The University of Chicago Law School Roundtable

Volume 6 | Issue 1 Article 10

1-1-1999

Measuring Discrimination in the Workplace: Strategies for Lawyers and Policymakers

Julie Lee

Caitlin Liu

Measuring Discrimination in the Workplace: Strategies for Lawyers and Policymakers

JULIE LEE[†]
CAITLIN LIU[‡]

I. INTRODUCTION

In light of dramatic changes in the legal and political landscape in recent years on affirmative action in employment, it has become more important than ever to gather facts about the nature and extent of discrimination in the workplace. Studies that aim to investigate, measure and analyze discrimination not only can provide valuable information to policymakers and the public about the state of society but can also be effective instruments for litigation and law enforcement. The goal of this article is to explain the different techniques through which employment discrimination can be measured, illustrate how they can be used, and identify the state-of-the-art methodologies.

We will begin by explaining why it is important to measure discrimination and then present a brief overview of the five main methods for investigating discrimination. This article will then provide an in-depth examination into the most commonly used techniques, explain their underlying methodological principles, and highlight some examples. We will also explore the latest innovations, analyze the strengths and weaknesses of each methodology, and discuss the implications such studies can have on developing better strategies for litigation, enforcement, research, and policymaking in the future.

---

†. B.A. 1994, University of California at Berkeley; M.P.P. 1997 John F. Kennedy School of Government, Harvard University.

‡. B.A. 1991, Stanford University; M.P.P. 1997, John F. Kennedy School of Government, Harvard University; J.D. 1998, University of California at Berkeley School of Law (Boalt Hall).

A. WHY MEASURE DISCRIMINATION?

Like taking the temperature of a feverish patient, measuring employment discrimination provides a diagnosis of the severity of an illness—the extent of discrimination or lack of equal opportunities in different aspects of employment such as hiring, promotion, termination, and wages. There are several reasons why these diagnostic techniques are vitally needed today. Measurement methods can be a potent instrument for law enforcement, whether the techniques are used to investigate complaints, monitor compliance with anti-discrimination laws, or gather evidence for litigation.[1] Research is also necessary to gather data and educate the public about the persistence of discrimination in the workplace; better information about the level of discrimination against your neighbors, or possibly against yourself, will likely lead to more prudent voting and advocacy. Measuring discrimination can further provide policymakers with valuable feedback on the amount of progress made in anti-discrimination efforts, and the data obtained can help identify problem areas for better targeting law enforcement efforts and policymaking in the future.

B. OVERVIEW OF DIFFERENT TECHNIQUES

There are five main ways through which discrimination in the workplace can be measured or gauged: disparity studies, multivariate analyses, matched-team testing, victimization studies and self-reports. The following is a brief overview of each.[2]

"Disparity studies" compare characteristics of the "affected" group to those of the "benchmark," or a peer group of comparable individuals, and use probability theory to test for the likelihood that differences in the outcome in question would have occurred by chance. This technique can be used to analyze allegations of discrimination in hiring, promotion, termination, initial placement and wages. Disparity studies can be used for research purposes but are primarily used to investigate allegations of discrimination.

---

1. To prove discrimination in a court of law, the plaintiff(s) must show *disparate treatment* by the employer, or *disparate impact* from a specific employer policy or practice. To prove disparate treatment, the plaintiff must prove an intent to discriminate against people because of their race, color, religion, or other protected characteristic, and in some cases intent may be inferred from the mere fact of differences in treatment. *Hazen Paper v Biggins*, 507 US 604, 609 (1993). Disparate impact cases involve employment rules or practices that are facially neutral but fall more harshly on one group than another, and courts outlaw those rules or practices that cannot be justified by business necessity. No showing of intent is necessary to prove disparate impact. Id. A prima facie case of disparate impact may be established through statistical evidence. Once the plaintiff establishes a prima facie case, the burden shifts to the defendant to show that its actions were based upon legitimate business needs. See Michael J. Piette and Douglas G. Sauer, *Legal and Statistical Approaches to Analyzing Allegations of Employment Discrimination*, 3 J Legal Econ 1, 4-5 (1993).

