Chapter Five

THE URBAN INSTITUTE AUDIT STUDIES: THEIR METHODS AND FINDINGS

James J. Heckman and Peter Siegelman

INTRODUCTION

Policy discussions often revolve around concepts that defy precise definition or measurement. The current controversy over the prevalence of discrimination by race, ethnicity, and gender and the remedies for such discrimination is fueled by the lack of hard evidence. Most people agree with the definition of labor market discrimination. It occurs if persons in one group with the same relevant productivity characteristics as persons in another group are treated unfavorably by the labor market solely as a consequence of their demographic status, which is assumed to be irrelevant to productivity.

A large and sometimes polemical literature has emerged about what characteristics of persons are relevant to their productivity and how they can be measured. Especially problematic is the possibility of statistical discrimination that may arise if the same levels of observed characteristics convey different information about true productivity for different demographic groups. In order to assess true productivity, we need to acquire much more information about individuals and jobs than is generated in standard data on labor market transactions. The available data on the operation of the labor market are meager and unsatisfactory; so is our understanding of the prevalence and sources of discrimination.

The current empirical literature in labor economics focuses most of its attention on widely available wage data. An enormous literature summarized in Cain (1986) documents the existence of demographic wage differentials even after controls are made for available measured "productivity" characteristics. Because of the absence of standardized, economywide data on hiring, promotion, and firing decisions, there is much less evidence on discrimination in these important dimensions of labor market activity. What we know about hiring and

promotion mostly comes from court cases or selected studies of firms, with their attendant uncertain generality.

Audit studies are a potentially promising method for extending our understanding of hiring discrimination. Although such studies can overcome some of the limits inherent in traditional analyses of discrimination, they also pose a number of important and subtle challenges. Both the generation of the audit pair data and its interpretation need to be conducted with extreme care if the potential usefulness of hiring audits is to be realized.

Despite suggestive rhetoric to the contrary, audit pair studies are not experiments or matched pair studies. Race or ethnicity cannot be assigned by randomization or some other device as in experimental or matched pair analyses. Race is a personal characteristic and adjustments must instead be made on "relevant" *observed* characteristics to "align" audit pair members. Some characteristics can be controlled by the audit analyst: e.g., which firms to sample. Audit pair studies are, then, a variant of statistical matching in which some of the match characteristics can be manipulated. Because there is partial control on auditor characteristics, such studies may improve on matching methods. But audit pair studies suffer from the main defect of matching methods: they do not account for unobservables that affect outcomes. Indeed, as we document below, matching on observables may exacerbate the problem of non-alignment of audit pairs by accentuating differences in unobservables among audit pair members. Audit pair evidence of the sort reported in this volume is not necessarily clear or convincing.

This chapter critically examines the methods and findings of two recent Urban Institute studies and a study of the Denver labor market patterned after them.[1] For brevity, the Urban Institute black/white study will henceforth be denoted UIBW, the Urban Institute Anglo/Hispanic study will be denoted UIAH, and the Denver study will be called just that. In this chapter, we draw heavily on our recent research on audit pair methodology (Heckman and Siegelman 1993). In that work, we develop explicit models of employment and use them to define discrimination. Both large sample and exact small sample tests are developed to analyze audit pair data. We also use this apparatus to analyze the available data. Our framework can also be used to design "optimal" audit pair studies. A major methodological finding of our research is the potentially misleading character of large sample statistical methods that are widely used to design and interpret audit pair studies. Conventional large sample methods, if correctly adapted to small samples, produce less evidence of discrimination

than is found when optimal exact small sample tests are used. If taken literally, conventional large sample methods produce misleading advice on optimal experimental design. Here we communicate the main ideas in our work.

The rest of this chapter is executed in seven sections. The second section presents a description of the audit pair methodology. The third section considers the potential of the UI studies for understanding discrimination in hiring. The fourth section presents the evidence set forth in the UI and Denver studies and an intuitive discussion of alternative measures of discrimination. We present large sample statistical tests for the absence of discrimination.

Using a variety of tests, and the measure of discrimination that seems most satisfactory, we find little evidence of discrimination against Denver blacks and Denver Hispanics. For Chicago blacks, the evidence in support of discrimination is at best marginal. For Washington, D.C. blacks and Chicago and San Diego Hispanics, our tests reveal evidence of what might be termed discrimination.

In the fifth section we subject this evidence to further scrutiny. We question the representativeness of the UI sampling frame. We raise concern about "experimenter effects," that is, inadvertent contamination of the audit studies by motivational devices. We also point out an important ambiguity that plagues the UI Anglo/Hispanic study: that all Hispanics had accents and facial hair while none of the Anglos possessed these characteristics.

In the sixth section we show that the UI studies do not shed much light on the magnitude of the wage gap or on the source of the 25-year decline in the labor force participation rate for black males relative to white males. We go on to note that the evidence in the Denver and UI studies does not support the claim that affirmative action has led to reverse discrimination, at least in low-wage labor markets.

We further note that the Denver and UI studies appear to differ in the extent of homogeneity across audit pairs. The Denver pairs are not comparable with each other. Since both Denver and UI claim to match testers within and across pairs at comparable levels, this conflict in the evidence demonstrates the difficulty in objectively verifying the quality of the matches, on which the employment audit method depends.

In the seventh section, we summarize a formal model of firms' employment decisions developed in appendix 5.D that enables us to assess the intuitive definitions of labor market discrimination. We establish the conditions required to justify them. Appendix 5.D is a condensed version of our companion paper. We make two important

observations about the limitations of audit pair methodology: (1) Standardization on *observed* productivity characteristics may either accentuate or attenuate measured racial differences in hiring rates compared with what occurs in actual labor markets; (2) The choice of a particular observed productivity level at which to standardize may produce much more (or less) evidence of discrimination than standardization at another level. More thought should go into the choice of a standardization level. A variety of levels across audit pairs should be used to present a more accurate picture of actual labor market hiring practices. The chapter concludes with suggestions for future research.

HOW AUDITS WORK

Audits have been adopted by social scientists from techniques employed by legal activists, who pioneered their use in the enforcement of fair-housing laws during the late 1960s. The audit procedure can be conveniently divided into two parts.[2] First is the selection and training of auditors. Groups of two individuals, one white or Anglo and one black or Hispanic, are selected from a group of applicants to resemble each other as closely as possible except for race. Testers are typically matched on such attributes as age, education, physical appearance (subjective level of attractiveness), physical strength, and level of verbal skills, as deemed relevant.[3] The goal is to produce pairs of testers who are identical in all *relevant* characteristics so that any systematic difference in treatment within each pair can be attributed only to the effects of race.[4] In addition to their outward similarities, testers are given training about how they are supposed to behave during the course of the audits. Such training typically includes developing synthetic biographies (current and past employment, references, education, and so on), behavioral alignment (e.g., level of aggressiveness and overall "presentation of self"), and experience in role-playing, simulating the kind of transaction being audited.

An important element of subjective judgement enters at this stage. Audit pair analysts assume that they know which characteristics are relevant to employers, and when such characteristics are "sufficiently" close to make majority and minority audit pair members "indistinguishable." Audit pair members must be matched on each of the relevant characteristics. Alternatively, audit analysts assume that they know how employers trade off characteristics. For the housing market,

where the original audit studies were conducted, fewer characteristics (wealth, income, credit status, etc.) are essential attributes of purchasers. For the case of the labor market, many more characteristics are likely to be relevant and different employers are likely to place different weight on those characteristics. Thus audit methods seem less well-suited to the labor market than to the housing market.

Given the current low level of factual knowledge about which characteristics employers value and how attributes trade off in productive content, and given the likely heterogeneity among employers in making these assessments, it is not obvious that audit analysts would possess the relevant information required to make perfect matches. There is a presumption of knowledge about "what is really important" that is difficult to demonstrate objectively. This inability to defend, or even fully enunciate, the criteria used to match audit pair members constitutes the Achilles heel of the audit pair methodology. In the UIBW study five audit pairs were chosen in Chicago and Washington from a group of male college students between the ages of 19 and 24 at the major universities in those environments (excluding junior colleges and community colleges) who applied to perform the audits. Job announcements were mailed to university employment and placement offices, social science departments, minority affairs offices, and selected professors. There were 23 applicants in Washington and 31 in Chicago, who were winnowed to five audit pairs in each site. The choice among potential audit pair partners was thus rather limited. One of two implicit assumptions must be made at this stage. Either there are (many) fewer than thirty relevant characteristics of workers valued by employers, so that exact matches can be formed among potential auditors, or else UI analysts know the isoproductivity tradeoff curves of firms which are assumed to be identical at all places of employment.

The small size of the list of potential match partners, coupled with the likely large number of relevant productivity attributes, makes the probability of five successful matches rather low. To take an example, suppose that as in Washington there are 10 potential white match partners for 10 potential black partners. Suppose that there are five productivity attributes that describe each worker, each independently distributed as binomial with probability of one-half for each value of each attribute. (The case of discrete attributes is most favorable to matching.) Black and white distributions are identical. Then the number of expected successful matches is less than 2.7, whereas UI reports 5 such matches in Washington. If the number of attributes is increased to 10, the expected number of successful matches is less than 0.1. It

seems unlikely that five exact matches were found in Washington given that there are likely to be more than 10 characteristics valued by employers, that some characteristics are continuously distributed, and that black and white productivity distributions are likely to be different.

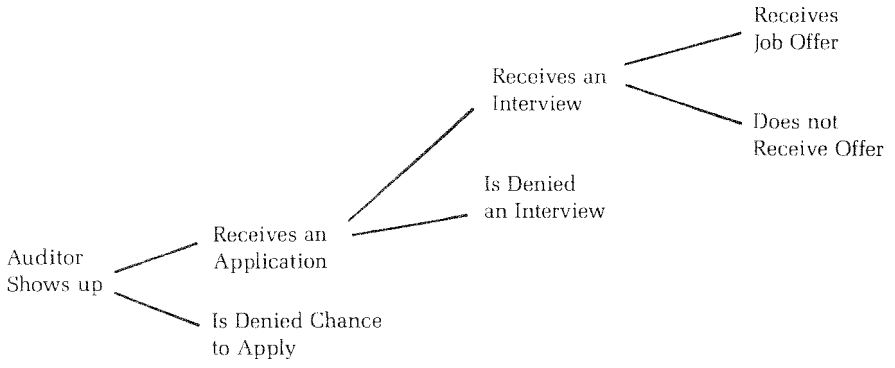
Payment was made to applicants for a fixed sum of $3,000 for six weeks of work. It was not made contingent on performance in the audit study. Each audit pair partner applied without knowledge of the employment outcomes of the other partner.

The second phase of an audit study is the generation of the data. Job openings to be audited are selected at random for certain types of entry level jobs sampled from help-wanted advertisements in local newspapers. Members of an audit pair are then sent in random order to apply for the job, typically within a few hours of each other. (If the first member of a pair is offered a job, he is instructed to turn down the offer so as to leave the vacancy available for his teammate.) Each pair typically conducts many tests, so repeated observations are available for each person in each pair. However, in these studies, the same firms are not visited by more than one audit pair.

The data generated by these tests document the way each member of an audit pair was treated in the process of applying for a job. In the employment context, the possible outcomes might look like those described in figure 5.1.

Outcomes are typically limited to only a few, discrete possibilities and there is usually only a small sample of observations for each audit pair. Statistical analyses must cope with the small numbers in these samples.

Figure 5.1 A SCHEMATIC REPRESENTATION OF OUTCOMES IN AN EMPLOYMENT AUDIT



Auditors often collect data on aspects of employer treatment apart from employment offers. Such supplemental data might include the amount of time the auditor had to wait for an interview, the kinds of remarks interviewers made, or other subjective features of interviewee treatment. While such data can be enlightening, and are especially useful in illuminating potential *causes* of discriminatory treatment (Yinger 1986), we concentrate on the analysis of the bottom line employment variable.

HOW THE AUDIT METHOD MIGHT IMPROVE OUR KNOWLEDGE OF LABOR MARKET DISCRIMINATION

Virtually all subsequent analyses of discrimination follow Gary Becker's *The Economics of Discrimination* ([1957] 1975), and examine employment segregation and wage differentials. They do not explicitly consider the hiring process that is the focus of the studies under review here.[5] One crucial benefit of audit studies is that they offer a chance to examine an important aspect of labor market behavior—discrimination in hiring—that has been largely inaccessible to social scientists until now.

The other major advantage of the audit technique is that it allows more control over the characteristics that are thought to be relevant to the employment decision than is possible in conventional ex-post regression analyses. For example, regression studies typically use years of education as a control variable in explaining wage discrimination. But this is an extremely crude control, ignoring as it does differences in educational quality and performance between workers with the same number of years of education. In an audit, by contrast, the two testers can be matched exactly on certain characteristics (by giving them identical educational histories, including schools attended, GPA, and so forth), providing a much cleaner measure of the demand-side response to race and ethnicity than techniques based on passive observation. In addition, by sending pairs of auditors to the same firms, one gains partial control over idiosyncratic differences in firm valuations of common bundles of characteristics that plague ordinary observational studies. In Heckman and Siegelman (1993), we demonstrate the gains in statistical power that accrue from sending pairs to the same firm rather than sending team members to separate firms. Eliminating common unobserved components makes it possible to construct better tests of the hypothesis of no discrimination.

Evidence for Discrimination in Hiring

Several measures of discrimination were developed in the UI studies, and several measures were not used that might have been. This section presents the main empirical findings. We then discuss ways of measuring discrimination, comment briefly on their theoretical justification, and show how they operate when applied to the audit data under consideration. A more extensive theoretical analysis of these methods and definitions is reserved for later sections of this paper.

Discrimination in What?

Table 5.1 presents summaries of outcomes for each black/white audit pair in Washington, D.C., and Chicago. Table 5.2 presents comparable summaries from the UIAH study, in which similar definitions are used. Table 5.3 presents the aggregate data from Denver.[6] See table 5.10 for the disaggregated data.

The first question one needs to ask in analyzing these data is, "What constitutes an 'outcome' that exhibits discrimination?" In one context, the question becomes whether one is interested only in differences within pairs in the rate of job offers, or whether one also cares about disparities in getting interviews or opportunities to apply—what we call "favorable audits."[7] Additionally, incidental treatment (what the UIBW study refers to as "opportunities-diminished" variables)—such as length of waiting time for an application, length of interview, number of interviewers, or positive and negative comments made during an interview—can be used to measure discrimination. Based on these variables, UIBW constructed a composite index for each auditor in each audit, with differences in this index between auditors indicative of discrimination. Given the ambiguity in interpreting an index, and the clear bottom-line nature of a job offer, we focus on the "get-a-job" measures of discrimination in this chapter.[8]

Unequal Treatment of "Identical" Pairs as a Measure of Discrimination

One intuitively plausible measure of the existence of discrimination is the proportion of times that two members of an audit pair who are identical (except for ethnicity or race) are treated differently by potential employers. This measure cannot be formalized into a statistical test until we have a precise measure or range of measures specifying whether any given sample proportion is large or small. Nevertheless,

Table 5.1 OUTCOMES IN THE URBAN INSTITUTE BLACK/WHITE STUDY IN CHICAGO AND WASHINGTON, D.C. (Get Job or Not)

Number of Audits	Pair	(a) Both get job	(b) Neither gets job	a + b	White yes, black no	White no, black yes
Chicago						
35	1	(5) 14.3%	(23) 65.7%	80%	(5) 14.3%	(2) 5.7%
40	2	(5) 12.5%	(25) 62.5%	75%	(4) 10.0%	(6) 15%
44	3	(3) 6.8%	(37) 84.1%	90.9%	(3) 6.8%	(1) 2.3%
36	4	(6) 16.7%	(24) 66.7%	83.4%	(6) 16.7%	(0) 0%
42	5	(3) 7.1%	(38) 90.5%	97.6%	(1) 2.4%	(0) 0%
197	Total	(22) 11.2%	(147) 74.6%	85.8%	(19) 9.6%	(9) 4.5%
Washington						
46	1	(5) 10.9%	(26) 56.5%	67.4%	(12) 26.1%	(3) 6.5%
54	2	(11) 20.4%	(31) 57.4%	77.8%	(9) 16.7%	(3) 5.6%
62	3	(11) 17.7%	(36) 58.1%	75.8%	(11) 17.7%	(4) 6.5%
37	4	(6) 16.2%	(22) 59.5%	75.7%	(7) 18.9%	(2) 5.4%
42	5	(7) 16.7%	(26) 61.9%	77.6%	(7) 16.7%	(2) 4.8%
241	Total	(40) 16.6%	(141) 58.5%	75.1%	(46) 19.1%	(14) 5.8%

Note: Results are percentages; figures in parentheses are the relevant number of audits.

Table 5.2 OUTCOMES IN THE URBAN INSTITUTE ANGLO/HISPANIC STUDY IN CHICAGO AND SAN DIEGO (Get Job or Not)

Number of Audits	Pair	(a) Both get job	(b) Neither gets job	a + b	Anglo yes, Hispanic no	Anglo no, Hispanic yes
Chicago						
33	1	(9) 27.3%	(16) 48.5%	75.8%	(7) 21.2%	(1) 3%
32	2	(4) 12.5%	(18) 56.3%	68.8%	(9) 28.1%	(1) 3.1%
39	3	(9) 23.1%	(20) 51.3%	74.2%	(7) 17.9%	(3) 7.7%
38	4	(4) 10.5%	(19) 50.0%	60.5%	(10) 26.3%	(5) 13.2%
142	Total	(26) 18.3%	(73) 51.4%	69.7%	(33) 23.2%	(10) 7.0%
San Diego						
39	1	(5) 12.8%	(24) 61.5%	74.3%	(6) 15.4%	(4) 10.3%
37	2	(8) 21.6%	(18) 48.7%	70.3%	(9) 24.3%	(2) 5.4%
44	3	(14) 31.8%	(17) 38.6%	69.4%	(11) 25%	(2) 4.6%
40	4	(9) 22.5%	(18) 45.0%	67.5%	(8) 20%	(5) 12.5%
160	Total	(36) 22.5%	(77) 48.1%	70.6%	(34) 21.2%	(13) 8.1%

Note: Results are percentages; figures in parentheses are the relevant number of audits.

Table 5.3 OUTCOMES IN THE DENVER STUDY (Aggregate Data)

Total	Pair	Majority favored	Minority favored	Neither favored
140	Hispanic/Anglo	(26) 18.6%	(36) 25.7%	(78) 55.7%
145	Black/White	(17) 11.7%	(15) 10.3%	(113) 77.9%

Notes: Results are percentages; figures in parentheses are the relevant number of audits. The study is reported in James and DelCastillo (1991).

one of the most striking features of all three tables is the relatively high proportion of trials in which there was no difference in treatment by race/ethnicity—roughly 80 percent by the get-a-job measure. The Denver Hispanic/Anglo study showed the smallest proportion of equal treatment, but in that study Hispanics were actually favored over Anglos. Compared with the housing audit studies of Yinger (1986) or the car negotiations tested by Ayres and Siegelman (1993), the proportion of tests in which applicants received equal treatment is very high. By focusing on the disparities between the treatment of majority and minority group members, the Urban Institute studies deemphasize the high proportion of audits in which equal treatment of both partners was found. An appropriate question, therefore, is "whether the glass is one-quarter empty or three-quarters full?" In all of the audit pair studies it seems quite full to us.

Symmetrical Treatment vs. Zero Differences

Any measure of discrimination can be justified only in terms of an implicit or explicit model of how discrimination arises. In a later section we present a model that seems rich enough to enable us to evaluate various measures. The UIBW study (though not the UIAH study) defines discrimination by the proportion of trials in which members of a pair of matched testers were treated differently.[9] While this definition has some appeal, it has some peculiar implications and is based on a somewhat implausible model of an error-free hiring process.

For example, consider the following hypothetical audit pair outcome:

	Tester's Race	
Firm	White	Black
1	Offer	No offer
2	No offer	Offer

Under the UIBW definition, if the white tester gets a job offer from firm 1, while the black tester doesn't, we conclude that there is dis-

crimination in favor of the white tester at firm 1. If the black tester gets a job offer from firm 2, and the white tester does not, we conclude that there is discrimination at firm 2 as well (although, of course, in the opposite direction). The UIBW definition therefore implies that we are observing discrimination at both firms. Suppose that, consistent with these results, firm 1 openly advertises that it will hire only white applicants, while firm 2 makes it clear that it will hire only black applicants.[10] These behaviors would both be a clear violation of race-neutral hiring, so the definition seems appropriate in this setting—we would want to sum the differences in treatment across firms to determine how much discrimination there was. Applying this definition to the numbers in the second panel of table 5.1 suggests that 14.1 percent of the audit observations in Chicago were consistent with discriminatory behavior on the part of firms.

