical discrimination model implies that "under-treatment" of minorities may be optimal given the current state of the world.

In the absence of prejudice, two patients receiving treatment who have the same propensity to get the treatment (as measured by clinical characteristics) should have the same expected benefit from the treatment; if there is prejudice, the treatment-on-the-treated effect should be larger for minorities (conditional on the propensity to get the treatment). This test is similar in spirit to Knowles, Persico and Todd (2001), who analyze racial bias in motor vehicle searches. Chandra and Staiger test this prediction using data on treatments for heart attacks (the setting is very similar to their earlier paper that we discussed, Chandra and Staiger (2007)).

Chandra and Staiger's results do not provide evidence of taste-based discrimination: if anything, women and minorities appear to have slightly *smaller* benefits from treatment relative to men and whites. Section VI of their paper discusses several potential explanations for their results. Although they are not able to definitely rule out other explanations, in the end their results appear most consistent with a model of statistical discrimination in which treatment disparities exist because of lower minority appropriateness for treatment. The mechanism for this finding that minorities and women are less appropriate for treatment is key to both interpreting their findings and to the public policy relevance of their work, but is not well understood on the basis of this paper.

References

Bertrand, Marianne and Sendhil Mullainathan, "Are Emily and Greg more employable than Latisha and Jamal? A field experiment on labor market discrimination," *American Economic Review*, 2004, *94* (4), 991–1013.

Chandra, Amitabh and Douglas Staiger, "Productivity spillovers in health care: Evidence from the treatment of heart attacks," *Journal of Political Economy*, 2007, *115* (1), 103–140.

_ **and** _ , "Identifying provider prejudice in healthcare," 2010. NBER working paper #16832.

Charles, Kerwin and Jonathan Guryan, "Prejudice and wages: An empirical assessment of Becker's The Economics of Discrimination," *Journal of Political Economy*, 2008, *116* (5), 773–809.

Fryer, Roland and Steven Levitt, "The causes and consequences of distinctively Black names," *Quarterly Journal of Economics*, 2004, *119* (3), 767–805.

Goldberger, Arthur, "Reverse regression and salary discrimination," *Journal of Human Resources*, 1984, *19* (3), 293–318.

_ , *A Course in Econometrics*, Harvard University Press, 1991.

Goldin, Claudia and Cecilia Rouse, "Orchestrating impartiality: The impact of 'blind' auditions on female musicians," *American Economic Review*, 2000, *90* (4), 715–741.

Heckman, James, "Lessons from the Bell Curve," *Journal of Political Economy*, 1995, *103* (5), 1091–1120.

_ **and Peter Siegelman**, "The Urban Institute Audit Studies: Their methods and findings," in Michael Fix and Raymond Struyk, eds., *Clear and Convincing Evidence: Measurement of Discrimination in America*, 1992, pp. 187–258.

Herrnstein, Richard and Charles Murray, *The Bell Curve: Intelligence and Class Structure in American Life*, Free Press, 1994.

Knowles, John, Nicola Persico, and Petra Todd, "Racial bias in motor vehicle searches: Theory and evidence," *Journal of Political Economy*, 2001, *109* (1), 203–232.

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

Lundberg, Shelly and Richard Startz, "Private discrimination and social intervention in competitive labor markets," *American Economic Review*, 1983, *73* (3), 340–3347.

Neal, Derek and William Johnson, "The role of premarket factors in black-white wage differences," *Journal of Political Economy*, 1996, *104* (5), 869–895.

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

Riach, Peter and Judith Rich, "Field experiments of discrimination in the market place," *Economic Journal*, 2002, *112* (483), F480–F518.

Smith, James and Finis Welch, "Closing the gap: Forty years of economic progress for blacks," 1986. RAND Report R-3330-DOL.