2. For a quick summary of the uses, strengths and weaknesses of each technique, see Appendices B and C.

"Multivariate analyses" refer to a grouping of statistical techniques using multiple regressions that attempt to separate out the qualification (e.g., educational attainment) and demographic (e.g., race, gender or age) factors that influence employment outcomes. Multivariate analyses can be used to examine discriminatory outcomes in hiring, promotion, termination, initial placement, and wages. This methodology has been used for both research and investigative purposes.

"Matched-team testing," also called *"testing,"* relies on the use of two or more testers, identically matched in all qualifications and personal characteristics except for the characteristic in question (e.g., race, gender or age), to investigate whether they have been accorded different, and perhaps discriminatory treatment. Testing for employment discrimination can be conducted for research or law enforcement purposes, but this technique is usually limited to measuring discrimination at the hiring level.

"Victimization studies" refer to surveys, polls or interviews that ask respondents whether they have experienced discrimination in the workplace. The results are generally used only for research, public education and policy feedback purposes and not for enforcing laws.

"Self-reports" are those studies that survey or interview employers to find out whether they harbor discriminatory attitudes, beliefs and perceptions, and whether they have engaged in discriminatory behavior. Like victimization studies, self-reports produce information that can provide the public and policymakers with better insights into the problem but are not used for law enforcement purposes.

As research and investigation techniques, disparity studies, multivariate analyses and matched-team testing are generally believed to be superior methods for generating quantifiable and robust data for measuring discrimination. The bulk of this paper will focus on in-depth discussions of these three techniques. Victimization studies and self-reports can also provide valuable information on discrimination in the workplace, but inherent methodological problems in these techniques lead to unreliable measurement results. Because victimization studies and self-reports are less useful for measuring the true extent of discrimination, this paper will provide only abbreviated summaries of those methodologies at the end.

II. DISPARITY STUDIES

A. OVERVIEW OF METHODOLOGY

Disparity studies have proliferated in recent years as a technique to detect and measure discrimination in the workplace. The method has been used by the federal government to investigate discrimination against government contractors and within contracting organizations. A proliferation of such studies were conducted after the 1989 decision by the Supreme Court in *Richmond v J. A. Croson*

Co.,[3] when state and local jurisdictions needed to examine whether their minority business enterprise programs comported with the new "strict scrutiny" judicial standards.[4] By one estimate, more than 100 disparity studies have been done nationwide, at a cost of over $45 million.[5]

1. General principles

All disparity studies include two main components: an appropriate "benchmark" group and an application of probability theory. The benchmark serves as a comparison group against the pool of individuals in question. Probability theory tests for the likelihood that a distribution as extreme as the one observed in the "affected" group would have occurred by chance.

2. Constructing benchmarks

To provide an accurate comparison, the benchmark group should closely resemble, in all relevant characteristics such as qualifications and career interests, the group against which the comparison will be made. Comparability between the "affected" and benchmark groups is crucial for disparity study results to be considered valid.

Depending on the type of study being conducted, internal or external benchmarks, or both, may be preferred. If promotions or terminations within an organization are being examined, an internal benchmark would likely be used: the "affected" pool of employees who seem to have differential rates of promotion or termination might be compared against benchmark pool of employees who are not "affected." External benchmarks are often used in cases involving allegations of discrimination in hiring; for example, the demographic characteristics of the individuals hired into a firm may be compared to the demographic character-

---

3. 488 US 469 (1989). The case involved a constitutional challenge to a city's minority contracting program, instituted to remedy past discrimination. The plan required that the city's prime contractors spend at least 30 percent of the dollar amount of every city contract on minority subcontractors.