But another possibility is also worth considering. Suppose that testers A and B are identical in all respects (except for race), that both firms saw the two candidates as very close, but that "random factors" led firm 1 to prefer the white tester and firm 2 to prefer the black. As long as there is some uncertainty about future productivity, or there are characteristics valued by a firm that cannot be perfectly aligned in an audit pair and there is scope for intrafirm variability in the ranking of the two candidates unrelated to their race or ethnicity, there is no reason to imagine that every instance of differential treatment constitutes discrimination. It could rather be the case that both testers face the same chance of getting a job at either of these two firms (and at every other firm in this hypothetical economy).

This possibility suggests a different definition of what constitutes discrimination. Discrimination exists whenever two testers in a matched pair are treated differently in the aggregate or on average. Thus, rather than suggesting discrimination at both firms, the hypothetical outcomes presented above could also support an inference of no discrimination in the aggregate.[11] Rather than summing the differences in treatment, we would subtract them to arrive at a net amount that reflects average treatment across all firms.[12]

The net, or symmetric treatment, definition is consistent with the view that whites experience no discrimination and that the proportion of trials in which blacks are hired and whites are not constitutes a benchmark measure for "randomness" in employment decisions. (It is also consistent with the view that one cannot measure an absolute level of discrimination in the labor market but that asymmetry of treatment of "identical" persons constitutes evidence of discrimination.) By this measure, there is evidence of discrimination in only 7.1

percent of the Chicago audits summarized in the first panel of table 5.1. This second definition asks whether in the aggregate the two testers in a pair are "exchangeable"—whether there is *symmetric* treatment of the two testers—while the first definition sees each instance of disparate treatment as evidence of discrimination. The second definition is only an aggregate measure and is consistent with an economy in which half the firms employ only blacks and half employ only whites and both black and white workers sample from the same universe.

The UIBW study essentially adds a measure of asymmetric treatment favoring whites (the proportion of times a white is hired and a black is not) and asymmetric treatment favoring blacks (the proportion of times a black is hired and a white is not) to arrive at a total amount of discrimination. But this makes sense only if one believes that asymmetric treatment of both blacks and whites is race based. If asymmetric treatment against blacks is race based but asymmetric treatment against whites simply reflects employer "errors" in the hiring process, it is meaningless to add the two measures to form an index of overall discrimination.[13]

The choice between the two definitions of discrimination is not entirely straightforward. We would not want to preclude the possibility of detecting reverse discrimination, should it be occurring. The UIBW study properly raises this possibility, although only implicitly.

However, the exchangeability or symmetry definition of no discrimination, which allows for random errors in hiring, is the natural point of departure for any statistical testing procedure. The first definition is a special case of the second in which blacks and whites are treated symmetrically both on average (as in the second definition) and in each instance. (The first definition of no discrimination requires symmetry in the sense that the probability that a black is hired and a white is not should equal the probability that a white is hired and a black is not, and *both should equal zero.*) In terms of testing hypotheses, it is desirable to start with the weaker version. If blacks and whites are treated asymmetrically by the second definition, they cannot be treated symmetrically by the first definition.

The choice of a measure of discrimination vitally affects the interpretation of the evidence in tables 5.1 through 5.3 and the choice of an appropriate test statistic. To see this it is useful to be more precise. Let P_i be the population probability of outcome i. There are four mutually exclusive and exhaustive outcomes in the second panel of table 5.1: (a) Both auditors get a job ($i = 1$); (b) neither gets a job ($i = 2$); (c) white gets a job but black does not 5($i = 3$); (d) black gets a job

but white does not ($i = 4$). The underlying probability model is thus multinomial.

By definition,

$$\sum_{i=1}^{4} P_i = 1.$$

Under the UIBW definition of discrimination, the null hypothesis being tested is H_0: $P_3 + P_4 = 0$. That is, $P_3 = P_4 = 0$, since both P_3 and P_4 are nonnegative. Using large-sample test statistics, UIBW presents evidence that rejects this hypothesis. The study therefore concludes that there is significant discrimination in hiring.

Under the alternative definition, the appropriate hypothesis is one of exchangeability or symmetry of outcomes: H_0': $P_3 = P_4$. This definition is intuitively more plausible because it allows for the possibility of race neutral chance or randomness in the hiring and screening process in a way that H_0 does not. Obviously H_0 implies H_0'. Rejection of H_0' implies rejection of H_0.

Hypothesis H_0 raises a technical problem that is finessed in the UIBW study by its use of normal approximations to multinomial cell means. The problem is that of testing for a zero probability. UIBW uses standard *t*-tests to compare sample proportions with zero. The study ignores the fact that the exact variance of the test statistic is zero under the null hypothesis, which implies that standard classical test statistics are invalid.[14] Thus, on purely technical grounds, H_0 is not an attractive null hypothesis. There is no nonrandomized test of the hypothesis with size α (> 0), although if the estimated cell probabilities are nonzero one can safely reject the null with size $= 0$ (i.e., the probability of a type I error equals zero). This problem is alleviated by testing the weaker null hypothesis H_0', the net definition.

Even in situations in which the maintained hypothesis bounds the P_i away from one or zero, some caution is required in using large-sample theory to test hypotheses given the small samples that are an intrinsic feature of audit pair studies. In both the design of the study (i.e., sample size and power calculations) and the interpretation of the evidence, it would be more appropriate to use exact small-sample theory for multinomial models. In Heckman and Siegelman (1993), we consider matters of optimal sample design and the issue of the suitability of large sample theory for small-sample audit studies. For the main tests of interest, large-sample theory is often very inappropriate. It is unfortunate that it has become enshrined in sampling

textbooks, and was used by UI to design its samples and interpret its evidence.

Tests for Homogeneity across Pairs (Pooling)

Before presenting tests of hypothesis H_0', it is necessary to address an important problem not considered in the UI studies: whether it is possible to pool the data across audit pairs. The UIBW and Denver studies assumed that simple pooling (i.e., adding up estimates across audit pairs using unweighted observations) produces valid statistical inferences. Both the UI and the Denver studies pooled observations across individual audit pairs within sites. Such pooling is valid if there is homogeneity in the selection of firms across pairs and if audit pairs are comparably matched (which means only that any discrepancies between minority and majority members are uniform across pairs, not that the two members of each pair are identically matched). Homogeneity across audit pairs is a testable hypothesis, which we now examine. Below we demonstrate that homogeneity in the skill level across pairs in any given site is not necessarily a desirable feature of an audit study. Homogeneous data characterizes only a single skill level and may not represent the dispersion across types in a given demographic group that typifies actual labor markets.[15] A study that forces homogeneity across pairs may not present a valid description of the market for all workers in the sites being studied. Evidence of homogeneity is, however, consistent with the claimed uniformity across audit pairs in the UI and Denver reports, and a test of homogeneity is a test of the efficacy of the auditor matching and selection procedures. Tests for inhomogeneity may indicate whether a single (possibly misaligned) audit pair drives the results of tests for discrimination at any site, although inhomogeneity may also arise if auditors are perfectly aligned within pairs but different levels are selected across pairs.

To examine this issue, we use a standard test for equality of cell proportions across rows for each of the panels in tables 5.1 and 5.2. More formally, we test for homogeneity of rows in contingency tables assuming product multinomial sampling for pairs—the scheme actually used in the UI studies. (See Bishop et al. 1975, 63, for a definition of this sampling scheme.) A detailed description of the test is given in appendix 5.A. Table 5.4 presents the results of these tests for the get-a-job tables. The evidence suggests homogeneity across audit pair teams except for Chicago blacks. (These results are confirmed

Table 5.4 COMPARISON OF χ^2, "AUGMENTED CELL" χ^2, AND CRESSIE-READ TESTS FOR POOLING OF OUTCOMES ACROSS PAIRS

	Blacks vs. Whites
Washington, D.C.	
$\chi^2_{(12)}$	3.27
Augmented[a]	3.02
Cressie-Read[b]	3.26
Chicago	
$\chi^2_{(12)}$	25.09*
Augmented	21.46*
Cressie-Read	24.49*
Both Cities	
$\chi^2_{(27)}$	38.68
Augmented	34.68
Cressie-Read	39.03
	Hispanics vs. Anglos
Chicago	
$\chi^2_{(9)}$	8.50
Augmented	7.49
Cressie-Read	8.48
San Diego	
$\chi^2_{(9)}$	9.01
Augmented	8.41
Cressie-Read	9.02
Both Cities	
$\chi^2_{(21)}$	18.29
Augmented	16.65
Cressie-Read	18.37

*Significantly different from zero at the 5 percent level for a χ^2 test with the appropriate degrees of freedom.

[a]Calculated by adding 0.5 to the value in each cell and computing a χ^2 statistic.

[b]Cressie-Read evaluated at $\lambda = 0.666$.

when Fisher's exact test for pooling is performed.) In the analysis of the Denver data, below, we test for, and reject, homogeneity across audit pairs.

For Chicago blacks, audit pairs 2 and 5 are the sources of the inhomogeneity. Interestingly, these appear to be the pairs that evidence little asymmetry of treatment between minority and majority auditors. Table 5.5 presents the cell-by-cell difference between the proportion predicted under homogeneity and the actual cell. These residuals are asymptotically normal and confirm the claim that the main source of inhomogeneity in the Chicago data is in audit pairs 2 and 5.

Appendix 5.B presents evidence that large sample theory is producing correct inferences in the homogeneity tests. The "Cressie-Read" statistics are a family of asymptotically equivalent test statistics that can produce different inferences in small samples. When they do not, we have greater confidence in applying large sample methods. Evidence from Fisher's exact test reported in Heckman and Siegelman (1993) confirms the validity of the asymptotic test statistics for testing homogeneity.

Tests for Symmetry of Treatment

Visual inspection of tables 5.1 through 5.3 reveals that one can decisively reject the strong hypothesis H_0—the "gross" definition of no discrimination in treatment ($P_3 = P_4 = 0$). This test has a zero type I error, but the hypothesis is implausibly strong. The evidence against the weaker symmetry in aggregate treatment hypothesis H_0' ($P_3 = P_4$) is less clear-cut.

Table 5.6 presents large-sample tests of symmetry for each audit pair and for sites as a whole. On a pair-by-pair basis, one can reject symmetry for only one Washington black audit pair and only one Chicago black audit pair using 5 percent significance levels. For the Washington pairs taken together, however, one rejects symmetry for the entire set of audit pairs. For Chicago pairs as a whole, one also rejects symmetry (although pooling the Chicago data is not legitimate, as indicated above). The source of the rejection is audit pair 4. For the Hispanic/Anglo comparisons in table 5.6, two of the four pairs in each site allow one to reject symmetry, so the hypothesis is rejected overall in both sites.

Elsewhere (Heckman and Siegelman, 1993) we investigate exact versions of the large-sample tests for symmetry on these data tests and demonstrate the fragility of the asymptotic statistics for all audit pairs, especially for Chicago blacks. This is already apparent from the Cres-

Table 5.5 RESIDUAL ANALYSIS OF AUDIT PAIR HOMOGENEITY

Pair	Both Got	Neither	W+, B−	W−, B+	Total	Residual*			
Blacks, Washington									
1	5	26	12	3	46	−1.16	−0.30	1.34	0.23
2	11	31	9	3	54	0.85	−0.19	−0.51	−0.09
3	11	36	11	4	62	0.28	−0.08	−0.31	0.25
4	6	22	7	2	37	−0.07	0.13	−0.03	−0.11
5	7	26	7	2	42	0.01	0.49	−0.44	−0.32
Total	40	141	46	14	241				
Blacks, Chicago									
1	5	23	5	2	35	0.65	−1.33	1.03	0.36
2	5	25	4	6	40	0.30	−1.97	0.09	3.54
3	3	37	3	1	44	−1.04	1.64	−0.72	−0.83
4	6	24	6	0	36	1.16	−1.21	1.58	−1.45
5	3	38	1	0	42	−0.93	2.66	−1.80	−1.60
Total	22	147	19	9	197				

Pair	Both Got	Neither	A+, H−	A−, H+	Total	Residual			
Hispanics, San Diego									
1	5	24	6	4	39	−1.66	1.93	−1.03	0.56
2	8	18	9	2	37	−0.15	0.07	0.52	−0.69
3	14	17	11	2	44	1.74	−1.48	0.71	−1.02
4	9	18	8	5	40	0.00	−0.46	−0.22	1.17
Total	36	77	34	13	160				
Hispanics, Chicago									
1	9	16	7	1	33	1.52	−0.38	−0.31	−1.03
2	4	18	9	1	32	−0.97	0.62	0.74	−0.98
3	9	20	7	3	39	0.90	−0.02	−0.92	0.19
4	4	19	10	5	38	−1.45	−0.20	0.52	1.72
Total	26	73	33	10	142				

*Residual calculated as $R_{ij} = \dfrac{N_{ij} - N_i N_j / N}{[(N_i N_j / N)(1 - N_i / N)(1 - N_j / N)]^{1/2}}$ where N_i is number in row i, N_{ij} is number in cell ij, etc.

sie-Read diagnostic statistics reported in appendix 5.B. Asymptotically equivalent test statistics produce very different inferences for the symmetry hypothesis. Using an exact small sample version of a conventional likelihood ratio test and standard 5 percent significance levels we cannot reject symmetry for all UI sites. From *this* test, we would conclude that the evidence of discrimination is less clear and convincing than the title of this volume seems to suggest. We also fail to reject symmetry for the Denver audit studies using the same test.

But another, and better, way to look at the data is to notice their sign pattern. In UIAH, the pattern of outcomes is always in the same direction. That is, the minority-favored proportion ($\hat{P}_4$) is always smaller than the majority favored proportion ($\hat{P}_3$). Conditioning on observations that are in a 3 cell or a 4 cell, under the null hypothesis of symmetry, the probability of a "3" or a "4" is one-half. Sizable departures of

$$\hat{\theta} = \frac{\hat{P}_3}{\hat{P}_3 + \hat{P}_4}$$

from 1/2 in one direction indicate asymmetry in that direction. The UIAH patterns in both sites could happen only with probability $(\frac{1}{2})^8 = 1/256 = 0.3\%$. For each site (Chicago/San Diego separately), the probability is $(\frac{1}{2})^4 = 1/16 = 6.25\%$. These are relatively rare events under the null of symmetry. In UIBW, there is one reversal in ten audits. This happens with probability $\binom{10}{9}(\frac{1}{2})^{10} = 0.97\%$. There are no reversals in Washington ($p = 3.1\%$) and one reversal in Chicago ($p = 15.62\%$). With the exception of the test for Chicago blacks, the conditional sign test appears to support the UI conclusions much more than do conventional tests, an intuition that we make rigorous in our companion paper.

Tests based on $\hat{\theta}$ are conditional tests. They are so named because they are conducted on samples defined by certain conditioning events. In our case, the conditioning event is that only one member of an audit pair receives a job offer. There are two main advantages of such tests: (a) their size (probability of a type I error) does not depend on specific values of P_1, P_2, P_3 and P_4 which are generally unknown, and (b) they are uniformly most powerful compared to other test statistics for either one-sided or two-sided versions of the hypothesis of no discrimination,

$$H_0\colon \theta = \frac{P_3}{P_3 + P_4} = 1/2.$$

Table 5.6 LIKELIHOOD RATIO TESTS FOR SYMMETRICAL TREATMENT OF TESTERS WITHIN PAIRS, AND ACROSS PAIRS AND CITIES

Pair	Both got job	Neither got job	W yes B no	W no B yes	Total	$\chi^2(1)$
Part I: Blacks vs. Whites						
Washington						
1	5	26	12	3	46	5.78*
2	11	31	9	3	54	3.14
3	11	36	11	4	62	3.40
4	6	22	7	2	37	2.94
5	7	26	7	2	42	2.94
Aggregated	40	141	46	14	241	17.98*
Total $\chi^2(5)$						18.20*
Chicago						
1	5	23	5	2	35	1.33
2	5	25	4	6	40	0.40
3	3	37	3	1	44	1.05
4	6	24	6	0	36	8.32*
5	3	38	1	0	42	1.39
Aggregated	22	147	19	9	197	3.65
Total $\chi^2(5)$						12.48*
Grand Total	62	288	65	23	438	20.89*
$\chi^2(10)$						30.68*

Part II: Hispanics vs. Anglos

San Diego						
1	5	24	6	4	39	0.40
2	8	18	9	2	37	4.82*
3	14	17	11	2	44	6.86*
4	9	18	8	5	40	0.70
Aggregated	36	77	34	13	160	9.72*
Total $\chi^2(4)$						12.78*
Chicago						
1	9	16	7	1	33	5.06*
2	4	18	9	1	32	7.36*
3	9	20	7	3	39	1.65
4	4	19	10	5	38	1.70
Aggregated	26	73	33	10	142	12.97*
Total $\chi^2(4)$						15.77*
Grand Total	62	150	67	23	302	22.46*
$\chi^2(8)$						28.55*

Notes: Likelihood ratio tests calculated as $2[ln(L_u) - ln(L_c)]$, where

L_u = unconstrained likelihood = $(N!/(N_1!\cdot N_2!\cdot N_3!\cdot N_4!))\cdot(N_1/N)^{N1}(N_2/N)^{N2}(N_3/N)^{N3}(N_4/N)^{N4}$

L_c = constrained likelihood = $(N!/(N_1!\cdot N_2!\cdot N_3!\cdot N_4!))\cdot(N_1/N)^{N1}(N_2/N)^{N2}([N_3+N_4]/2N)^{N3+N4}$

* = Significant at the 5% level (critical value for $\chi^2(1) = 3.84$)

Aggregated is computed summing the first four audit pairs and treating them as one pair.

Total and Grand Total refer to the sums of the likelihood ratio statistics for the audit pairs in each table, and overall, respectively.

(See, e.g., Lehmann 1986, or Pratt and Gibbons 1981.)[16]

Neither property is shared by the classical and widely used likelihood ratio test statistic or the small sample *t*-test approximation to it. Because the small sample distribution of the likelihood ratio test statistic depends on particular values of P_3 and P_4, while the distribution of our conditional test statistic does not, the large sample properties of the likelihood ratio test statistic are a poor guide for small sample inferences. Within the classical (large sample) statistical model, this dependence on P_3 and P_4 is handled by picking as a testing point the *worst case* null (among all possible combinations of P_3 and P_4 that set $\theta = \frac{1}{2}$) so that the probability of a type I error is *no greater than* α (where α is usually .05 or .10). Such procedures are inherently conservative—there are many values of P_3 and P_4 for which the true type I error rate is much less than α. (In our companion paper we show that there are values of P_3 and P_4 near zero that make the type I error rate arbitrarily small.)

A more efficient use of the inherently small sample information available from audit pair studies is based on conditional inference. When it is used, the evidence, on a pair by pair basis, *rejects* H_0: $\theta = \frac{1}{2}$ more frequently (*i.e.*, finds more evidence of disparity in treatment) than when widely used likelihood ratio tests are correctly employed. In our companion paper we develop in detail the argument that conditional inference is the more appropriate framework for analyzing small sample audit pair data. Details of the construction of the conditional test are given in appendix 5.C.

All of this discussion is conducted for a single pair. What is the appropriate procedure for combining evidence across pairs within a given study? If the pairs are homogeneous (identical under the null and identical under the alternative $\theta \neq \frac{1}{2}$), the appropriate way to pool the data across pairs is to combine the results from all pairs to make a single "synthetic" pair and then use the conditional procedures just described to test H_0: $\theta = \frac{1}{2}$. When the alternatives are not the same for all pairs, the matter is more delicate. A full analysis is presented in our companion paper.

One way to combine information in this more general setting is to use the tools of meta-analysis (see, e.g., Hedges and Olkin, 1985). An easy way to deal with heterogeneous alternatives is to use Fisher's pooling method for *p* values, adjusting for continuity using the method of Pearson (1950).[17]

Table 5.7 presents evidence on these issues. For each pair, for the synthetic pair created by aggregating over all pairs, and for two pooling procedures, we examine the conflict in inference that arises from

using alternative testing procedures. The first column records the study. The second records the pair number. The third reports the conventional likelihood ratio test for each pair, for the synthetic pair formed by adding the results for each of the pairs ("Aggregate"), and for the sum of the likelihood ratios for each pair ("Σ2lik ratio").