4. In *Croson*, the Court stated that government-sponsored race-based measures would be subject to "strict scrutiny." For a plan such as Richmond's to pass muster under strict scrutiny, the city must demonstrate a "compelling governmental interest" in justifying the plan. Holding that a "generalized assertion that there has been past discrimination in the entire construction industry" is insufficient justification, the Court indicated that the city should have determined "the precise scope of the injury it seeks to remedy." Id at 470. This suggests that the Court wants not just *proof* but also to know the *extent* of past discrimination – which could be ascertained only through data gathering, measuring and documentation. The Court also stated that the remedy had to be "narrowly tailored" to target only the effects of past discrimination. Id at 469-71.

5. Oversight Hearing on the Impact of *Adarand v Peña*: The Constitutionality of Race-Based Preferences, Joint Hearing before the Subcommittee on the Constitution, Federalism, and Property Rights of the Senate Committee on the Judiciary and the Subcommittee on the Constitution of the House Committee on the Judiciary, 104th Cong, 1st Sess 41 (Sept 22, 1995) (statement of George R. LaNoue, Professor of Political Science and Director of the Policy Sciences Graduate Program at the University of Maryland-Baltimore).

istics of the local work force. Examining pay differentials may require the use of either or both internal and external benchmarks. The earnings of the "affected" group may be compared to the earnings of a benchmark group in the firm (if, for instance, the benchmark and "the affected" group in the firm are employed at the same level) and a benchmark group outside the organization.

3. Employing probability theory

a. Chi-square statistic

Once the benchmark has been constructed, a common starting point for establishing a prima facie case of employment discrimination is the chi-square test, which is a simple comparison of two or more distributions.[6] An example of how this technique can be used is an investigation of hiring discrimination by a company against African-Americans. The condition of the affected group (the percentage of African-Americans in the pool of total hired) is compared against the condition of the benchmark (percentage of African-Americans in the applicant pool). If the chi-square statistic indicates that there is a disparity and that the difference found probably did not occur by chance, then further investigations would be warranted.[7]

b. Binomial test

Continuing the process of establishing a prima facie case, a binomial test may be used. This technique provides additional information on the extent of disparities. Continuing the example of the allegation of discrimination against African-Americans, a binomial analysis would yield data on the predicted number of African-Americans hired by the firm if there has been no discrimination, the differences between the predicted results and actual numbers, and the statistical likelihood that this shortfall occurred by chance. If the gap between the actual number and the expected number is statistically significant, an inference of discrimination may be drawn, and the burden of evidence would then shift to the defendant firm to refute the prima facie case.[8]

When establishing whether a disparity is statistically significant, a 95 percent confidence interval is generally used by courts, as it is the convention of social science. [9] In other words, discrimination may be inferred only if the disparity could have happened by chance in less than or equal to five percent of the time. A high significance level implies that the difference found in the sample is not

---

6. Piette and Sauer, 3 J Legal Econ at 6 (cited in note 1).

7. Id.

8. Id at 7.

9. See *Dalley v Michigan Blue Cross/Blue Shield, Inc.*, 612 F Supp 1444, 1451 n 18 (E D Mich 1985). But see *Watson v Fort Worth Bank & Trust*, 487 US 977, 995 n 3 (1988) ("We have emphasized the useful role that statistical methods can have in Title VII cases, but we have not suggested that any particular number of 'standard deviations' can determine whether a plaintiff has made out a prima facie case in the complex area of employment discrimination.")

likely to have occurred at random, given that there is no such difference in the benchmark population. It also means that there is a 1 in 20 chance that the discrimination observed happened by chance.[10]

B. COMPONENTS OF THE STATE-OF-THE-ART METHODOLOGY

Because disparity studies are primarily used for enforcement purposes, it is important that the technique is used in a way that complies with court expectations. An important example could be found in the *Croson* case mentioned earlier. To justify a program that set aside 30 percent of the municipal contracts for minority contractors, the City of Richmond cited a disparity study that had found that although 50 percent of the city's population was black, only 0.67 percent of its prime contracts were awarded to minority businesses. The city argued that this statistical disparity – between the population at large and the amount of contract dollars – was evidence of discrimination.[11] The Supreme Court, however, could not have disagreed more.