The fourth column reports the inference from a 5 percent exact one-sided test: H_0: $\theta = \frac{1}{2}$ vs. H_0: $\theta > \frac{1}{2}$. "R" means reject. "$R(b)$" means reject $b \times 100\%$ of the time using a randomized testing procedure. (Randomization is required to produce exact α percent values.) The higher is "b", the more confident we would be in rejecting the hypothesis (*i.e.*, the more likely it is that the outcome in hand would produce a rejection on a randomized test). The fifth column reports the inference from an exact two-sided test (H_0: $\theta = \frac{1}{2}$ vs. H_A: $\theta \neq \frac{1}{2}$). The last column reports the Fisher pooled p value test using one-sided tests (and adjusting for discontinuity in the p values). This test recognizes that the pairs may be heterogeneous, as is certainly the case for Chicago blacks, and Denver blacks and Hispanics. We do not report the conservative small sample adjustment to the likelihood ratio statistics. These adjustments mark up the size of the "true" (least favorable) type I error associated with the conventional likelihood ratio test statistic. The individual Denver data are presented in table 5.10 (p. 221).

The general pattern is that the exact tests tend to reject the null H_0: $P_3 = P_4$ at the 5 percent level more often than the conventional likelihood ratio tests on a pair-by-pair basis. This is even true for the synthetic pairs. The summed likelihood ratio tests are larger than the synthetic pair likelihood ratio tests, but this difference is misleading because the latter are $\chi^2(1)$ whereas the former are $\chi^2(4)$ or $\chi^2(5)$. Pooled p values aggregated in the manner of Fisher, produce much less sharp rejections of the null. The "optimal" pooling procedure across pairs is discussed in our companion paper. Pooling data by forming synthetic pairs overstates the strength with which we can reject the null hypothesis of symmetry.

Appendix 5.C presents evidence from another statistical model that allows for heterogeneity among the audit pairs in each UI site. This heterogeneity can arise as a consequence of sampling variability in the set of firms selected for a particular audit pair member. If we assume that the P_i from each site are observations from a Dirichelet distribution, it is possible to estimate the *distribution* of the P_i and test the hypothesis of symmetry. The test in this case is of the hypothesis that the marginal *distributions* of P_3 and P_4 are equal. The hypothesis is rejected in all studies except for Chicago blacks.

Table 5.7 INFERENCE ON H_0: $P_3 = P_4$ FROM LARGE SAMPLE, EXACT TESTS, AND POOLED P VALUES

Study	Pair	Likelihood Ratio Statistic	Exact One Sided[c] 5% Test	Exact Two Sided 5% Test	Fisher Pooled Test with Pearson Continuity Correction (One Sided Test)[g]
Black vs. Whites	1	5.78	R[d]	R[d]	
Washington	2	3.14	R(.57)[e]	R(.11)[e]	
	3	3.40	R(.77)	R(.18)	
	4	2.94	R(.77)	R(.77)	
	5	2.94	R(.77)	R(.77)	
	Aggregate[a]	17.98 ($\chi^2(1)$)	R	R	28.61
	Σ2lik ratio[b]	18.20 ($\chi^2(4)$)			($\chi^2(10)$)
Blacks vs. Whites	1	1.33	NR	NR	
Chicago	2	.40	NR	NR	
	3	1.05	NR	NR	
	4	8.32	R	R	
	5	1.39	R(.10)	R(.05)	
	Aggregate[a]	3.65 ($\chi^2(1)$)	R	R(.28)	15.40
	Σ2lik ratio[b]	12.48 ($\chi^2(5)$)			($\chi^2(10)$)
Hispanics vs. Anglos	1	.40	NR	NR	
San Diego	2	4.82	R	R(.71)	
	3	6.86	R	R	
	4	.70	NR	NR	
	Aggregate[a]	9.72 ($\chi^2(1)$)	R	R	20.24
	Σ2lik ratio[b]	12.78 ($\chi^2(4)$)			($\chi^2(8)$)
Blacks vs. Whites	1	5.06	R	R(.68)	
Washington	2	7.36	R	R	
	3	1.65	NR	NR	
	4	1.70	NR	NR	

	Aggregate[a]	12.97 ($\chi^2(1)$)	R	R	23.07
	Σ2lik ratio[b]	15.77 ($\chi^2(4)$)			($\chi^2(8)$)
Black vs. Whites	1	6.93	NR	R(.8)	
Denver	2	13.86	R	R	
	3	1.39	R(.1)	R(.05)	
	4	0.34	NR	NR	
	9	0	—[f]	—	
	Aggregate	1.33 ($\chi^2(11)$)	NR	NR	15.25
	Σ2lik ratio[b]	22.52 ($\chi^2(4)$)			($\chi^2(10)$)
Hispanics vs. Anglos	5	1.39	R(.1)	R(.05)	
Denver	6	7.72	NR	R	
	7	0	—[f]	—	
	8	1.02	NR	NR	
	Aggregate[a]	3.06 ($\chi^2(4)$)	R(.64)	NR	2.75
	Σ2lik ratio[b]	13.19 ($\chi^2(4)$)			($\chi^2(10)$)

Notes: [a]Aggregate treats data from all pairs as a single pair. This is the correct procedure if P_1, P_2, P_3, P_4, values are the same across pairs under the null and the alternative.

[b]Sum of $2ln$ likelihood across audit pairs: $\chi^2(4) = 9.49$ at 5%; $\chi^2(5) = 11.07$ at 5%.

[c]The one-sided alternative considered is $P_3 > P_4$.

[d]R means reject at 5% significance level using Uniformly Most Powerful Unbiased Test (UMPU). NR means we do not reject equality.

[e]R(.57) means that 57% of the time one would reject the hypothesis at a 5% significance level using a randomized test.

[f]Cannot be evaluated.

[g]This is the Fisher Aggregation Test for Pooling p-values for one-sided test $\frac{P_3}{P_3 + P_4} = 1/2$ vs. $\frac{P_3}{P_3 + P_4} > 1/2$ using Pearson's (1950) continuity correction as obtained from the median of 1,000 trials to correct P values for discontinuity in the data. It is distributed $\chi^2(2K)$ where K = number of pairs. $\chi^2(10) = 18.31$ at 5% level; $\chi^2(8) = 15.51$ at 5% level; $\chi^2(1) = 3.84$ at 5% level. The test statistic is $-2\sum_{k=1}^{K} ln\ p_k \sim \chi^2(2K)$ where p_k is the one-sided p value (for $\theta > 1/2$) and K is the number of audit pairs. See Hedges and Olkin (1986).

One should be cautious about this particular piece of evidence, however, because the justification for it is based on large sample statistics. This test is discussed further in Heckman and Siegelman (1993). In that paper we also present exact small-sample tests based on the multinomial distribution. Accounting for the composite nature of the null hypothesis $P_3 = P_4$ (see Lehmann 1989) and the composite nature of the alternative ($P_3 \neq P_4$), these tests do not support the inferences based on the standard asymptotic distributions. The classical large sample tests conventionally applied understate the size of tests, sometimes badly so. The bias in using conventional testing methods is toward rejecting equality of treatment, i.e., toward finding discrimination.

Our reanalysis of the data produces the following conclusions. For Chicago blacks, Denver blacks and Denver Hispanics, conventional likelihood ratio test statistics do *not* reject the hypothesis of symmetry of treatment. When adjusted for their small sample dependence on P_3 and P_4, the evidence produced from these statistics is much less strong (see our companion paper). However, use of exact small-sample conditional statistics produces sharper evidence against the null of symmetric treatment.

In this paper, we have not presented a thorough analysis of aggregation across pairs when there is interpair heterogeneity. Tests based on synthetic pairs reject the null of symmetry much more strongly than tests that allow for heterogeneity. Our companion paper pursues this point at length.

In sum, our analysis to date demonstrates (a) the value of making pairs homogeneous and (b) the value of using exact conditional testing procedures to test the hypothesis of symmetry. Large sample tests are misleading in light of their true small sample performance. Tests that rely on large sample methods to design and evaluate an intrinsically small sample problem fail to exploit the full promise of the audit method.

We next turn to the question of what the evidence from the large sample tests would mean about the prevalence of discrimination in the tested labor markets if it were taken at face value.

QUALIFYING THE URBAN INSTITUTE CONCLUSIONS: SOME LIMITATIONS OF THE EVIDENCE

This section addresses four aspects of the audit methods used in the employment discrimination studies. The first is the question of the

proper sampling frame for selecting the firms and jobs to be audited. The second is the possibility of "experimenter effects." The third is the possible bias induced by using false credentials to align audit pair members within a pair and to the skill level of the market being studied. The fourth is the possible problem posed by the presence of facial hair and accents among the Hispanic testers.

Sampling Frame

The Urban Institute studies presented persuasive reasons for sampling jobs from newspaper advertisements. Employers seeking applicants through this route clearly signal the availability of jobs. For audit studies with limited budgets and operating over limited time frames, it is clearly much more cost-efficient to sample firms with jobs than it is to sample the universe of all firms to determine subsamples of firms that are hiring.

An important drawback to the use of newspaper advertisements in constructing the sampling frame, however, is that relatively few actual jobs are obtained through this route, even for youth participating in the unskilled labor markets analyzed in the Urban Institute and Denver studies. Recent evidence by Holzer (1988), presented in tables 5.8 and 5.9, indicates that friends and relatives and direct contact of firms by applicants are much more common sources of jobs in searches by both employed and unemployed youth. For employed youth, only 26 percent of search time is spent on contacts generated by newspaper advertising. For unemployed youth, the corresponding figure is only 18 percent (see table 5.8.). Table 5.9 documents that job acceptances from newspaper-generated searches are much less common than acceptances from other sources. In sum, youths' job searches are characterized by primary use of informal networks.[18] Sampling from these networks poses a major challenge to the audit pair methodology.

Holzer's evidence suggests that the sampling frame adopted in the UI studies is not representative of the job search process followed by most workers. The UI studies claim that this lack of representativeness leads to an understatement of the extent of labor market discrimination as measured by their studies. Their argument is that discriminatory firms are more likely to use informal employment sources, rather than publicly advertising for applicants, in an effort to conceal their discriminatory practices and avoid prosecution under employment discrimination laws.

The claim that firms that hire through networks are inherently more discriminatory than firms that hire from newspaper advertisements

Table 5.8 SEARCH METHODS OF EMPLOYED AND UNEMPLOYED JOB SEEKERS: MEANS AND STANDARD DEVIATIONS

Search Method	Employed		Unemployed	
Number of methods used	2.723		3.285	
	(1.283)		(1.261)	
Percentage using:				
Friends/relatives	.873		.852	
Direct contact	.693		.796	
State agency	.303		.538	
Newspaper	.449		.578	
Other methods[a]	.409		.524	
Time[b] spent by those using:				
Friends/relatives	167.81		295.97	
	(309.71)		(516.37)	
Direct contact	271.71		363.28	
	(238.69)		(536.28)	
State agency	147.07		212.23	
	(148.37)		(298.09)	
Newspapers	251.04	26%	237.74	18%
	(454.99)		(309.11)	
Other methods[a]	121.55		218.65	
	(113.46)		(251.11)	
Total minutes	959.2		1328.00	

Sources: National Longitudinal Survey New Youth Cohort, 1981 panel. Holzer, July 1987, table 1.

Notes: All means are weighted. Sample sizes are 438 for employed seekers and 609 for unemployed seekers.

[a]Other methods include private agencies, school placement offices, labor unions, and community organizations.

[b]Minutes spent on the search method.

must be treated with caution, however. The use of informal networks may simply be an efficient means of screening prospective workers, in that firms may prefer to rely on the information and reputation of existing workers when considering new applicants. In light of the great gains made by minority workers in unskilled markets in the past 25 years, it is not obvious that informal networks now act to exclude minorities. Indeed, in unskilled markets like those studied by the Urban Institute, it seems likely that there are both majority and minority networks that certify applicants by word of mouth.

Nevertheless, Holzer (1987) reports that the fraction of blacks using each search method is virtually identical to that of whites but that job offers from friends and relatives and direct contacts are greater for whites than for blacks. Whites spend more time searching by direct

Table 5.9 OUTCOMES OF SEARCH METHODS USED BY UNEMPLOYED YOUTH

Outcome	Percentage
Percent of job seekers who reported:	
One offer	.220
Two or more offers	.120
Percent of job seekers who reported:	
Friends/relatives	.177
Direct contact	.186
State agency	.089
Newspaper	.099
Other methods	.078
Percent of job seekers who reported:	
One job acceptance	.243
Two or more job acceptances	.043
Percent of job seekers who reported acceptances from use of:	
Friends/relatives	.143
Direct contact	.121
State agency	.048
Newspaper	.040
Other methods	.050

Source: Holzer (1987).
Notes: Samples for those reporting offers and acceptances for each method include only those who used each one. All means are weighted.

application and through leads generated by friends and relatives than do blacks (397 minutes vs. 252 minutes, respectively, for searches via friends and relatives). Blacks spend more time on searches at state agencies (292 minutes vs. 187 minutes), newspaper searches (292 minutes vs. 223 minutes) and other methods (266 minutes vs. 205 minutes). Assuming rational search behavior by both blacks and whites, this pattern provides indirect support for the UI claim that firms advertising in newspapers are less discriminatory.

Future research should make an effort to audit jobs obtained through informal networks, such as those obtained by word of mouth. Such jobs are clearly much more difficult to identify, and the paired-tester approach is unlikely to be a feasible way to sample most of the hiring action in the labor market. Accordingly, evidence from labor market audit studies has uncertain generality.[19]

Experimenter Effects

By experimenter effects, we mean that "the experimenter is not simply a passive runner of subjects, but can actually influence the results"

of an experiment (Lindzey and Aronson 1975, 66). When it exists, such influence is not exerted by any deliberate or conscious actions on the part of the experimenter but, rather, occurs because of unconscious motivations or because subjects may have a desire to conform to (what they perceive as) the experimenter's wishes.

Social psychologists, who have a much longer and more sophisticated tradition of behavioral experiments than do economists, take experimenter effects extremely seriously. In one experiment involving learning in rats, for example, "each experimenter was randomly assigned a rat after being told that the animal had been specially bred for brightness or dullness. Lo and behold, when the results were tabulated, the so-called bright rats learned more quickly than the so-called dull rats" (Lindzey and Aronson 1975, 67).[20]

Anyone who doubts the importance of experimenter effects need only consider the importance of controls in the testing of new drugs. According to one expert, "it is not at all unusual to find placebo effects that are more powerful than the actual chemical effects of drugs whose pharmacological action is fairly well understood" (Rosenthal 1976, 134). Studies evaluating the effects of drugs are always double blind (neither the patient nor the experimenter knows whether the drug being administered is a placebo or the real drug) precisely to minimize such effects.

Both of the Urban Institute studies, as well as the Denver audits, are potentially subject to experimenter effects. All three studies made a point of stressing the nature of the experiment and its expected findings to the testers in several days of training. This was done in part to minimize the psychological impact of discriminatory treatment on minority auditors, but it may have had perverse unintended effects. (However, the performance of the other testers in the project was not revealed to any tester, so at least there was no contamination from direct feedbacks.) An explicit part of the training of auditors was a general discussion of the pervasive problem of discrimination in the United States.

In the UIBW study, part of the first day of the five-day auditor training session included an introduction to employment discrimination and equal employment opportunity and a review of project design and methodology. Similar protocols were used in UIAH. In both studies, participants were warned about the employer bias they might encounter and how they should react to it. We would prefer an experimental design in which the testers themselves were kept ignorant of the hypothesis being tested (discriminatory hiring) and the fact that they were operating in pairs.

Posing and the Use of False Credentials

All of the studies supplied applicants with partially false credentials in order to make audit pair members resemble each other more closely. All of them used college students who masqueraded as blue collar workers seeking entry level jobs. Apart from the ethical issues involved, this raises the potentially important problem that the tester/actors may not have experienced what actually occurs in these labor markets among real participants. For one thing, the auditors may have been overqualified, or have been perceived as overqualified by employers. These suspicions may have been reinforced by the time of year at which the applicants appeared at the door. The audits were conducted in the summer, in order to accommodate the summer schedule of the auditors. At such times, it would be natural for employers seeking long-term workers to be suspicious that the auditors were actually college students who would not stick around in the fall. Credentials may be more closely checked in summer months for this reason, especially for persons who appear to be overqualified. If race or ethnicity is a factor in arousing employer suspicion, differential checking rates by race/ethnicity of applicant, and subsequent discovery of falsified data, might account for the UI findings of no discrimination. Since whites are more likely to attend college than blacks (Cameron and Heckman 1992), white credentials may be more likely to be checked. If the discovery of forged credentials leads to lower white hiring rates, black/white differentials in job offer rates would be understated. On the other hand, discrimination on the part of employers may take the form of greater suspicion of blacks and their credentials than of whites and their credentials. More checking of black credentials, and greater subsequent discovery of false credentials, could lead to an overstatement of true black-white racial disparity in hiring.

If credentials are not, in fact, checked by firms, as UI staff have claimed is the case, this would allay our concerns. It would be helpful to document that the concerns raised in this subsection are irrelevant, if in fact they are.

Facial Hair and Accents

In the Urban Institute study, all the Hispanic testers in San Diego had facial hair and strong Hispanic accents.[21] The presence of accents, facial hair, or any other characteristic across *all* testers of one type is unfortunate because it means that the hypotheses of discrimination

against "accents" or "hair" are observationally equivalent to (indistinguishable from) the hypothesis of discrimination against Hispanics per se. There is some evidence that employers are sensitive to the general appearance or "attractiveness" of applicants.[22] Thus, it is interesting and important to know whether Hispanic men without beards and/or accents do better relative to whites than do those with these characteristics.[23] The fact that in Denver, Hispanic men were more like Anglo men in these characteristics, and did not experience discrimination, indicates that this problem is potentially serious.

Using the presence of facial hair or an accent—rather than race/ethnicity itself—to make hiring decisions could constitute discrimination that would subject an employer to "disparate impact" liability under Title VII of the 1964 Civil Rights Act.[24] Since effect, not intent, is what is at issue in such cases, if Hispanics were disproportionately hurt by a "no beards" or "no accents" rule, one might be able to mount a legal challenge to such a rule. Still, it makes an important difference to policy whether employers are using ethnicity directly in making hiring decisions, or are instead relying on apparently neutral rules that disproportionately hurt minorities. Moreover, in the context of the audit pair methodology, in which Anglo and Hispanic testers were virtually identical in all other observable productivity-related characteristics, slight differences in accents (or facial hair) could have ended up being more important in employers' decisions than they ordinarily are. In other words, employers might have used accents or facial hair only to "break a tie" between candidates who were otherwise identical. This kind of behavior might well constitute discrimination, but it is probably an unusual kind of discrimination compared to what typically occurs in the market, with very different policy implications.[25]

ADDITIONAL COMMENTS ON THE AUDIT STUDIES

Evidence on Reverse Discrimination

Those who see reverse discrimination as a more serious problem than discrimination against racial/ethnic minorities will find no support in any of these findings. In virtually no dimension did white and Anglo auditors consistently do worse than their black and Hispanic partners.

Evidence on the Role of Hiring Discrimination in Explaining the Wage Gap

Using large sample test statistics, a clear pattern of discrimination against blacks is found in the UIBW Washington study. Since government jobs were excluded from the universe from which employers were sampled, it is not clear how seriously these findings are to be taken as a general description of that market.

Taken at face value, the disparity between black and white job offer rates is substantial. Whites were offered jobs in 35.7 percent of the interviews, while blacks were offered jobs in only 22.4 percent of the interviews. These figures imply that blacks would have to sample about 50 percent more jobs than whites to get an offer.

In a simple model of job search with identical fixed-costs components for blacks and whites, as well as common wage offer distributions and discount rates, these results indicate that blacks would have lower reservation wages and lower accepted wages than whites. Wages observed for working blacks would be lower than wages observed for working whites even if employers made identical wage offers to accepted blacks and whites.

However, it does not follow that blacks would necessarily have higher unemployment rates than whites, although it does follow that with identical nonmarket opportunities, blacks would have lower labor force participation rates. For hiring discrimination to produce higher black unemployment rates requires special conditions on the shape of the wage offer distribution (see Flinn and Heckman 1983 or Van den Berg 1991). Lower reservation wages may offset the effect of lower job offer rates. For the Urban Institute to translate its evidence into convincing stories about unemployment, and about the contribution of hiring discrimination to the wage gap, it will be necessary to collect information on offered wages. In order to gain a more complete understanding of labor market disparities between blacks and whites, it is necessary to know the rate of arrival of potential job offers to race groups. The audit studies produce no information on this question because majority and minority partners are necessarily sent to the same firms at the same rate. It would be necessary to supplement the audit data with conventional labor market surveys in order to parse out the roles of individual search and firm behavior in accounting for majority/minority differences in unemployment and labor force attachment.