In the Court's eyes, the study was fatally flawed because it did not use the appropriate benchmark. Justice O'Connor explained the following criteria: "Where there is a significant statistical disparity between the number of qualified minority contractors willing and able to perform a particular service and the number of such contractors actually engaged by the locality or the locality's prime contractors, an inference of discriminatory exclusion could arise."[12] (emphasis added) The city was comparing the number of contracts awarded with the black population at large, but the population that should have been used was qualified minority contractors who were "willing and able" to do those jobs. The city also failed to give specific evidence of discrimination against blacks or other minorities in the contracting business.

Though *Croson* focused on public contracting, the language and reasoning of the Court points to the criteria that all disparity studies must meet to pass judicial muster: constructing a benchmark group that is not only qualified but also "willing and able," using statistically sound methodologies, and having additional evidentiary information to bolster statistical data and identify the sources of discrimination.[13]

1. Good benchmarking

A good disparity test is one that relies on an appropriate benchmark. As the *Croson* decision indicated, it is not sufficient to compare the affected group to the population at large; the entities of a benchmark and the "affected" pools should be similarly "qualified" as well as "willing and able."[14] For example, when inves-

10. See *Segar v Smith*, 738 F 2d 1249, 1282 (D C Cir 1984).
11. *Croson*, 488 US at 499 (cited in note 3).
12. Id at 509.
13. Id at 509-10.
14. Id at 509.

tigating allegations of discrimination in hiring, one may construct a sophisticated benchmark using an organization's "applicant flow data," which are the profiles of people who have applied for a job with the defendant organization and their success rates. Qualifications may be ascertained from education and years of experience, and willingness to work may be presumed from their voluntary entry into the application process. As the employee stock of both benchmark and "affected" firms can vary over time, it is important to gather data on the benchmark firm specific to the time period at which an alleged discriminator conducted the challenged behavior. It is also important to ensure that the benchmark group is not over-inclusive.

Constructing a good benchmark group can sometimes be difficult, as the data may be ambiguous or not available. Indeed, the usual attack on the validity of a disparity test is showing that the benchmark relied upon is inappropriate.[15] For instance, the defendant may assert the importance of unquantifiable factors such as specialized work experience or productivity-related differences in the "affected" group to explain differences in hiring, pay, or promotion.[16]

2. Sound study design

The means to determine disparities must be statistically sound. If a sample group is to be used for benchmarking purposes, the sample must be sufficiently large and drawn at random. To analyze for significance, a one-tail, rather than a two-tail test should be used, because the differential outcome under examination moves in only one direction, making its results less vulnerable to attack.[17] If disparities are found, a high level of confidence—generally, at the 95 percent level, or even at the 99 percent level if the disparity observed is greater than three standard deviations—should be used for courts to find "significant" statistical disparity.

3. Additional data

As statistical disparities between the benchmark and "affected" groups can provide only an inference of discrimination, the data should be supplemented with additional information to identify the source of discrimination to provide a basis for a "narrowly-tailored" remedy as required by the courts.[18] Moreover, where the court believes the statistical evidence is weak, anecdotal information provided by witnesses or victims of discrimination becomes necessary to establish a persuasive case.

---

15. Piette and Sauer, 3 J Legal Econ at 7 (cited in note 1).

16. Id at 8-9.

17. A two-tail test would be appropriate only if the disparity could go either way; for example, when examining disparities in public contracts awarded to minorities and non-minorities, a two-tail test should be used only if there is uncertainty over who gets more, i.e. that minorities could be given more contracts than non-minorities. Id at 13.