No evidence is offered in any of the studies under consideration that discrimination in hiring has increased over time or that the post-

1966 decline in relative labor force participation rates for black males compared with white males is a consequence of increased discrimination in hiring.

Comparing the Denver and Urban Institute Studies

An interesting problem posed by the collection of audit studies considered here is the apparent disparity in the results between the two Urban Institute studies on one hand and the Denver studies on the other.[26] While the Urban Institute studies based on large sample testing methods apparently find evidence of discrimination against Hispanics in San Diego and Chicago, and in Washington, D.C. against blacks, the Denver study suggests virtually no discrimination against either of these groups.

Two explanations for these divergent results should be considered. One possibility is that the differences are simply artifacts of methodological differences between the two groups of studies. Although the methodology used in Denver was patterned after that of the Urban Institute studies, there were some differences that may have had an effect on the results. The second possibility is that there is actually less discrimination (against both blacks and Hispanics) in Denver than in Chicago, San Diego, or Washington, D.C.

Of course, these are not mutually exclusive explanations—both could be operating at the same time. In fact, our view is that the differences between Denver and the other cities seem small enough to be explained by either of the two sources, or both together.

The disaggregated data from the Denver experiments are presented in table 5.10. While we do not resolve this issue definitively in this paper (but see Heckman and Siegelman 1993), the data strongly suggest that the pairs of Denver auditors were much more heterogeneous (diverse) than are the pairs in either UIBW or UIAH. Thus, while the aggregate experience for *all* Denver audit pairs reveals little evidence of discriminatory treatment, this aggregation conceals large differences in the way certain pairs were treated. A Fisher exact test rejects the hypothesis of across-pair homogeneity, as does a large sample test.

Table 5.10 reveals that overall, black auditors got a job when their white partner did not in 7 out of 145 audits (4.8 percent); white auditors were favored in 12 audits (8.3 percent). According to the analysis presented in Table 5.7, this relatively small difference does not provide statistically significant evidence of the existence of discrimination at the aggregate level or at the individual level.

Table 5.10 DISAGGREGATED DENVER DATA: "GET A JOB" MEASURE

			Black/White		
Pair	Both get job	Neither gets job	White yes, Black no	White no, Black yes	Total
1	(2) 11.1	(11) 61.1	(0) 0.0	(5) 27.8	(18)
2	(2) 3.8	(41) 77.4	(10) 18.9	(0) 0.0	(53)
3	(7) 21.2	(25) 75.8	(0) 0.0	(1) 3.0	(33)
4	(9) 60.0	(3) 20.0	(2) 13.3	(1) 6.7	(15)
9	(3) 11.5	(23) 88.5	(0) 0.0	(0) 0.0	(26)
Total	(23) 15.8	(103) 71.1	(7) 4.8	(12) 8.3	(145)
			Hispanic/Anglo		
Pair	Both get job	Neither gets job	Anglo yes, Hispanic no	Anglo no, Hispanic yes	Total
5	(0) 0.0	(11) 91.7	(0) 0.0	(1) 8.3	(12)
6	(4) 7.8	(30) 58.8	(3) 5.9	(14) 27.5	(51)
7	(1) 2.8	(35) 97.2	(0) 0.0	(0) 0.0	(36)
8	(2) 4.9	(30) 73.2	(6) 14.6	(3) 7.3	(41)
Total	(7) 5.0	(106) 75.7	(18) 12.8	(9) 6.5	(140)

Note: Results are percentages; figures in parentheses are the relevant number of audits.
Source: Denver study.

The aggregate results conceal a widely disparate set of outcomes among the different pairs of testers, however. For example, consider pairs 1 and 2. In pair 1, the black tester was favored over his white partner in 5 out of the 18 tests, while the white tester was never favored. For pair 2, the results are dramatically opposite: the white tester was favored in 10 of the 53 tests, while the black tester was never favored. Roughly similar patterns can be observed for Hispanic/Anglo pairs 6 and 8. In the notation developed earlier, the quantity $(P_3 - P_4)$ ranges from 28 percent to -19 percent for the black audit pairs, and from 21.6 percent to -7.3 percent for the Hispanic pairs. There are large differences among the different audit pairs that are masked when the experiences of all the pairs are aggregated.

The heterogeneity found in the Denver data raises three important issues. First, it demonstrates the importance of providing data at the disaggregated (pair-by-pair) level. Aggregation of audit pair results by city can obscure some important differences in the way individual pairs were treated, which can in turn influence the interpretation of the aggregate results. We note that the authors of all three studies were willing to furnish us with the disaggregated data. But these data were not part of the main bodies of the reports prepared by any of the audit teams. (The data were included in an appendix to UIAH.) We

strongly urge that pair-by-pair data be included, as a matter of course, in future studies of this kind.

Second, both the Denver and UI studies show how difficult it is to draw inferences about the homogeneity of pairs from verbal descriptions of the selection and matching procedures by themselves. Looking only at the *descriptions* of the rigorous and careful procedures used in the Denver (or UI) studies, one would naturally be inclined to assume that the pairs were all quite homogeneous. In fact, however, we reject homogeneity in three of the six race/city sites (Denver blacks, Denver Hispanics, and Chicago blacks). In spite of the extremely careful efforts made by all of the researchers to ensure that all the pairs of testers resembled each other, the pairs appear to have been treated quite differently in their respective labor markets.

This raises a final problem. Why should the audit pair analysts have found it relatively difficult to control for heterogeneity across pairs? The answer must rest on the difficulty of comparing and matching large numbers of auditor characteristics, many of which are intangible or difficult to describe. If mismatching in characteristics occurs in half of the city/race sites despite the best efforts of the testers to prevent it, it would seem that our knowledge of the hiring process being audited is rife with uncertainty. One should be wary of assuming that much is known about how hiring decisions are actually made. The burden of proof that audit pairs are properly aligned must be assumed by audit pair analysts. More objective demonstrations of the matching methods actually used would increase the value of audit pair evidence.

THE VALUE OF HAVING MORE THAN TWO TESTERS

There is potentially great value in having two or more testers from each race/ethnic group. (See McIntyre et al., 1980, who use this method in a resume audit study.) Access to evidence from such audit pair studies would permit calibration of background noise and would provide valuable corroborating evidence on the capacity of UI to match on "relevant" productivity characteristics. It would also help to settle the choice of an operational definition of discrimination.

Thus for two majority group members it would be possible to estimate P_3 and P_4 (proportions that one is hired and the other is not) and determine whether a $P_3 + P_4 = 0$ definition of no discrimination is reasonable. A similar check would be of value for minority group

members. Evidence that $P_3 \neq P_4$ would indicate problems in making identical matches. If, in fact, "identical" matched pairs show $P_3 \neq P_4$, even after accounting for sampling variation, tests of symmetry ($P_3 = P_4$) should be modified. It would be more appropriate to test whether $P_3 - P_4$ lies in an *interval* determined from the "noise band" estimated from the analyses of audit pair members of the same demographic group. Such a band would make audit evidence both intuitively and formally more convincing.

A BEHAVIORAL MODEL OF FIRM'S JOB OFFERS TO HETEROGENEOUS WORKERS

We have discussed the UI and Denver studies on their own terms: as intuitive models of the job offer process. Intuitive models are informative guides to data but can take one only so far. When the researchers in UIBW use the models developed by McFadden (1973) to analyze their data, they implicitly appeal to a particular decision process for firm job offers. In appendix 5.D, we present a rigorous model of firm hiring decisions; we evaluate the methods used by the Urban Institute; and we consider what specifications of these models justify the definitions of discrimination used in the UI studies. A more complete analysis is presented in Heckman and Siegelman (1993).

Three important conclusions emerge from this analysis. First, our model reveals some of the underlying assumptions needed to support each of the various definitions of (and tests for) discrimination. In particular, we demonstrate that the distributions of unobserved (by the UI) worker characteristics within an audit pair must be identical to justify the "symmetrical treatment" definition of discrimination adopted in the UIAH study; an even stronger assumption is necessary to support the "zero difference" definition of discrimination in UIBW—namely that there are no relevant unobservables omitted by the UI analysts.

Second, we point out that standardizing on *observed* productivity characteristics for members of an audit pair may, paradoxically, accentuate the discrepancy in job offer rates between the two audit pair members as compared with what would be measured from observing two randomly selected job searchers. Under certain circumstances, data from randomly selected pairs would provide closer approximations to job offer patterns in an actual labor market.

Third, our model demonstrates that the choice of an observed productivity level for an audit pair may drive the resulting estimates of job offer discrimination produced by the audits. In other words, standardizing the observed productivity level of two auditors at one level can produce much more (or less) evidence of discrimination than standardizing at another level.

By a suitable choice of an observed productivity level, it is always possible to produce evidence of no discrimination, of favoritism toward minorities, or of discrimination against minorities, provided only that minority and majority distributions of unobservables are not identical. A corollary of our analysis is that any evidence of firm-based discrimination can also be interpreted as evidence of differences in the variance of unobservables between majority and minority group members.

These considerations suggest that much more thought should be given to the choice of an appropriate level of auditor productivity characteristics in designing audit pair studies and that a variety of levels across audit pairs should be used to present a more complete picture of labor market hiring practices. The relatively weak evidence of discrimination found in Chicago might simply be a consequence of an unfortunate choice of observed productivity level. It also suggests that audit studies are crucially dependent on an unstated hypothesis: that the distributions of unobserved (by the testers) productivity characteristics of majority and minority workers are identical. The UIBW study is based on the even stronger assumptions that there are no relevant unobservables and that observables can be perfectly matched within audit pairs.

IMPLICATIONS FOR POLICY AND FUTURE RESEARCH

The key item on the agenda for future research is to design audit studies that provide more effective answers to the question of *why* discrimination occurs. Without such answers, policy recommendations are unlikely to be very helpful, because the kind of policies one would favor depend on the kind of discrimination one observes.

We strongly encourage the use of more than two testers at each firm.[25] This would allow analysts to distinguish between random and race-based explanations for differences in treatment. If both white males consistently got job offers while the black males, for example, did not, one would have a stronger case for the race-based explanation.

If only one of the white males consistently got the job, however, the random explanation would seem more plausible. Use of additional applicants would further facilitate measurement of randomness or "background noise" at firms. Use of extra testers would also enable analysts to determine how well relevant characteristics have been aligned. Using extra testers, it is possible to adjust the gross index of discrimination used by UIBW by subtracting the summed gross index for the two demographically comparable patterns from the gross index for two demographically disparate partners.

More attention should be paid to firm-specific variables. The multinomial logit analyses reported in the appendixes to both Urban Institute reports head in the right direction (although they inappropriately pool across audit pairs in Chicago), but we need to go further. What, if anything, distinguishes firms that discriminate against blacks or Hispanics from those that discriminate against whites or do not discriminate at all? Are some firms operating under a quota constraint while others are free to practice discrimination against blacks? Obtaining answers to questions such as these is vital before strong policy recommendations can be made.

We also urge further consideration of the attributes of the testers. It is essential to present more objective evidence on the comparability of the white and minority testers in a pair. The difference between each tester's characteristics and the employer's stereotype or belief about the characteristics of the average applicant of that race may also matter. Thus, it would be useful to know if it makes a difference whether the testers both resemble the typical white applicant or both resemble the typical black applicant in the relevent productivity attributes. In light of the analysis summarized in the previous section, and developed in appendix 5.D and in our companion paper, it is essential to consider a range of observed skill levels across the audit pairs in order to sample market outcomes. Otherwise, there is great danger in picking an anomalous value of skill levels that may understate or overstate actual differences in firm's hiring rates.

As a relatively new research technique, audit studies may benefit from some nontraditional strategies for documenting their results. For example, ABC Television's recent PRIMETIME LIVE feature "True Colors," (9/26/91) used only a single audit pair, filming the two testers as they went about many of the activities of everyday life. This was an exercise in investigative journalism, rather than social science, and we should not judge academic research by the standards it set (or vice versa). Nevertheless, one of the reasons the results were so powerful is that the audience could clearly see how similar the two testers

actually were. Perhaps employment audits could also make use of some of these techniques and provide both visual and oral evidence of alignment by making videotapes of audit pairs available.

The UI studies provide suggestive evidence on the prevalence of discrimination in job hiring. Much more information must be collected on the job offer process, especially on offered wages, before the UI studies can be said to document discrimination on the basis of race or ethnicity.

Richer sampling plans for firms must be undertaken in order to get a more representative picture of the labor market. Mechanisms for sampling of job search methods other than newspaper advertising must be devised if we are to achieve a truly representative view of the labor market. Concerns about the use of falsified credentials should be met. Audits should be conducted using real rather than simulated low-skilled market participants. To understand the conflict between the findings of the UIAH study and the Denver Hispanic study, it is essential to unbundle accents, facial hair, and Hispanic status.

We also urge that the Denver study be replicated following the Urban Institute protocol as closely as possible. Ultimately, this will be the only way of deciding whether the apparent difference in results stems from lower levels of discrimination in Denver or is attributable to the differences in research design. Such replication would be extremely useful to others planning future work in this field; given our limited experience with audits, we are almost completely ignorant about how sensitive the findings are to small changes in methodology. A replication of the Denver audits using the strict UI methods would give us some important information about the robustness of the audit technique.

Finally, we urge further work on the statistical aspects of the design of audit studies and the testing of hypotheses using audit data. Because audit studies are expensive to conduct, sample sizes in such studies are likely to be relatively small, especially for studies conducted by civil rights organizations. Moreover, most of the outcomes being studied, at least in the employment context, are discrete (either the tester is offered a job or he is not) rather than continuous. The combination of small sample size and discrete outcomes means that researchers should make more use of small sample, multinomial statistical techniques in sample design and hypothesis testing, rather than relying on large sample normal approximations. We have explored some of these questions in appendix 5.D, and do so in greater depth in our companion paper (Heckman and Siegelman 1993). We

establish that conventional large sample methods do not efficiently exploit audit pair data.

None of these comments detracts from our admiration for these pioneering studies. The audit pair methodology, properly augmented, promises to shed much valuable new light on the nature of racial and ethnic disparities in labor market outcomes.

Notes

We would like to thank Jerry Marschke and Allison Sylvester for competent research assistance. Moshe Buchinsky, Hector Cordero-Guzmann, Robert LaLonde, Allan Lind, Richard Robb, John Yinger, and Mahmoud Zaidi made valuable comments on this work. We are especially grateful to Gary Becker for insightful comments.

1. The work we focus on is a study of black-white pairs in Washington and Chicago reported in Turner, Fix, and Struyk, *Opportunities Denied, Opportunities Diminished: Racial Discrimination in Hiring* (Urban Institute, September, 1991); a study of Hispanics and Anglos in San Diego and Chicago by Cross, *Employer Hiring Practices: Differential Treatment of Hispanic and Anglo Job Seekers* (Urban Institute, 1990) and the study of blacks, whites and Hispanics in the Denver labor market by James and Del Castillo (March, 1991).

2. We concentrate on employment audits. For more detailed description of the audit techniques used in this context, see the methodological appendixes to Turner et al. (1991) and Cross et al. (1991). Newman (1978) and McIntyre et al. (1980) used resumes, rather than live auditors, to test for differences in the treatment of applicants. For the use of the audit technique in other contexts, see Ayres (1991) (cars), Ayres and Siegelman (1993) (cars), and Yinger (1986) (housing).

3. Fair-housing testers have typically been matched on relatively few characteristics, while employment testers have been matched on many more, reflecting the greater variety of characteristics likely to be relevant in hiring decisions.

4. The similarity of testers *across* pairs is also relevant if, as is almost always the case, data from the pairs are to be aggregated for statistical analysis. We discuss this issue at greater length below.

5. The empirical and theoretical literature on labor market (mostly wage) discrimination is surveyed by Cain (1986). His pessimistic conclusion is that " . . . the results are so varied that they reveal as much about our ignorance as about our knowledge of labor market discrimination" (p. 743). Our ignorance of hiring discrimination is at least an order of magnitude greater than it is with regard to discrimination in wages.

6. See table 5.10 below for the Denver microdata and Heckman and Siegelman (1993) for their extensive analysis.

7. That is, if tester A is denied an interview, while tester B is interviewed but is nevertheless rejected for the job, should one consider this outcome as evidence of discrimination?

8. For one thing, the raw, pair-by-pair data for this index were not available to us for analysis. Moreover, while looking at "nonoutcome" aspects of the hiring process can produce some potentially useful evidence, the relationship between the categories and discrimination is somewhat tenuous. Is being interviewed for a shorter period of time necessarily evidence of discrimination? Is the number of positive comments (as opposed to their strength) necessarily a measure of favoritism? Under the circumstances, these crude measures are probably the best we can hope for, but they should be interpreted with caution. It seems unwise to build much of the case for the existence of discrimination on such subjective measures.

9. Turner et al. (1991).

10. We should note that reverse discrimination could come from black-owned firms, or firms with largely black customers or work forces, that dislike white employees. Or, as seems more likely, this kind of discrimination could be the result of legal pressures to increase hiring of blacks. See, for example, Leonard (1984) for further analysis.

Interestingly, Title VII and other antidiscrimination legislation could also have a negative impact on black hiring. The reason, as suggested by Donohue and Siegelman (1991), is that most litigation under Title VII now contests discriminatory discharge rather than discriminatory failure to hire. Thus, employers may see little costs for failing to hire a minority applicant but potentially high costs to hiring and possibly firing at some later date.

11. This point is developed more formally in the model of job hiring presented in appendix 5.D.

12. One could also take the ratios of the number of tests in which there was disparate treatment. In a companion to this paper, we show that the choice of which test statistic to use may not be innocuous.

13. For an interesting analysis of errors in the hiring process, see McIntyre et al. (1980), who conclude, using employment audits based on resumes rather than live testers, that identical whites are treated differently approximately 12 percent of the time.

14. Let $X_N = (X_{N1}, \ldots, X_{NT})$ have the multinomial distribution $M(N, \underset{\sim}{P})$ where $\underset{\sim}{P} = (P_1, \ldots, P_T)$. Then

$$E(X_N) = N(\underset{\sim}{P})$$

$$Cov(X_N) = N(D_{\underset{\sim}{P}} - \underset{\sim}{P}'\underset{\sim}{P})$$

where $D_{\underset{\sim}{P}} = \text{Diag } \underset{\sim}{P}$.

$$\hat{P} = \left(\frac{1}{N}\right) X_N.$$

$\sqrt{N}(\hat{\underset{\sim}{P}} - \underset{\sim}{P})$ converges under random sampling to

$$\sqrt{N}\,(\hat{\underset{\sim}{P}} - \underset{\sim}{P}) \doteq N(0, D_{\underset{\sim}{P}} - \underset{\sim}{P}'\underset{\sim}{P}).$$

Thus, asymptotically, under H_0

$$\sqrt{N}\,(\hat{P}_3 - P_3) \doteq N(0, 0)$$

$$\sqrt{N}\,(\hat{P}_4 - P_4) \doteq N(0, 0),$$

degenerate normal random variables.

15. Homogeneity is desirable from a statistical point of view in the sense that it justifies a simple pooling of observations. If pooling is not possible and each audit pair must be treated separately, the power of audit methods to detect discrimination is reduced. See Heckman and Siegelman (1993) for further discussion of this point.

16. The test statistic in one-sided form is to check if $\hat{\theta} \geq c$, where c is determined so that the probability of a type I error is α percent. In two-sided form, the test is to determine c_1, c_2 such that $\hat{\theta} \geq c_1$ or $\theta \leq c_2$ for c_1 and c_2 determined so that the overall error rate is α percent. For more details, see descriptions in the Lehmann (1986) or Pratt and Gibbons (1981) or the example in Appendix 5.C.

17. More specifically, in the results we report below, we perform a Monte Carlo analysis on the Pearson adjustment to the Fisher method and use the median of 1000 draws using the method. For more details, see our companion paper. The Fisher method is based on the observation that the p value for a one-sided test is distributed uniformly. Minus twice the log of the p value for a one-sided test is distributed uniformly. Minus twice the log of the p value is distributed $\chi^2(2)$. Twice the sum of logs of the p values is distributed $\chi^2(2K)$ where K is the number of audit pairs, i.e.,

$$-2 \sum_{k=1}^{K} \ln p_k \sim \chi^2(2K).$$

18. Mahmoud Zaidi has pointed out to us that the use of major newspapers as the sampling frame may be problematic. His research suggests that black and Hispanic youth often use community or local newspapers in their job searches, rather than the major newspapers from which the Urban Institute studies sampled.

19. Studies that use resumes, rather than (or in addition to) live auditors, could be potentially useful in this regard. See Newman (1978) and McIntyre et al. (1980).