18. Oversight Hearing on the Impact of *Adarand v Peña* (statement of George R. LaNoue) at 43 (cited in note 5).

C. EXAMPLE OF A DISPARITY STUDY: *EEOC V SEARS, ROEBUCK AND CO.*

Constructing a good disparity study is not easy; much can go wrong. Perhaps the classic example of a disparity study that failed is the 1988 *EEOC v Sears, Roebuck & Co.* case,[19] from which many lessons may be gleaned on what should be done, and just as important, what to avoid in disparity testing.

The Equal Employment Opportunity Commission suspected Sears, Roebuck & Co., a nationwide department store chain, of discriminating against women; the EEOC found that most of the company's female sales employees worked in low-paying, non-commission jobs while male employees worked in better paying, commission jobs. To gather evidence of discriminatory conduct by Sears, the EEOC compiled voluminous data on hiring and constructed a benchmark to use as its comparison group. The EEOC then examined whether there were differences between the proportion of women hired for commission sales positions and the proportion of women in the applicant pool of all sales positions.[20]

To construct the benchmark group, the EEOC used applicant flow data. It first culled employment applications of 33,000 rejected sales applicants from 33 randomly selected Sears stores and the applications of approximately 1,920 persons hired into full-time and part-time commission sales positions at approximately 210 Sears stores, between the years 1973 and 1980. Next, using payroll records, the EEOC estimated the female proportion of full-time and part-time commission sales hires in all Sears stores in the United States for those same years. The job applications did not distinguish between commission and non-commission sales jobs. Therefore, to estimate the proportion of full-time and part-time commission sales applicants who were women, the EEOC analyzed the sample of applications, and counted as commission sales applicants all applicants who had applied for any job at Sears, except those persons who specifically requested non-sales jobs. The EEOC then compared the estimated percentage of women hired into commission sales jobs ("actual percent female") with the percent of women in the "sales" applicant pool ("expected percent female"), on a nationwide and territorial basis from 1973 through 1980.[21]

Ambitious as this study was, the plaintiff's case contained two fatal flaws. First, the benchmark was over-inclusive and thus inappropriate. The job applications did not distinguish between commission-selling and non-commission-selling positions or account for differences in interests or qualifications among applicants. Rather, all the applications were pooled together to form the comparison group. The court found the benchmark over-inclusive and unreliable and thus affirmed the district court's finding that the EEOC "presented no credible evidence" to support its assumption that all applicants who indicated interest in sales were also interested in commission sales.[22]

---

19. 839 F2d 302 (7th Cir 1988).
20. Id.
21. Id at 302-09.
22. Id at 326.

The first mistake by the EEOC was compounded by a second—EEOC did not call on former job applicants or current or past Sears employees, to serve as witnesses testifying about Sears' alleged discriminatory conduct. The flawed statistical evidence, coupled with a complete absence of victim testimony, led the court to conclude that the EEOC's allegation of discrimination was groundless.[23]

The morals of this story are these: It is difficult but critical to construct good benchmark groups for disparity studies. It is also important to obtain witness or victim testimony to bolster the court's confidence in statistics, especially when the statistical evidence is believed to be weak.

D. THE VERDICT ON DISPARITY STUDIES

1. Strengths

As a technique for investigating and measuring discrimination, disparity studies have many strengths.

It can be a powerful tool for examining employment discrimination at any level. Unlike the matched-team testing methodology, which, as will be explained later, is only useful for investigating discrimination at the job entry level, disparity studies can be used to examine discrimination throughout the job cycle—at hire, promotion, and termination as well as wage outcomes.

If done correctly and the disparities found are large, results from disparity studies are legally persuasive. According to the Supreme Court, "where gross statistical disparities can be shown, they alone may in a proper case constitute prima facie proof of a pattern or practice of discrimination." [24]

Conducting disparity studies can be relatively inexpensive. If the relevant data already exist, the cost of conducting disparity studies would be the expense of analyzing the data. Costs would be higher, of course, if data are not readily available for analysis and extensive information-gathering is necessary.

Disparity studies are not subject to experimenter bias. The analysis relies on data that illuminate the conditions and trends in the affected and benchmark groups and would not be subject to such attacks to discredit the findings.