20. For a massive collection of additional evidence on experimenter effects, see Rosenthal (1976).

21. The accents were deemed appropriate because the study was designed to uncover whether reform of the immigration laws caused employers to discriminate against "foreign-sounding" Hispanics. The UIAH study is careful to note that its results may not generalize to the larger population of Hispanics in the United States. Others, however, have interpreted the results more broadly than the authors presumably intended they would.

22. Dipboye et al. (1975).

23. Note that the Denver study found virtually no evidence of discrimination against Hispanics; it appears that Denver testers had less strong accents than those in the Urban Institute study, although this is not clear. No mention was made of facial hair in the Denver study. There were other differences between the studies as well, as we discuss later.

24. See *Bradley v. Pizzaco of Nebraska* 926 F.2d 714 (8th Cir. 1991) (a no-beards rule meets plaintiff's prima facie burden for asserting a disparate impact challenge).

25. This point is illustrated by the model of the hiring decisions developed below.

26. One should also mention two other studies that use a modification of the audit technique. McIntyre et al. and Newman used resumes, rather than live job applicants, to test for discrimination. Both found statistically significant discrimination in *favor* of black and women applicants. The studies were done more than a decade ago, however, and did not allow applicants to apply for jobs (and hence to receive job offers). The sampling frames were also different (consisting of employers who advertised in *College Placement Annual* that they were accepting written job applications). It is conceivable that employers discriminate in favor of women and minorities at the early stages of the hiring process, but nevertheless discriminate against them when it comes to the actual hiring decision itself.

APPENDIX 5.A

MATHEMATICAL APPENDIX

Our test for homogeneity in the contingency table of audit pairs is standard. Assume product multinomial sampling. For row i and column j, we construct the estimated number in cell ij (assuming no auditor effects) as

$$N_{ij} = \frac{N_i}{N}\frac{N_j}{N}N \qquad \text{all } i, j$$

where N_i is the row i total (summing across i) and N_j is the column j total (summing across j). This can be tested using standard $\chi^2[(R-1)(C-1)]$ statistics. (R is the number of rows; C is the number of columns.) Also, one can form cell-by-cell asymptotically normal deviates to examine which cells produce the departure from normality if there is one.

The likelihood ratio statistic is a special case of a whole class of "Cressie-Read directed divergence" statistics that nest classical Pearson, modified Pearson, and likelihood ratio statistics as special cases. It produces a class of test statistics that should produce equivalent inferences if the asymptotic theory used to conduct conventional tests is any good. Computation of different versions of the statistic therefore provides evidence on the validity of the asymptotics. Let $\hat{P}_{ij}$ be the sample proportion in cell ij. Let $\hat{P}_{ij}^{(r)}$ be the restricted sample proportion in cell i, j under a particular hypothesis. Then the directed divergence test statistic is, for sample size N,

$$I(\hat{P}_{ij};\, \hat{P}_{ij}^{(r)},\, \lambda) = \frac{\sum_{i,\,j}^{R,\,C} N\,\hat{P}_{ij}}{\lambda(\lambda+1)}\left(\left[\frac{\hat{P}_{ij}}{\hat{P}_{ij}^{(r)}}\right]^{\lambda} - 1\right),$$

$$-\infty < \lambda < \infty \qquad \lambda \neq -1. \quad (1)$$

$2I$ is χ^2 with the number of degrees of freedom specified by the null. In the case of independence, the number of degrees of freedom is $(R-1)(C-1)$. When $\lambda = 0$, the classical likelihood ratio is produced. When $\lambda = 1$, the classical Pearson χ^2 is produced. High values of λ downweight high ratios of $(\hat{P}_{ij}/P_{ij}^{(r)})$ that may be due to sampling variation.

For testing symmetry within audit pairs, the same type of statistic can be used. Let there be K outcomes for each audit pair. Let the hypothesis in question be $P_{K-1} = P_K$ (i.e., equality in audit pair outcomes for outcomes K and $K-1$). In our case, these outcomes are those in which one demographic group is treated differently from another.

The test of symmetry for any audit pair is based on

$$2I(\hat{P}_i; \hat{P}_i^{(r)}, \lambda) = \frac{2}{\lambda(\lambda+1)} \sum_{i=1}^{K} N \hat{P}_i \left(\left[\frac{\hat{P}_i}{\hat{P}_i^{(r)}} \right]^{\lambda} - 1 \right). \tag{2}$$

This statistic is chi-square one. In the case of the symmetrical treatment hypothesis, there is one degree of freedom for each audit pair and

$$\hat{P}_i^{(r)} = N \left[\frac{N_i}{\sum_{i=1}^{K-2} N_i + 2 \left(\frac{N^{\frac{1}{\lambda+1}}_{K-1} + N^{\frac{1}{\lambda+1}}_{K}}{2} \right)^{\lambda+1}} \right]$$

for $i \neq K, K-1$.

$$\hat{P}^r_{K-1} = \hat{P}^r_K = \frac{1}{2} \left(1 - \sum_{i=1}^{K-2} \hat{P}_i^{(r)} \right).$$

APPENDIX 5.B

PLOTS OF CRESSIE-READ STATISTICS FOR TESTS OF HOMOGENEITY AND TESTS OF SYMMETRY

We present plots of the Cressie-Read statistics for the values of λ indicated. The labels identify the sites and demographic groups. The figures with one line are for the tests of homogeneity, and the figures with multiple lines are for the tests of symmetry. Each line corresponds to a particular audit pair. A flat line indicates that the same asymptotic inference is being produced by the range of asymptotically equivalent Cressie-Read statistics. Flatness generates confidence in the asymptotic inference. Curvature (as in the audits of Chicago blacks for a single audit pair) casts doubt on the validity of the asymptotic inference.

For $\lambda = 1$, the Cressie-Read statistic is a classical χ^2 statistic, with degrees of freedom indicated in each figure.

Figure 5.B.1 CRESSIE-READ STATISTICS TESTING THE HYPOTHESIS OF HOMOGENEITY ACROSS PAIRS OF TESTERS FOR BLACK/WHITE PAIRS, WASHINGTON, D.C.

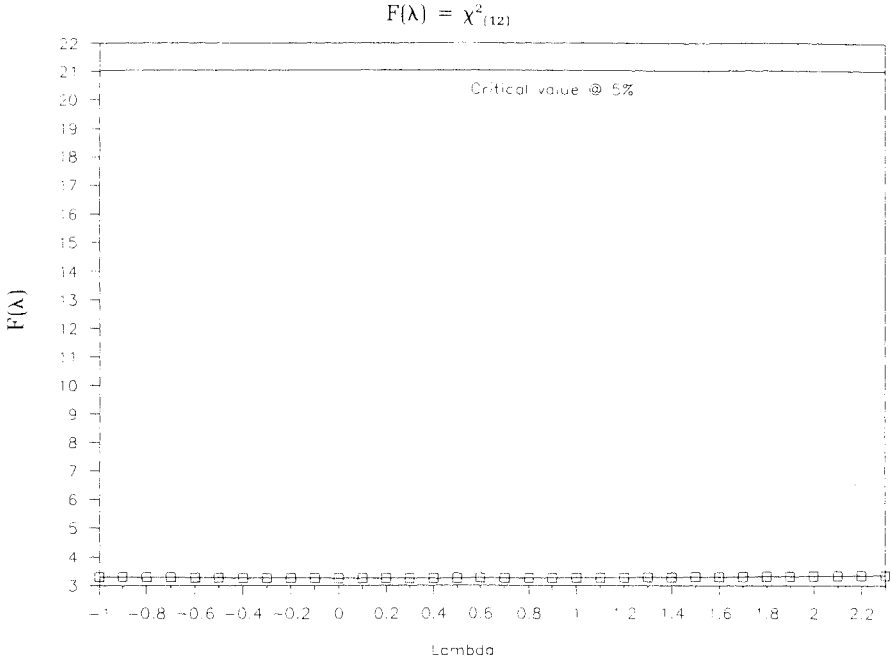


Figure 5.B.2 CRESSIE-READ STATISTICS TESTING THE HYPOTHESIS OF HOMOGENEITY ACROSS PAIRS OF TESTERS FOR BLACK/WHITE PAIRS, CHICAGO

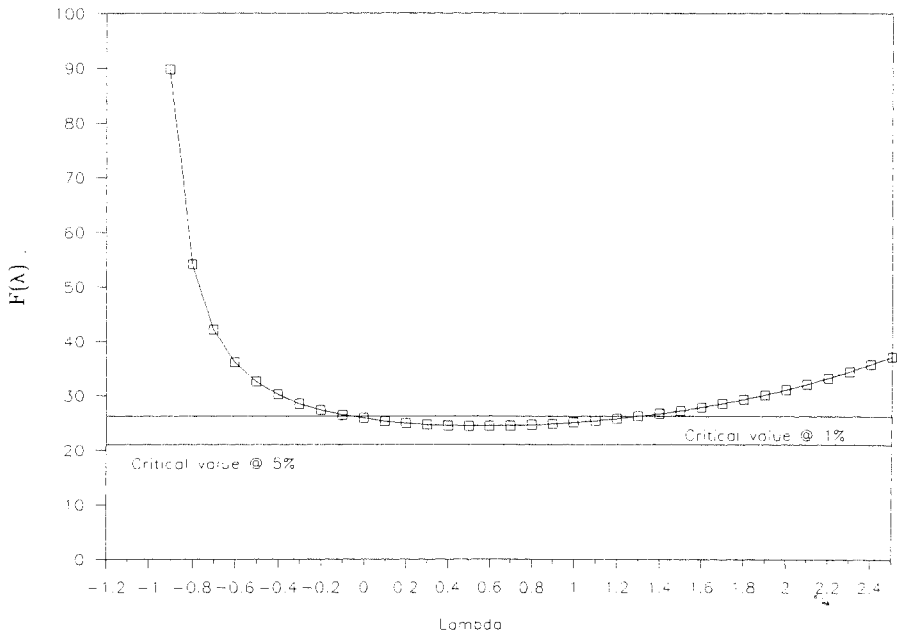


Figure 5.B.3 CRESSIE-READ STATISTICS TESTING THE HYPOTHESIS OF HOMOGENEITY ACROSS PAIRS OF TESTERS FOR ANGLO/HISPANIC PAIRS, SAN DIEGO

$$F(\lambda) = \chi^2_{(9)}$$

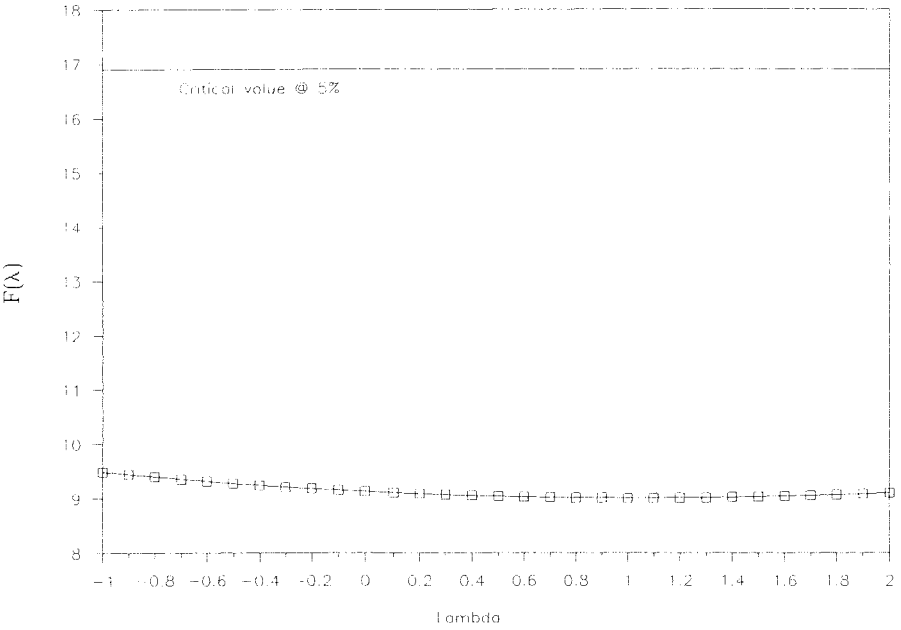


Figure 5.B.4 CRESSIE-READ STATISTICS TESTING THE HYPOTHESIS OF HOMOGENEITY ACROSS PAIRS OF TESTERS FOR ANGLO/HISPANIC PAIRS, CHICAGO

$$F(\lambda) = \chi^2_{(9)}$$

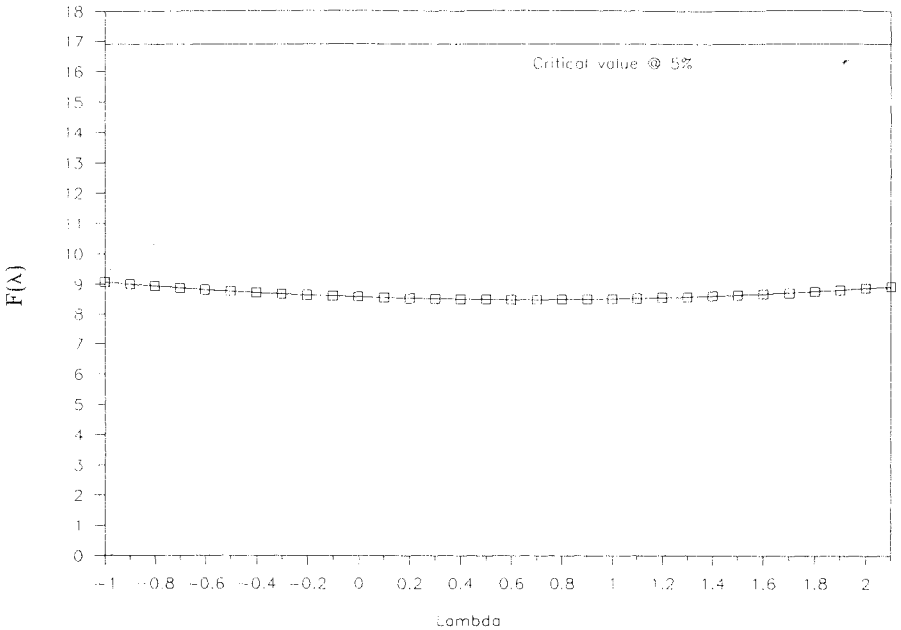


Figure 5.B.5 CRESSIE-READ STATISTICS TESTING THE HYPOTHESIS OF HOMOGENEITY ACROSS CITIES FOR BLACK/WHITE PAIRS

$F(\lambda) = \chi^2_{(27)}$

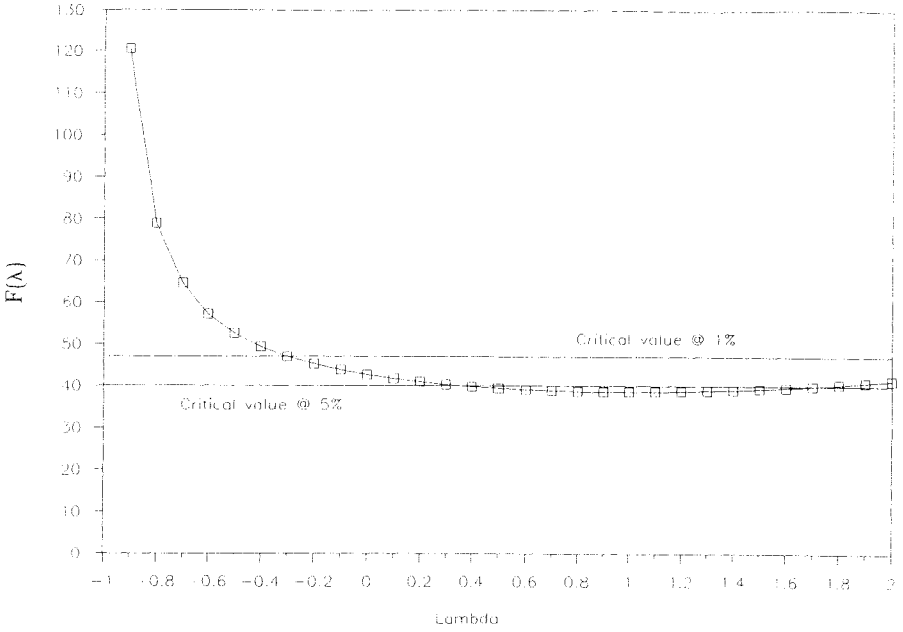


Figure 5.B.6 CRESSIE-READ STATISTICS TESTING THE HYPOTHESIS OF HOMOGENEITY ACROSS CITIES FOR ANGLO/HISPANIC PAIRS

$F(\lambda) = \chi^2_{(21)}$

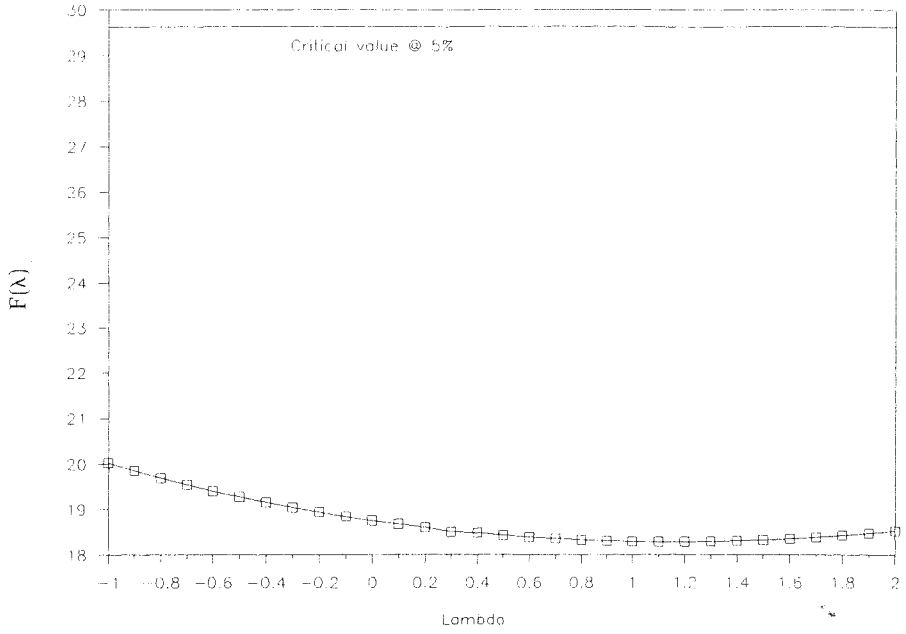


Figure 5.B.7 CRESSIE-READ STATISTICS TESTING THE HYPOTHESIS OF SYMMETRICAL TREATMENT OF MAJORITY AND MINORITY AUDITORS FOR BLACK/WHITE PAIRS, WASHINGTON, D.C.

$F(\lambda) = \chi^2_{(1)}$

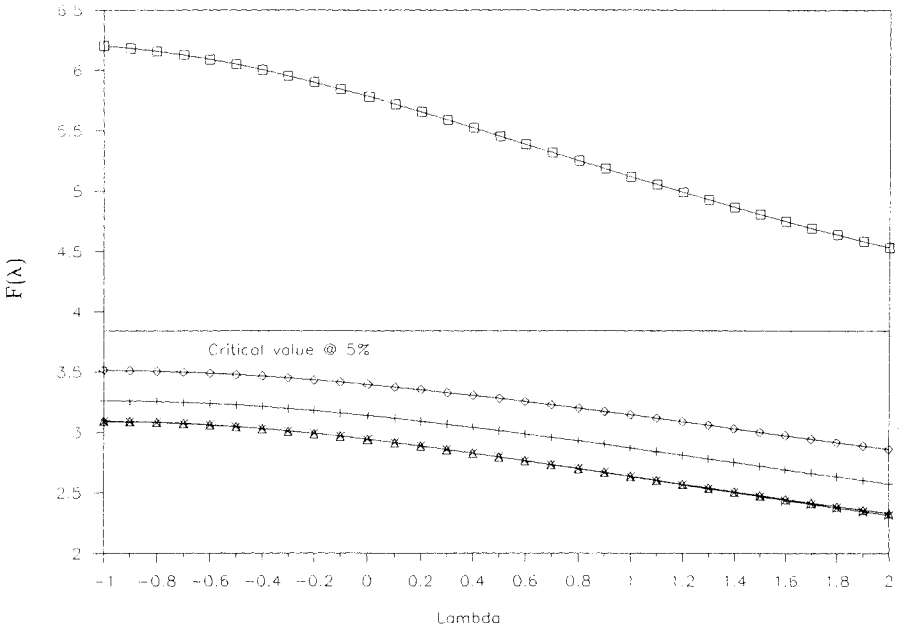


Figure 5.B.8 CRESSIE-READ STATISTICS TESTING THE HYPOTHESIS OF SYMMETRICAL TREATMENT OF MAJORITY AND MINORITY AUDITORS FOR BLACK/WHITE PAIRS, CHICAGO

$F(\lambda) = \chi^2_{(1)}$

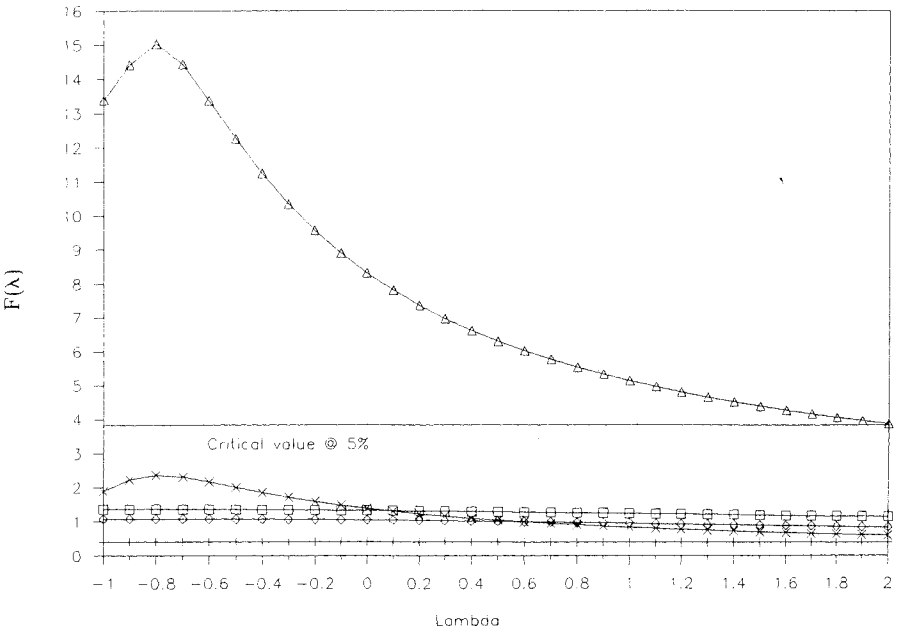


Figure 5.B.9 CRESSIE-READ STATISTICS TESTING THE HYPOTHESIS OF SYMMETRICAL TREATMENT OF MAJORITY AND MINORITY AUDITORS FOR ANGLO/HISPANIC PAIRS, SAN DIEGO

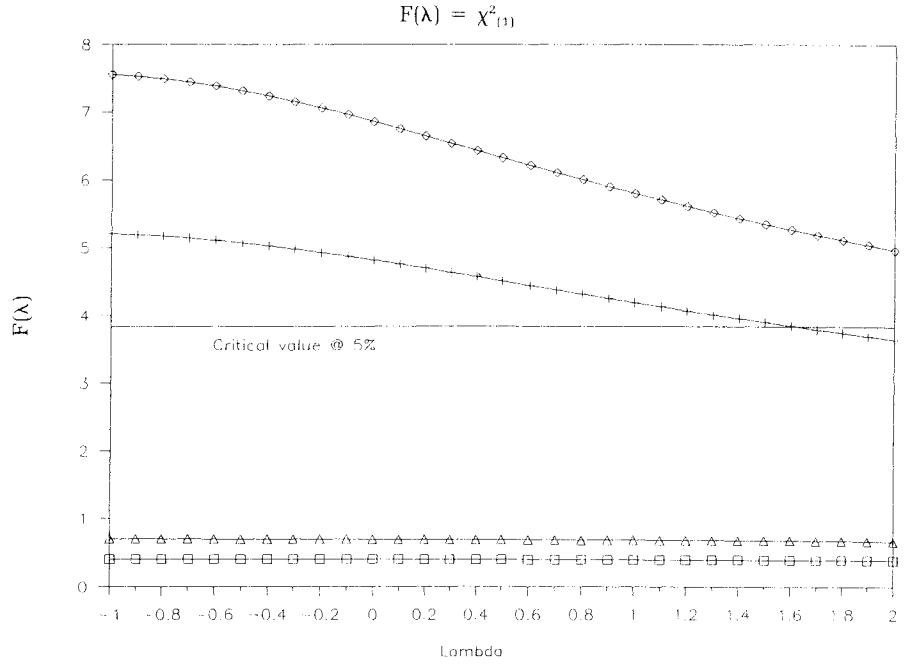


Figure 5.B.10 CRESSIE-READ STATISTICS TESTING THE HYPOTHESIS OF SYMMETRICAL TREATMENT OF MAJORITY AND MINORITY AUDITORS FOR HISPANIC/ANGLO PAIRS, CHICAGO

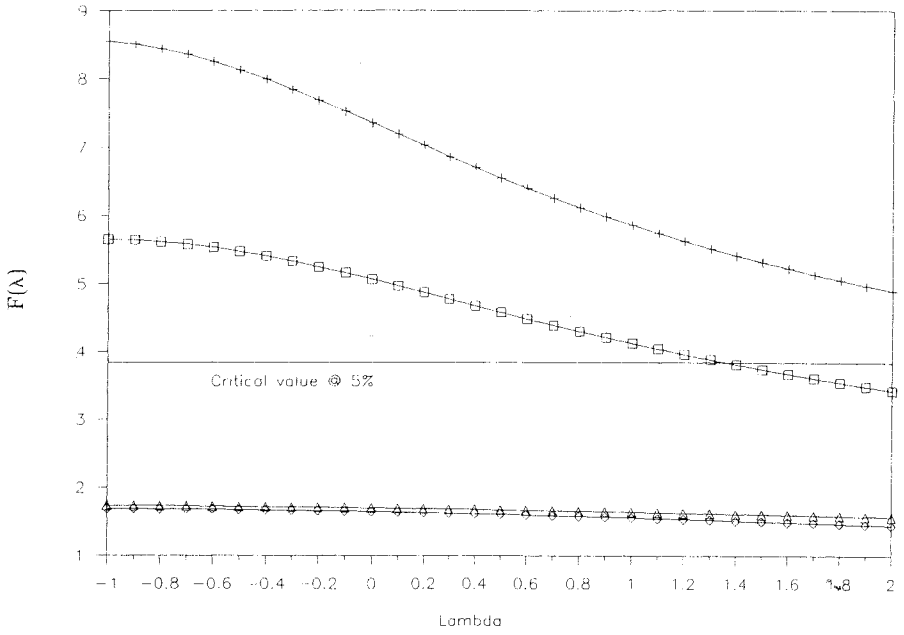


Figure 5.B.11 CRESSIE-READ STATISTICS TESTING THE HYPOTHESIS OF SYMMETRICAL TREATMENT OF MAJORITY AND MINORITY AUDITORS FOR BLACK/WHITE PAIRS AGGREGATED BY CITY

$F(\lambda) = \chi^2_{(1)}$

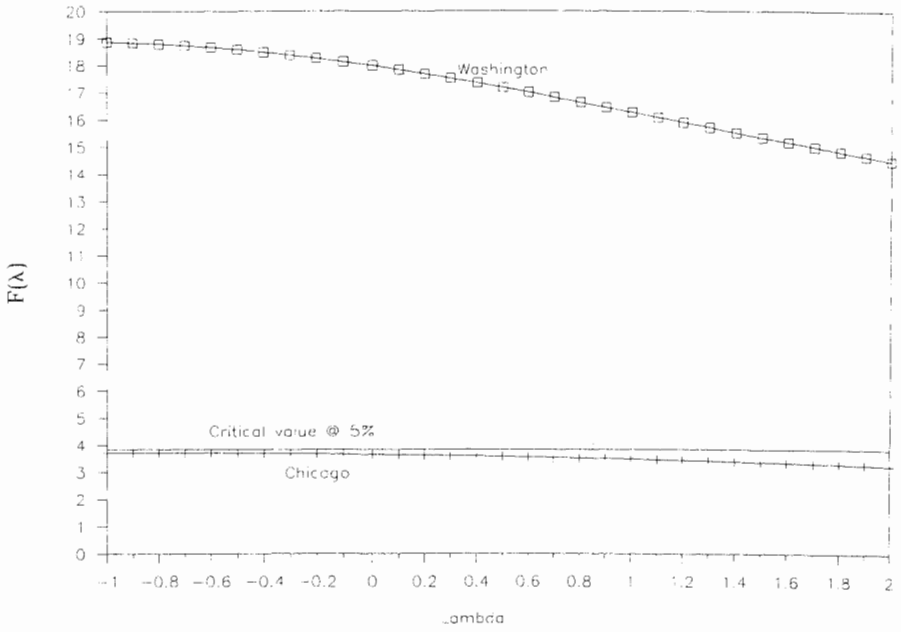
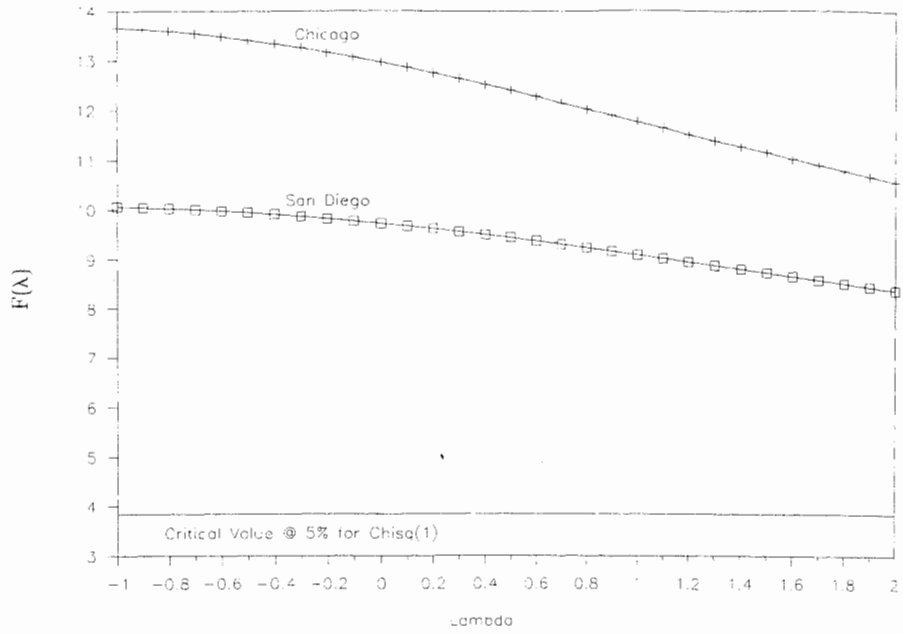


Figure 5.B.12 CRESSIE-READ STATISTICS TESTING THE HYPOTHESIS OF SYMMETRICAL TREATMENT OF MAJORITY AND MINORITY AUDITORS FOR ANGLO/HISPANIC PAIRS AGGREGATED BY CITY

$F(\lambda) = \chi^2_{(1)}$



APPENDIX 5.C

EMPIRICAL BAYES METHODS AND THE SIGN TEST

Empirical Bayes procedures postulate a distribution of outcome probabilities $F(\underset{\sim}{P})$. These can arise from heterogeneity across pairs in the match of auditors or in the firms they sample. For multinomial data, the Dirichelet prior is a conjugate prior widely used in the literature (see, e.g. Good 1965 or DeGroot 1970).

The density of $\underset{\sim}{P}$ is

$$f(\underset{\sim}{P}) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma(\Sigma\alpha_i)} \prod_{i=1}^{K} P_i^{\alpha_i - 1} \qquad \sum_{i=1}^{K} P_i = 1, \alpha_i \geq 0 \ \forall_i .$$

Γ is the gamma function. The density for the multinomial likelihood assuming random sampling with samples of size N and with N_i observations in cell i is

$$f = \binom{N}{N_1, N_2, \ldots N_K} \prod_{i=1}^{K} P_i^{N_i}.$$

Ignoring inessential constants, the expected probability given observation vector $(N_1, \ldots, N_K)$ is

$$E(P_1^{N_1} P_2^{N_2} \ldots P_K^{N_K}) = \frac{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \cdot \frac{\prod_{i=1}^{K} \Gamma(\alpha_i + N_i)}{\Gamma\left(\sum_{i=1}^{K} (\alpha_i + N_i)\right)}.$$

This expression is the basis for the empirical Bayes material we report below. The test of the hypothesis of equality in mean probabilities for outcome i and outcome j is a test of $\alpha_i = \alpha_j$, $i \neq j$.

$$E(P_i) = \alpha_i^* \Big/ \sum_{j=1}^{K} \alpha_j^*$$

$$\text{Var}(P_i) = (\alpha_i^*)\left(\sum_{j=1}^{K} \alpha_j^* - \alpha_i^*\right)\Big/\left(\sum \alpha_i^*\right)^2\left(\sum \alpha_j^* + 1\right).$$

where

$$\alpha_i^* = \alpha_i + N_i.$$

The Sign Test

Details of this test are given in Pratt and Gibbons (1981) or Lehmann (1986). Here we illustrate its computation by way of an example. Consider Chicago black audit pair 3. Of 44 interviews, only four firms treated pairs differently. $N_3 = 3$ (W yes; B no) and $N_4 = 1$ (W no; B yes). The conditional sign test proceeds *conditional* on $N_3 + N_4 = 4$.

Under the null that $P_3 = P_4$ ($\theta = \frac{1}{2}$), the probability of j outcomes where whites are hired and blacks are not is

$$\Pr(j \text{ whites hired} \mid N_3 + N_4 = 4) = \binom{4}{j}(1/2)^4.$$

Suppose that we seek values c_1 and c_2, $c_1 \leq c_2$, such that if $\theta = \frac{1}{2}$,

$$\Pr(j \leq c_1 \text{ or } j \geq c_2 \mid \theta = 1/2, N_3 + N_4 = 4) = .10$$

which is the probability of a type I error. Thus c_1 and c_2 are critical values.

Because of the binomial nature of j, in general it is not possible to find values of c_1 and c_2 that exactly set the type I error to 10%. If we set $c_1 = 0$ and $c_2 = 4$, the error rate is 12.5%. To attain exact error rates of 10%, we must randomize. Thus, if 80% of the time we reject if $j = 0$ or $j = 4$, we produce a randomized version of the test with the desired size (type I error rate). It can be shown that this test is uniformly most powerful and unbiased. The higher the randomized rate, the more likely are we to reject the null that $\theta = \frac{1}{2}$ if $j = 0$ or $j = 4$. In this sense, the test produces results closer to those obtained from conventional non-randomized tests. Observe that there is *no* 10% nonrandomized test in this example.

Table 5.C.1 LIKELIHOOD RATIO TESTS OF SYMMETRY

Study	Site	Outcome	Test Statistic $\chi^{2*}_{(1)}$
UIBW	Washington, D.C.	Get - job	17.90
UIBW	Chicago	Get - job	3.1
UIAH	Chicago	Get - job	11.04
UIAH	San Diego	Get - job	8.32

*Significant at the 5 percent level (critical value for $\chi^2_{(1)} = 3.84$).
Notes: UIBW = Urban Institute black/white audit; UIAH = Urban Institute Anglo/Hispanic audit. An interesting feature of the likelihood maximizing values of the α_i for each site is that the implied variances of α_i are very small, indicating that homogeneity among audit pairs is a valid description of the data.

APPENDIX 5.D

A MODEL OF THE HIRING DECISION

In our companion paper (Heckman and Siegelman 1993) we develop the following model of firm hiring decisions more completely.

The Model

Let $\underset{\sim}{C}^{MIN}$ be the skill vector of a minority member and $\underset{\sim}{C}^{MAJ}$ be the skill vector of a majority group member. Let $\underset{\sim}{\gamma}^{MIN}$ be a firm's evaluation vector of the productivity of attributes of minorities and $\underset{\sim}{\gamma}^{MAJ}$ be the firm's evaluation of the productivity of attributes of majority workers.[1] A constant is included among the elements of the C. The γ coefficient associated with the constant is a measure of the perceived productivity of the demographic group. Differences among groups could be due to animus-based discrimination. Thus, $\underset{\sim}{\gamma}^{MIN}\,\underset{\sim}{C}^{MIN}$ (a number) is the perceived increment to firm output from hiring a minority applicant and $\underset{\sim}{\gamma}^{MAJ}\,\underset{\sim}{C}^{MAJ}$ is the perceived contribution to firm productivity of a majority applicant. Assume that both applicants must be offered the same wage and impose the same hiring costs, K.

For a firm processing workers sequentially without recall of previous applicants, no explicit comparison of pairs of workers is made. In this case, a majority applicant is hired if $K \leq \underset{\sim}{\gamma}^{MIN}\,\underset{\sim}{C}^{MIN}$ and the majority group member shows up before any minority applicant.[2] Similarly, $K \leq \underset{\sim}{\gamma}^{MAJ}\,\underset{\sim}{C}^{MAJ}$ characterizes the case when a minority group member is hired. In the UI studies, inequalities may hold at a particular firm for each member of an audit pair if the first person offered the job declines it, so both persons may be offered a job.[3]

Differences between components of $\underset{\sim}{\gamma}^{MAJ}$ and $\underset{\sim}{\gamma}^{MIN}$ may arise because of discrimination in evaluating specific attributes or against a group as a whole (recall that $\underset{\sim}{C}$ includes a constant). Alternatively, discrimination might reflect statistical information processing in the event of

incomplete information about $\underset{\sim}{C}$, if the same observed characteristics convey different productivity information about majority and minority group members. A crucial unstated assumption in the UI analysis is that the absence of discrimination implies that $\underset{\sim}{\gamma}^{MAJ} = \underset{\sim}{\gamma}^{MIN}$ for each firm.[4] Audit pair studies as currently conducted cannot distinguish between animus-based discrimination and statistical discrimination, although it is of scientific and policy interest to do so.

Even if on average all firms treat majority and minority group members identically, firms may still differ in their evaluations of bundles of characteristics. Thus it is plausible that $\underset{\sim}{\gamma}$ is a realization or a draw from a population distribution

$$P(\underset{\sim}{\Gamma} \leq \underset{\sim}{\gamma}) = F(\underset{\sim}{\gamma}).$$

The characteristics of applicant pairs are drawings from the distributions $G(\underset{\sim}{c}^{MIN})$ and $H(\underset{\sim}{c}^{MAJ})$ and $\underset{\sim}{C}^{MIN}$ and $\underset{\sim}{C}^{MAJ}$ are assumed to be statistically independent in the population. For simplicity, we assume that all firms offer the same wage, say as a consequence of sample design (i.e., by the restriction of samples to low wage markets). If some firms place a large, negative weight on the constant for minority group members, they are unlikely to hire minorities. As the weight becomes large in absolute value, the probability of hiring a minority group member becomes arbitrarily small. Such firms would not hire a minority at any finite price.

In the labor market at large, if job seekers randomly sample firms, and there are many firms and many workers of both types, the probability that a firm will offer a job to a randomly sampled minority group member in small time interval Δt is

$$\text{Prob}(\underset{\sim}{\gamma}^{MIN}\, \underset{\sim}{C}^{MIN} > K)\lambda_{MIN}\,(\Delta t),$$

where λ_{MIN} is the rate of arrival of minorities to the firm. Similarly, the probability that the firm will offer a job to a majority group member is

$$\text{Prob}(\underset{\sim}{\gamma}^{MAJ}\, \underset{\sim}{C}^{MAJ} > K)\, \lambda_{MAJ}\,(\Delta t).$$

Under standard conditions, the probability that both a majority and minority member arrive at the firm in a small time interval Δt is λ_{MIN} $\lambda_{MAJ}(\Delta t)^2$, which is negligibly small. Assuming that $\underset{\sim}{\Gamma}$ and $\underset{\sim}{C}$ are statistically independent, the job offer rate to minority group members is higher compared to that of majority group members (a) the higher their rate of arrival at firms relative to majority members, (b) the greater their skill endowments, and (c) the higher is $\underset{\sim}{\gamma}^{MIN}$ compared to $\underset{\sim}{\gamma}^{MAJ}$.

Observe that minority job offer rates depend on the distribution of minority skills, the distribution of firm skill evaluation functions and the rate of arrival of minorities to vacancies. In the actual labor market, discrepancies in hiring rates between majority and minority group members can arise from any of these sources, only some of which plausibly constitute labor market discrimination. For example, minorities may do worse because they have lower skills ($\underset{\sim}{C}$), or lower firm encounter rates, which may not reflect employment or hiring discrimination.[5] Alternatively, minorities may fare worse because of differences in the way firms evaluate minority and majority members of equal skill.

The Urban Institute studies standardize the search component of job arrival rates by sending one minority member and one majority group applicant to the same firm in a short time period. Thus no information is gained about sources of racial differences in λ, since all attention is focused on the job offer rate to a given applicant pool of at least two workers.

In terms of this notation, the goal of the audit pair studies is to separate the effects of $\underset{\sim}{\gamma}$ and $\underset{\sim}{C}$ (firm evaluations and worker skills) on hiring. To test if $\underset{\sim}{\gamma}^{MIN} \neq \underset{\sim}{\gamma}^{MAJ}$, the UI studies attempt to align $\underset{\sim}{C}^{MIN}$ and $\underset{\sim}{C}^{MAJ}$ to a common value $\underset{\sim}{c}^*$. That is, the studies begin by matching the characteristics of the testers in an audit pair (including height, verbal facility, previous job history, and so on). If all characteristics could be aligned and wages and hiring costs were standardized across firms, and $\underset{\sim}{\gamma}^{MAJ} = \underset{\sim}{\gamma}^{MIN}$ for each audit trial at each firm, then

$$\text{Prob}(\underset{\sim}{\gamma}^{MAJ}\, \underset{\sim}{c}^* > K) = 0 \text{ or } 1$$

and

$$\text{Prob}(\underset{\sim}{\gamma}^{MIN}\, \underset{\sim}{c}^* > K) = 0 \text{ or } 1.$$

When all the characteristics of two applicants in an audit pair are perfectly matched, it is trivially true that the only way the members of the pair could be treated differently would be if $\underset{\sim}{\gamma}^{MAJ} \neq \underset{\sim}{\gamma}^{MIN}$. This model justifies the test of discrimination used in the UIBW study: $P_3 = P_4 = 0$.[6]

It seems more plausible, however, that only a subset of worker characteristics can be aligned by any pair matching procedure. In this case, it is fruitful to decompose the vector of characteristics relevant for hiring into observed and unobserved components, so that $\underset{\sim}{C} = (\underset{\sim}{C}_o, \underset{\sim}{C}_u)$. The audit pair researchers can control $\underset{\sim}{C}_o$, the characteristics they observe, but they cannot control $\underset{\sim}{C}_u$. Thus, for any two "paired" appli-

cants, only a *part* of their productivity—the $\underset{\sim}{C}_o$ part of $\underset{\sim}{\gamma}_o\underset{\sim}{C}_o$—can be observed and aligned. The unobserved and uncontrolled productivity is $\underset{\sim}{\gamma}_u\underset{\sim}{C}_u$. $\underset{\sim}{C}_u$ is assumed to be known to the firm but not to those designing an audit pair study.

This interpretation assumes that firms consider a much wider variety of characteristics than the audit designers can observe or control. This assumption seems valid in light of our current factual ignorance about the way matching works in the labor market, the great heterogeneity across tasks at firms and in firm skill requirements, and the enormous heterogeneity in skills among persons. At the end of this appendix, we briefly discuss the implications of an alternative assumption about the information possessed by the firm implicitly used in UIBW.

In the absence of discrimination on the part of firms, the probability that a minority is offered a job at a particular firm given $\underset{\sim}{C}_o^{\text{MIN}} = \underset{\sim}{C}_o^{\text{MAJ}} = c_o$ and firm productivity and cost vectors ($\underset{\sim}{\gamma}$, K) is

$$\text{Prob}(\underset{\sim}{\gamma}_u\ \underset{\sim}{C}_u^{\text{MIN}} > K - \underset{\sim}{\gamma}_o\ \underset{\sim}{c}_o \mid \underset{\sim}{\gamma}_u, \underset{\sim}{\gamma}_o, K, \underset{\sim}{c}_o),$$

while the probability that the majority member is offered a job is

$$\text{Prob}(\underset{\sim}{\gamma}_u\ \underset{\sim}{C}_u^{\text{MAJ}} > K - \underset{\sim}{\gamma}_o\ \underset{\sim}{c}_o \mid \underset{\sim}{\gamma}_u, \underset{\sim}{\gamma}_o, K, \underset{\sim}{c}_o).$$

These rates will be equal for all values of K and c_o if and only if

$$\underset{\sim}{\gamma}_u\ \underset{\sim}{C}_u^{\text{MAJ}} \text{ and } \underset{\sim}{\gamma}_u\ \underset{\sim}{C}_u^{\text{MIN}}$$

have the same conditional distribution.[7] Ironically, the two conditional job offer probabilities may be *more* unequal than the job offer probabilities that arise from *not* standardizing on the observable characteristics, c_o. For minority applicants at a firm with skill evaluation vector $\underset{\sim}{\gamma}$, this probability is

$$\Pr(\underset{\sim}{\gamma}\ \underset{\sim}{C}^{\text{MIN}} > K \mid \underset{\sim}{\gamma}, K)$$

and for majority applicants

$$\Pr(\underset{\sim}{\gamma}\ \underset{\sim}{C}^{\text{MAJ}} > K \mid \underset{\sim}{\gamma}, K).$$

The same can be said for the counterparts to these probabilities for the entire population of firms. That is, allowing for the distribution of $\underset{\sim}{\gamma}$ across firms, the population hiring rate for minority group members is

$$\Pr(\underset{\sim}{\gamma}\ \underset{\sim}{C}^{\text{MIN}} > K \mid K)$$

and for majority group members it is

$$\Pr(\underset{\sim}{\gamma}\ \underset{\sim}{C}^{MAJ} > K \mid K)$$

which are obtained by integrating the previous expressions that condition on $\underset{\sim}{\gamma}$ with respect to the population distribution of $\underset{\sim}{\gamma}$. The population counterparts to the conditional distributions (with respect to c_o) are

$$\Pr(\underset{\sim}{\gamma}_u\ \underset{\sim}{C}^{MIN} > K - \underset{\sim}{\gamma}_o\ \underset{\sim}{c}_o \mid K,\ \underset{\sim}{c}_o)$$

and

$$\Pr(\underset{\sim}{\gamma}_u\ \underset{\sim}{C}^{MAJ} > K - \underset{\sim}{\gamma}_o\ \underset{\sim}{c}_o \mid K,\ \underset{\sim}{c}_o)$$

obtained by integrating the previously given conditional probabilities with respect to the population distribution of $\underset{\sim}{\gamma}$, invoking independence between $\underset{\sim}{C}$ and $\underset{\sim}{\gamma}$.

There is no guarantee that standardizing on (conditioning on) a subset ($\underset{\sim}{C}_o$) of the components of $\underset{\sim}{C}$ makes the difference (or ratio) between majority and minority hiring rates smaller or larger than the difference (or ratio) between the population rates that do not condition on $\underset{\sim}{c}_o$. By standardizing on observed characteristics, it is possible to greatly exaggerate the role of differences in productivity characteristics that play only a minor role in actual labor markets, a phenomenon we illustrate below. Evidence of exchangeable treatment $P_3 = P_4$ (as defined in the text) is consistent with $\underset{\sim}{\gamma}^{MAJ} = \underset{\sim}{\gamma}^{MIN}$ at each firm, or at least in equality in the population distributions of $\underset{\sim}{\gamma}^{MAJ}$ and $\underset{\sim}{\gamma}^{MIN}$, and equality in the distributions of $\underset{\sim}{C}_u^{MAJ}$ and $\underset{\sim}{C}_u^{MIN}$. However, a test of $P_3 = P_4$ has no power against an alternative in which $\underset{\sim}{\gamma}_o^{MAJ} = \underset{\sim}{\gamma}_o^{MIN}$ but $\underset{\sim}{\gamma}_u^{MAJ} \neq \underset{\sim}{\gamma}_u^{MIN}$, and the distribution of $\underset{\sim}{C}_u^{MAJ}$ does not equal the distribution of $\underset{\sim}{C}_u^{MIN}$, but the distributions of $\underset{\sim}{\gamma}_u^{MAJ}\underset{\sim}{C}_u^{MAJ}$ and $\underset{\sim}{\gamma}_u^{MIN}\underset{\sim}{C}_u^{MIN}$ are identical (i.e., the $\underset{\sim}{\gamma}_u$ and $\underset{\sim}{C}_u$ distributions "offset" each other). Thus, even in the presence of discrimination, P_3 can equal P_4.

A Normal Characteristics Model

To illustrate the problems that partial standardization of tester characteristics can create, we present a simple example of the hiring model. We use this example to illustrate two points. First, standardizing on only a *subset* of relevant productivity characteristics may greatly exaggerate the measured level of disparity in racial hiring rates compared with what occurs in actual market settings. Second, the skill *level* at which standardization occurs greatly affects the results.

Depending on the level of skill at which one standardizes, it is possible to produce a wide array of disparities in majority vs. minority hiring rates.

Suppose that there are only two productivity characteristics that are relevant to the hiring decision (C_o, C_u), and both are normally distributed in the population. For simplicity, and without loss of any fundamental generality, the components are assumed to be statistically independent of each other. Both components are observed by the employer, but the testing organization observes only C_o and aligns tester pairs only on this characteristic. For example, suppose that the two relevant characteristics are height and "motivational level." We assume that researchers observe (and standardize) the height of both members in an audit pair, but do not observe the auditors' motivation levels.[8] We further assume that firms observe both auditors' heights and their motivation levels, and that the latter, because they are not standardized, may differ between the two auditors.

In our model, firms turn worker characteristics into expected output by means of an evaluation vector, $\underset{\sim}{\gamma}$ (which has two components, γ_o and γ_u). The components of the vector can be thought of as the expected marginal product of the worker characteristics in the $\underset{\sim}{C}$ vector. Thus, in the above example, γ_o is the expected marginal product of height—the extra output that is produced by one extra inch of height. The expected marginal product of an applicant with the vector of characteristics $\underset{\sim}{c}$ $(= (\underset{\sim}{c}_o, \underset{\sim}{c}_u))$ is $\underset{\sim}{\gamma}\underset{\sim}{c}$, a scalar.

To keep this example simple, suppose that no firms discriminate and that all firms place the same value on C_o and C_u in assessing productivity. Let us assign that value at $\gamma_o = \gamma_u = 1$ for both majority and minority attributes so there is no discrimination across firms. Later we introduce discriminatory behavior.

In the population at large,

$$C_o^{MIN} \sim N(\mu_o^{MIN}, \sigma_o^{2MIN}),$$

$$C_o^{MAJ} \sim N(\mu_o^{MAJ}, \sigma_o^{2MAJ}),$$

$$C_u^{MIN} \sim N(\mu_u^{MIN}, \sigma_u^{2MIN}), \text{ and}$$

$$C_u^{MAJ} \sim N(\mu_u^{MAJ}, \sigma_u^{2MAJ}),$$

where "$\sim$" denotes "is distributed as" and $N(a, b)$ denotes a normal random variable with mean a and variance b. σ_o^{2MIN} is the *variance* in observables for minorities. Let Z denote a standard normal random variable: $Z \sim N(0, 1)$. Thus the productivity of minority members is

$$P^{MIN} = C_o^{MIN} + C_u^{MIN}, \text{ where} \tag{1}$$

$$P^{MIN} \sim N(\mu_o^{MIN} + \mu_u^{MIN}, \sigma_o^{2MIN} + \sigma_u^{2MIN})$$

while

$$P^{MAJ} = C_o^{MAJ} + C_u^{MAJ}, \text{ with} \tag{2}$$

$$P^{MAJ} \sim N(\mu_o^{MAJ} + \mu_u^{MAJ}, \sigma_o^{2MAJ} + \sigma_u^{2MAJ}).$$

Suppose that firms face a fixed cost of hiring, which includes the wage paid plus any costs associated with the hiring transaction itself. Then it follows that in the population at large, the probability that a randomly selected minority will be hired is

$$\text{Prob}(P^{MIN} > K). \tag{3(a)}$$

Because both components are normally distributed, their sum is as well, so we can write

$$\text{Prob}(P^{MIN} > K) = \text{Prob}(C_o^{MIN} + C_u^{MIN} > K)$$

$$= \text{Prob}\left(Z > \frac{K - \mu_o^{MIN} - \mu_u^{MIN}}{(\sigma_o^{2MIN} + \sigma_u^{2MIN})^{1/2}}\right)$$

$$= \Phi\left(\frac{(\mu_o^{MIN} + \mu_u^{MIN}) - K}{(\sigma_o^{2MIN} + \sigma_u^{2MIN})^{1/2}}\right),$$

where Φ is the cumulative distribution function for Z. For the majority population, the comparable hiring rate is

$$\text{Prob}(P^{MAJ} > K) = \Phi\left(\frac{(\mu_o^{MAJ} + \mu_u^{MAJ}) - K}{(\sigma_o^{2MAJ} + \sigma_u^{2MAJ})^{1/2}}\right). \tag{3(b)}$$

Thus, *ceteris paribus*, the higher the mean productivity, the higher the probability of hiring a minority or majority applicant. Higher variability in skill ($\sigma_o^2 + \sigma_u^2$) lowers the hiring probability if $\mu_o + \mu_u > K$, while greater variability increases the hiring probability if $\mu_o + \mu_u < K$.[9] The audit pair methodology aligns values of $C_o = c_o$ for both groups. Then the *conditional* (on $C_o = c_o$) hiring rates for majority and minority group members are, respectively,

$$\text{Prob}(P^{MIN} > K \mid C_o^{MIN} = c_o) = \text{Prob}(C_u^{MIN} > K - c_o \mid C_o^{MIN} = c_o)$$

$$= \Phi\left(\frac{\mu_u^{MIN} + c_o - K}{\sigma_u^{MIN}}\right) \tag{4(a)}$$

and

$$\text{Prob}(P^{MAJ} > K \mid C_o^{MAJ} = c_o) = \Phi\left(\frac{\mu_u^{MAJ} + c_o - K}{\sigma_u^{MAJ}}\right). \qquad 4(b)$$

At issue is the comparison of the ratio of 4(*a*) to 4(*b*) (the standardized hiring rate) to the population hiring ratio (3(*a*) divided by 3(*b*)).

In the population the ratio of the minority to the majority hiring rate is thus

$$h = \frac{\Phi\left(\frac{\mu_o^{MIN} + \mu_u^{MIN} - K}{(\sigma_o^{2MIN} + \sigma_u^{2MIN})^{1/2}}\right)}{\Phi\left(\frac{\mu_o^{MAJ} + \mu_u^{MAJ} - K}{(\sigma_o^{2MAJ} + \sigma_u^{2MAJ})^{1/2}}\right)}. \qquad 5(a)$$

The ratio of hiring rates when the auditors have been matched on the observable characteristic (but not on the unobservable characteristic) is

$$h(c_o) = \frac{\Phi\left(\frac{\mu_u^{MIN} + c_o - K}{\sigma_u^{MIN}}\right)}{\Phi\left(\frac{\mu_u^{MAJ} + c_o - K}{\sigma_u^{MAJ}}\right)}. \qquad 5(b)$$

Using 5(*a*) and 5(*b*), the intuitive remarks made above can be justified. Suppose that the observable characteristic, height, is the major source of productivity variability. By this we mean that the variance of height is greater than the variance of motivation ($\sigma_o^2 > \sigma_u^2$) for both majority and minority group members. Letting σ_o^{2MIN} and σ_o^{2MAJ} become arbitrarily large, holding the other parameters fixed, both the numerator and the denominator of 5(*a*) approach one-half, so h approaches one. In this case C_o "swamps" C_u. Conditioning on $C_o = c_o$ eliminates this effect. Suppose that $\mu_u^{MIN} = \mu_u^{MAJ} = 0$. If $c_o > K$, and σ_u^{MIN} and σ_u^{MAJ} becomes arbitrarily small, but in such a way that the minority variance is relatively bigger than the majority variance, i.e.,

$$\lim_{\substack{\sigma_u^{MIN}\to 0 \\ \sigma_u^{MAJ}\to 0}} \frac{\sigma_u^{MIN}}{\sigma_u^{MAJ}} > 1,$$

a standard argument verifies that $h(c_o)$ approaches infinity. That is, minority hiring rates become infinitely large relative to majority hiring rates. To take a less extreme case, set $c_o = \mu_o^{MAJ} = \mu_o^{MIN}$ so the numerator within each probability in 5(*b*) is the same as in 5(*a*) but the denominators are different. Even small differences in the numerators, which are "swamped" in 5(*a*), become large in 5(*b*) if $\sigma_u^{MIN} = \sigma_u^{MAJ}$.

To gain further insight into the model, set $\mu_u = 0$ for both groups. For the special case $\mu_o^{MIN} = \mu_o^{MAJ} = K$, 3(*a*) and 3(*b*) reveal that in the population the hiring rate is the same for both majority and minority group members and equals ½. Thus *h*, the relative hiring rate, equals 1. Suppose that the component of variance for the unobservable differs between the two populations, with the minority trait more dispersed ($\sigma_u^{2MIN} > \sigma_u^{2MAJ}$). Then, depending on the value of c_o, $h(c_o)$ may be bigger or smaller than one. See figure 5.D.1, which plots $h(c_o)$ for the special case $\sigma_o^2 = 1$, $\sigma_u^{2MIN} = 1.5$, and $\sigma_u^{2MAJ} = 1$. For values of c_o less than zero, $h(c_o) > 1$. As c_o gets smaller, $h(c_o)$ becomes much bigger than one. As c_o becomes large, $h(c_o)$ approaches one. The smallest value of $h(c_o)$ is .89. For values of c_o above zero, $h(c_o) < 1$.[10] The choice of the level of c_o determines both the magnitude and the direction of the bias in the relative hiring rate as detected by audit pair methods. In this example, any relative hiring rate between .89 and infinity can be produced. Standardizing on only a subset of the relevant productivity characteristics may produce a severely distorted impression of actual labor market discrimination, which in this example does not exist in the labor market. It should be obvious that the same phenomenon of dramatic misrepresentation of population hiring rates by audit-measured hiring rates will appear if we reverse the assumption about the ordering of the variances of the unobservables ($\sigma_u^{2MAJ} < \sigma_u^{2MIN}$) or if the variances of the unobserved components are smaller than the variances of the observed components, Max(σ_0^{2MIN}, σ_0^{2MAJ}).

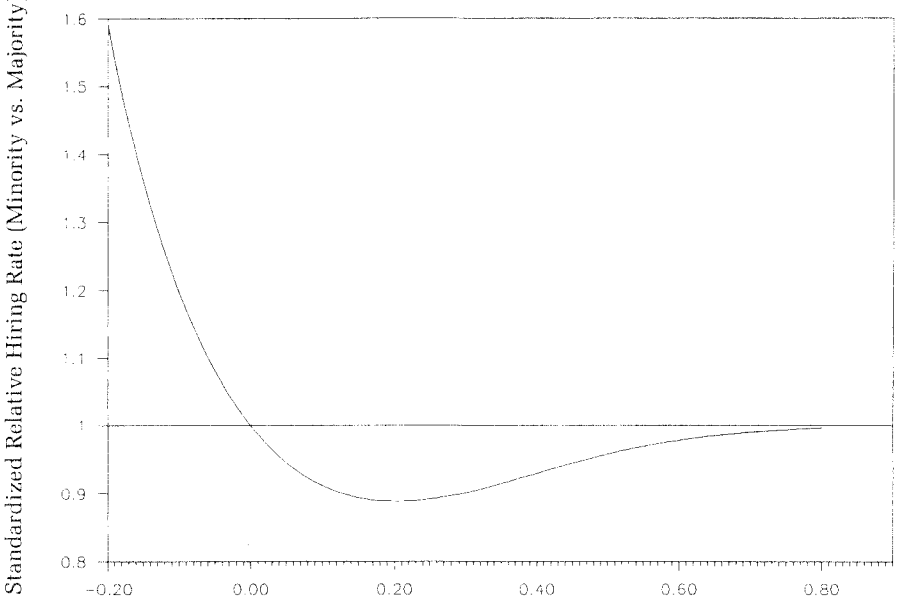
A somewhat less trivial example is produced when $K = 1$, $\mu_o^{MIN} = \mu_o^{MAJ} = 0.9$. For this case $h = .75$, so that there is a sizable disparity between minority and majority hiring rates in the population. Again, depending on the level at which the observable skill is standardized (c_o), the standardized hiring ratio can be set as large as we like and as small as .662 and can attain any value in between, as figure 5.D.2 shows.

The gist of the argument presented so far can be seen in the following example. Suppose that there are two audit pairs. In one, both the auditors are 5 feet tall. In the other both are 6′6″ tall. The model shows

Figure 5.D.1 STANDARDIZED RELATIVE HIRING RATE (MINORITY VS. MAJORITY) $h(c_o)$, AS A FUNCTION OF c_o, FOR:

$$K = \mu_u^{MIN} = \mu_u^{MAJ} = \mu_o^{MIN} = \mu_o^{MAJ} = 0;$$

$$\sigma_o^{2MIN} = \sigma_o^{2MAJ} = 1; \sigma_u^{2MIN} = 1.5, \sigma_u^{2MAJ} = 1$$



Level at which the observable characteristic is standardized (c_o)

$\text{Min}(h(c_o)) = 0.89$ at $c_o = 0.2$

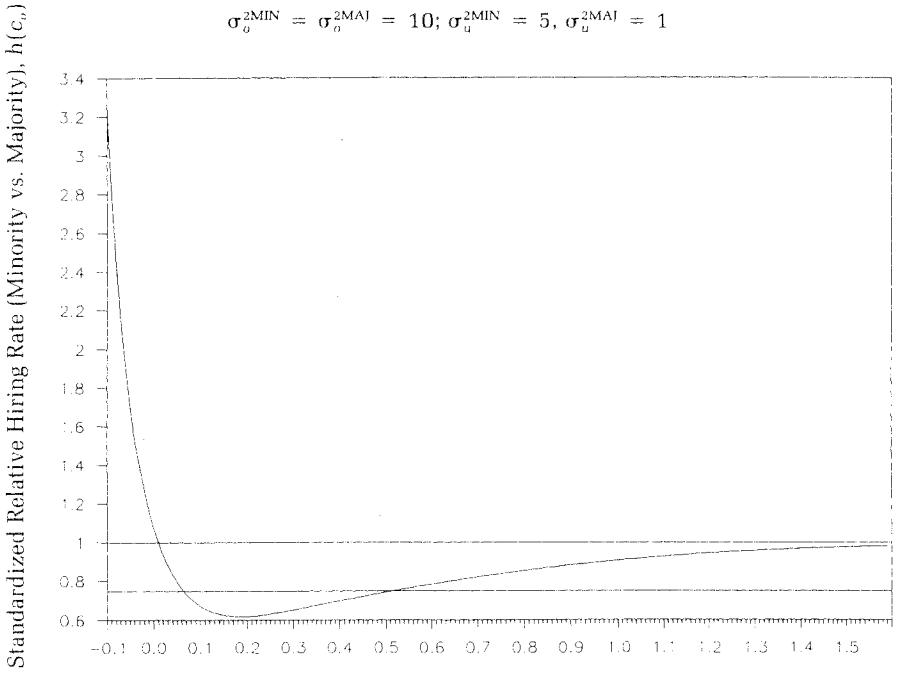
that under certain plausible conditions, the amount of "discrimination" (as measured by the difference in hiring rates within each of the two pairs) will be different for the two pairs.

Why should standardizing the level of the observable characteristic (say, height) at some value (c_o^*) below its mean favor black applicants, while standardization at a level above the mean favor whites? Recall that total productivity is derived from two components, height and motivation, only the first of which is observable by the researcher. Both height and motivation are observable by the firm that is evaluating job applicants, however, so when we use "unobservable," we refer only to what the researchers can observe. We further assume that minorities have a higher variance in the unobservable component than do whites, although the average motivation level is the same for both races.

Figure 5.D.2 STANDARDIZED RELATIVE HIRING RATE (MINORITY VS. MAJORITY) $h(c_o)$, AS A FUNCTION OF c_o, FOR:

$$K = 1;\ \mu_o^{MIN} = \mu_o^{MAJ} = 0.9;\ \mu_u^{MIN} = \mu_u^{MAJ} = 0;$$

$$\sigma_o^{2MIN} = \sigma_o^{2MAJ} = 10;\ \sigma_u^{2MIN} = 5,\ \sigma_u^{2MAJ} = 1$$



Level at which the observable characteristic is standardized (c_o)

$\text{Min}(h(c_o)) = 0.662$ at $c_o = 0.2$

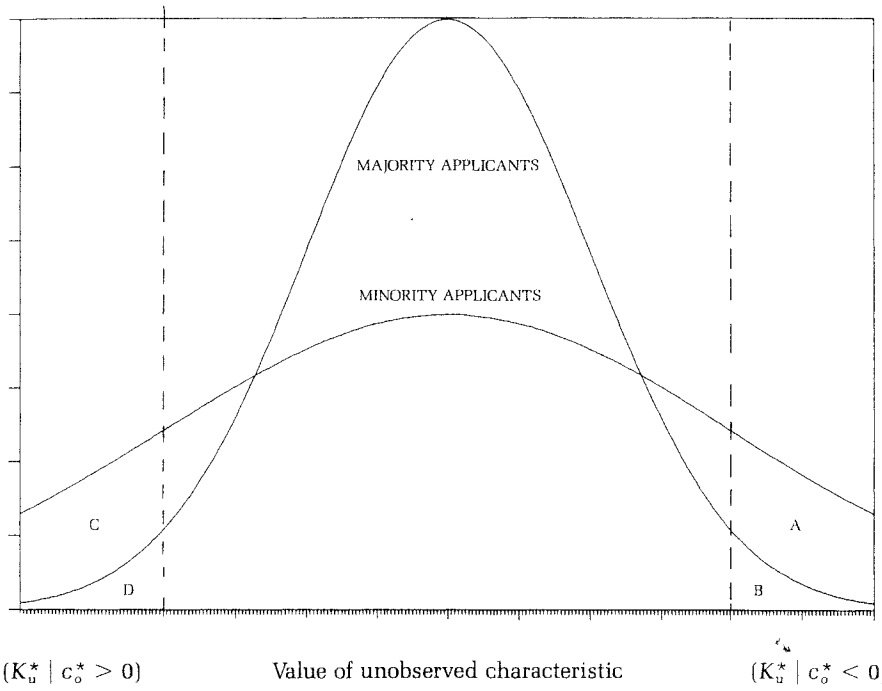
The model of hiring developed earlier suggested that there is some threshold value of productivity, K, such that any applicant whose productivity is above the threshold will be hired. Imagine that the experimenters pick some value for the observable component of total productivity. All auditors will have the identical value for the observable characteristic, height, which we will imagine is set at some value c_o^*, say 6 feet. In order for any 6-foot applicant to be hired, the sum of the two components must be greater than K, which implies that his motivation level must be greater than $(K - c_o^* = K - 6)$. In other words, conditional on a given height, the threshold for motivation (the component unobservable by the researchers) by itself is simply $K_u^* = K - c_o^*$.

Suppose we rescale so that the mean of the observable characteristic is now set to zero. If c_o^* is set below its mean of zero, then the threshold value for the unobservable variable, K_u^*, will be large and positive. All this says is that an applicant who is very weak in one of the two

dimensions must be correspondingly strong in the other in order to overcome his handicap and meet the hiring threshold. Even though minorities and whites both have the same average level of motivation, our earlier variance assumptions imply that there are relatively fewer whites with either extremely high or extremely low values of this variable. Thus, for $c_o^* < 0$, a higher fraction of minority applicants will have motivation greater than the threshold (K_u^*) than will white applicants. Fixing c_o less than zero will therefore tend to favor the group with the larger variance in unobservables—in this case, minority applicants. The reverse is true for a positive value of c_o. An employer considering a very tall applicant need only worry whether the worker has an extremely low motivation level. Those with unusually low (or unusually high) motivation are more common among applicants from the high-variance applicant pool (in this case, minorities) than for those from the low variance pool.

These points are illustrated graphically in figure 5.D.3, which depicts the distribution of "motivation" for majority and minority applicants. The threshold value for the unobserved characteristic (K_u^*)

Figure 5.D.3 EFFECTS OF STANDARDIZING THE OBSERVED CHARACTERISTIC AT DIFFERENT LEVELS ON RELATIVE HIRING RATES FOR MINORITY AND MAJORITY TESTERS



is the minimum value the applicant must have in order to be hired, conditional on some given value of height which has already been chosen by the auditors. The probability of hiring a randomly selected applicant from either of the two populations is simply the area under the relevant curve to the right of K_u^*. Suppose we set the standard height for the audit pairs very low, implying that K_u^* will be positive (so that applicants will need to be very highly motivated to be hired). Then the probability of hiring is greater for the minority applicant than for the majority applicant, as can be seen by comparing area $A + B$, the probability of hiring a minority applicant, to area B, the probability of hiring a white applicant. Conversely, if the audit pair members are both very tall, then K_u^* will be less than zero. Thus, the probability of being hired will be greater for the white applicant than for the minority applicant (compare $1 - (C + D)$ to $1 - D$)).

The following points are general features of the normal model. From 5(*b*), as long as $\sigma_u^{MIN} \neq \sigma_u^{MAJ}$, there always exists a c_o, c_o^* that sets $h(c_o^*) = 1$. That value is given by

$$c_o^* = \frac{\sigma_u^{MAJ}(\mu_u^{MIN} - K) - \sigma_u^{MIN}(\mu_u^{MAJ} - K)}{\sigma_u^{MIN} - \sigma_u^{MAJ}}.$$

Thus

$$h(c) > 1 \text{ for } c < c_o^* \text{ if } (\sigma_u^{MIN} > \sigma_u^{MAJ})$$

$$\text{or for } c > c_o^* \text{ if } (\sigma_u^{MIN} < \sigma_u^{MAJ}).$$

If unobserved characteristics are more heterogeneous in the minority population ($(\sigma_u^{MIN} - \sigma_u^{MAJ}) > 0$), it is always possible to find $h(c_o) > 1$ by picking a low enough level of c_o. That is, it is possible for audits to establish that there is no discrimination against minority members, even if such discrimination exists. This is evident from figures 5.D.1 and 5.D.2.

Comparing 5(*a*) and 5(*b*), it is clear that if $\sigma_u^{MIN} > \sigma_u^{MAJ}$, there exists a value of c_o such that $h(c_o) > h$. The proof entails use of L'Hospital's rule. It notes that for this case, $\lim_{c_o \to -\infty} h(c_o) \to \infty$, and by the continuity of $h(c_o)$, $h(c_o)$ assumes all values in the interval $(1, \infty)$. For the special case $\mu_u^{MIN} = 0$ and $\mu_u^{MAJ} = 0$, $h(c_o)$ reaches a global minimum at some $c_o > 0$. For the case $\mu_u^{MIN} = \mu_u^{MAJ}$ and $\mu_o^{MIN} = \mu_o^{MAJ}$ and $\sigma_o^{2MIN} + \sigma_u^{2MIN} > \sigma_o^{2MAJ} + \sigma_u^{2MAJ}$, it is always possible to find a value of c_o such that $h(c_o) > h$, since $h > 1$. For the case of $\sigma_u^{MIN} < \sigma_u^{MAJ}$ an obvious reversal of the previously stated results will hold.

Suppose that $h < 1$, so that the population hiring ratio exhibits discrimination against minorities. Regardless of the parameter values

that produce this result, there always exists a value of c_o such that $h(c_o) > h$. That is, it will always be possible to choose a skill-level for the observable variable such that the audits produce less evidence of discrimination than actually exists. This follows from the observation that $\lim_{c_o \to \infty} h(c_o) = 1$ for all values of the parameters (assuming $\sigma_u^{MAJ} > 0$ and $\sigma_u^{MIN} > 0$). The absence of discrimination found in some of the UI and Denver audits may be a consequence of the implicit choice of c_o.

Note, finally, an alternative interpretation of any pair study. Suppose that C_o and C_u have unit variances for both minority and majority members. Then the previous examples also apply to the following model of discrimination in hiring. Let σ_u^{MIN} be the *firm* valuation placed on C_u^{MIN} while σ_o^{MIN} is the firm evaluation placed on C_o^{MIN}. Similarly, σ_u^{MAJ} and σ_u^{MAJ} are the firm's evaluations of C_u^{MAJ} and C_o^{MAJ} respectively. All firms place the same valuation on the characteristics. The case $\sigma_u^{MIN} > \sigma_u^{MAJ}$ is thus equivalent to the case in which minority unobservables are valued more than majority unobservables. This alternative interpretation highlights a fundamental ambiguity in audit pair methodology for the normal example considered here. Greater variability in minority unobserved skills and greater market valuation of minority skills are indistinguishable. Every result stated in terms of differential variability in skills has a dual ordering in terms of differences in the market valuation for minority and majority skills. From the audit studies, one cannot distinguish variability in unobservables from discrimination.

THE CASE MOST FAVORABLE TO THE URBAN INSTITUTE INTERPRETATION OF THE AUDIT PAIR STUDIES

The case most favorable to audit pair methodology arises when the audit analysts have access to a larger set of productivity characteristics than firms use in making hiring decisions *and* they have access to large pools of audit partners so that near perfect matches can be made. In this case, audit pair members are perfectly aligned. Firm by firm, audit pair members should be treated identically in hiring decisions. That is the benchmark that was used in UIBW. But this assumption seems implausibly strong in light of our current ignorance about the nature of firms' hiring decisions and the small samples of potential auditors available in the UI studies.

Notes, appendix 5.D

1. The individual components of the evaluation vector, $\underline{\gamma}$, can be thought of as the marginal product of the associated component in the individual's skill vector, C. That is, the gamma vector maps an individual's skills (strength, reliability, etc.) into output.

2. We assume that because of its past history with job applicants, the firm has set K so as to equate the marginal benefits of making an offer to the current applicant (the immediate increase in output) with the expected marginal benefits of waiting for another applicant who may have higher productivity than the current candidate. For simplicity we assume a stationary evironment.

3. Note that this model explains the finding in the UI data that the first member of an audit pair who approaches a firm has a greater probability of receiving a job offer. The reason is that other applicants, some of whom are successful, may approach firms between the time that the first audit member leaves and the second applies.

4. Actually, all that is required under the definition of the absence of discrimination in the Anglo/Hispanic study, which defines equality of treatment by symmetry of treatment of majority and minority group members over a series of trials, is that the distributions of $\underline{\gamma}^{MAJ}$ and $\underline{\gamma}^{MIN}$ are equal. Thus, firms could draw a fresh $\underline{\gamma}$ each time a new applicant comes in. Although two applicants with identical bundles of characteristics ($\underline{C}$) may be treated differently in any individual instance, as long as the draws are from the same distribution for all workers, so that the two members of a pair are treated the same on average, there is no discrimination. We forego this generality here and assume that firms have stable preferences over attributes. As we point out below, the definition of discrimination in the black/white study requires a much stronger assumption, namely that $\underline{\gamma}^{MIN} = \underline{\gamma}^{MAJ}$. In Heckman and Siegelman (1993) we test for, and reject, this random-draw model of employment.

5. The extent to which skill differences or differences in search behavior are attributable to the labor market or to forces outside of the labor market is difficult to sort out with either audit pair or standard observational data. Lower skills for minorities may be a consequence of their lower market rate of return to skill or due to nonmarket factors. Lower search intensity by minorities can be rationalized in the same conflicting ways.

6. Operationally, this definition requires that $\text{Prob}(\tilde{\underline{\gamma}}^{MAJ}\, c^* > K \wedge \tilde{\underline{\gamma}}^{MIN}\, c^* \leq K) = 0$ so that majority and minority group members are treated alike. In the absence of statistical or animus-based discrimination, members of a matched pair must be treated the same at all firms and will both be hired or rejected depending only on the values of K and $\underline{\gamma}$ for the firm.

7. The conditioning is on c_o, K, $\underline{\gamma}_u$ and $\underline{\gamma}_o$.

8. In fact, all three of the studies did make efforts to control for demeanor, appearance, assertiveness, and other factors that might reflect motivation levels. We use this purely as an example of the kinds of characteristics that are difficult to control for.

9. This is trivially true algebraically: when the numerator of the term in brackets is positive ($\mu_o + \mu_u > K$), a large variance (in the denominator) makes the whole expression smaller. When the numerator is negative, an increase in the variance makes the entire expression larger. We offer a more intuitive explanation below.

10. Choices of c_o below zero are not implausible. If productivity is multiplicative in its components, c_o is log skill and can be negative, even if skill is positive.

References

Ayres, I. 1991. "Fair Driving: Gender and Race Discrimination in Retail Car Negotiations." *Harvard Law Review* 104 (4): 817–872.

Ayres, I., and P. Siegelman. 1993. "Race and Gender Discrimination in Negotiating for the Purchase of a New Car." Chicago: American Bar Foundation Working Paper, 1993.

Becker, G. 1975. *The Economics of Discrimination*. Chicago: University of Chicago Press.

Bishop, Y., S. Feinberg, and P. Holland. 1975. *Discrete Multivariate Analysis: Theory and Practice*.

Cain, G. 1986. "The Economic Analysis of Labor Market Discrimination: A Survey." In *Handbook of Labor Economics*, edited by O. Ashenfelter and R. Layard, vol. 1. Amsterdam: North Holland.

Cox, P. 1987. *Employment Discrimination*. New York: Garland Law Publishing.

Cressie, N., and N. Read. 1988. *Goodness of Fit Statistics for Discrete, Multivariate Data*. Berlin: Springer.

Cross, H., G. Kenny, J. Mell, and W. Zimmerman. 1990. *Employer Hiring Practices: Differential Treatment of Hispanic and Anglo Job Seekers*. Washington, D.C.: Urban Institute Press.

DeGroot, M. 1970. *Optimal Statistical Decisions*. New York: McGraw Hill.

Dipboye, R. 1975. "Relative Importance of Applicant Sex, Attractiveness, and Scholastic Standing in Evaluation of Job Applicant Resumes." *Journal of Applied Psychology* 60: 39–43.

Domencich, T., and D. McFadden. 1975. *Urban Travel Demand*. Amsterdam: North Holland.

Donohue, J., and P. Siegelman. 1991. "The Changing Nature of Employment Discrimination Litigation." *Stanford Law Review* 43(3): 983–1033.

Flinn, C., and J. Heckman. 1983. "Are Unemployment and Out of the Labor Force Behaviorally Distinct States." *Journal of Labor Economics* 1: 65–93.

Good, I.J. 1965. *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*. Cambridge: MIT Press.

Heckman, J., and P. Siegelman. 1993. "The Audit Pair Methodology for Assessing Labor Market Discrimination: A Critical Assessment." Unpublished manuscript, University of Chicago.

Hedges, L., and I. Olkin, *Statistical Methods for Meta-Analysis*, 1985, San Diego: Academic Press.

Holzer, H. 1987. "Informal Job Search and Black Youth Unemployment." *American Economic Review* 77(3): 446–452.

———. 1988. "Search Methods Used By Unemployed Youth." *Journal of Labor Economics* 6(1): 1–20.

James, F., and S.W. DelCastillo. 1991. "Measuring Job Discrimination by Private Employers Against Young Black and Hispanic Males Seeking

Entry Level Work in the Denver Metropolitan Area." Unpublished report. Denver: University of Colorado, Denver, March.

Lehmann, E. 1986. *Testing Statistical Hypotheses*. 2nd ed., New York: Wiley.

Leonard, J. 1984. "Anti-Discrimination or Reverse Discrimination: The Impact of Title VII, Affirmative Action, and Changing Demographics on Productivity." *Journal of Human Resources* 19: 145–169.

Lindzey, G., and E. Aronson. 1975. *The Handbook of Social Psychology*. 2nd ed., vol. 2. Reading, Mass.: Addison-Wesley.

Madden, J. 1987. "Gender and Race Differences in the Cost of Displacement: An Empirical Test of Discrimination in the Labor Market." *American Economic Review* 77: 246–251.

McFadden, D. 1973. "Conditional Logit Analysis of Qualitative Choice Behaviour." In *Frontiers of Econometrics*, edited by P. Zarembka. New York: Academic Press.

McIntyre, S., et al. 1980. "Preferential Treatment in Preselection Decisions According to Race and Sex." *Academy of Management Journal* 23.

Newman, J. 1978. "Discrimination in Recruitment: An Empirical Analysis." *Industrial and Labor Relations Review* 32: 15–23.

Pearson, E.S. 1950. "On Questions Raised by The Combination of Tests Based on Discontinuous Distributions." *Biometrika*, 37: 89–99.

Pratt, J., and J. Gibbons. 1981. *Concepts of Nonparametric Theory*. Berlin: Academic Press.

Rosenthal, R. 1976. *Experimenter Effects in Behavioral Research*. 2nd ed. New York: Irvington Publishers.

Thurstone, L. 1927. "A Law of Comparative Judgement." *Psychological Review* 34: 273–286.

Turner, Margery A., M. Fix, and R. Struyk. 1991. *Opportunities Denied, Opportunities Diminished: Racial Discrimination in Hiring*. Washington, D.C.: Urban Institute Press.

Van den Berg, G. 1991. "Results on the Rate of Arrival of Job Offers in a Search Model." *Review of Economic Studies* 62: 263–271.

Yinger, John. 1986. "Measuring Racial Discrimination With Fair Housing Audits." *American Economic Review* 76 (5): 881–893.