There are no ethical concerns. Unlike testing, ethically questionable techniques of examining information are not inherent in the regression methodology.

2. Weaknesses

But this methodology also has many weaknesses.

The relevant data often does not exist, or the data are inadequate. Without good data, good benchmarks cannot be constructed. Incomplete data, for example, can lead to over-inclusiveness in benchmark groups.

There can be variability in availability of certain types of individuals in "qualified applicant pools" of suspect firms. For example, the demographic composition of applicant

23. Id.

24. *Hazelwood School District v United States*, 433 US 299, 307 (1977).

pools may change over time, such as when there are structural changes in the economy or ups and downs due to business cycles.[25]

The statistical standard for establishing a prima facie case seems arbitrary and may be overstringent. The existence of a statistical cutoff at five percent for many courts means that some, albeit a small percentage, of innocent firms will be falsely accused of discrimination and would be subjected to further litigation. Shrinking the cut off to below five percent would decrease false accusations, but increase false absolutions: a large percentage of firms that are in fact guilty will be "freed" because the disparities would not be recognized as significant.

Disparity studies may lead courts to err by excluding chance as a possible cause of the disparity. Courts assume that statistical analysis can reveal the probability that observed work-force disparities were produced by chance. This assumption is based on the error that the probability of an observed disparity given random selection (which is what courts are looking for) is the same as the probability of random selection given an observed disparity (which is the data created from disparity studies).[26] As a result, this error may lead courts to exclude chance as a cause when such doubts may in fact be warranted.[27] The corollary to this error is that courts sometimes assume that the work force of a nondiscriminating employer would mirror the demographic composition of the relevant labor force. Although not a deficiency of disparity studies per se, the way in which the results are presented and explained may lead to such assumptions.[28]

III. MULTIVARIATE ANALYSES

A. OVERVIEW OF METHODOLOGY

Multivariate analyses, which include the simple OLS regression model as well as the logit and probit, are commonly used techniques for measuring disparate impact. Much of the multivariate analysis literature on discrimination focuses on measuring wage differentials, but this methodology can also measure discrimination in other areas of employment such as hiring, promotion, termination and job level.[29]

When studying employment discrimination using multivariate analyses, economists distinguish between two types of discriminatory outcomes: those within the labor market and those taking place beyond its boundaries. Labor market discrimination exists when transactions are conducted in such a way that

---

25. Kingsley R. Browne, *Statistical Proof of Discrimination: Beyond "Damned Lies,"* 68 Wash L Rev 477 (1993).

26. Under Bayesian principles, the previous statement in most cases will not be true, unless the probability of an observed disparity occurs at the same frequency as the probability of an event happening with random selection.

27. Browne, 68 Wash L Rev at 484 (cited in note 25).

28. Id.

29. Robert S. Follett, Michael P. Ward and Finis Welch, *Problems in Assessing Employment Discrimination*, 83 Am Econ Rev 73, 73-78 (1993).

employers minimize or eliminate contact with people of a protected category. Non-labor market discrimination refers to the differential treatment and experiences of individuals of a protected class before they enter the labor market that cause them to be less prepared for employment and therefore less marketable.[30]

The distinction between labor market discrimination and non-labor market discrimination is important because of the different remedies they imply. Attempting to separate out the "labor market" effects from the "non-labor market" effects, researchers control for "productivity" gaps (the portion of the earnings differential attributable to members of a protected class having smaller human capital endowments than individuals in nonprotected classes) and "wage" gaps (the remaining earnings difference where equally qualified and productive members of a protected class are paid less than individuals in unprotected classes) of the relevant workforce.[31]

B. COMPONENTS OF MULTIVARIATE ANALYSES

Regression analyses of employment discrimination control for human capital characteristics and demographic characteristics in an attempt to determine all the relevant factors influencing the dependent variable: employment measurement. A typical, simple regression equation resembles the following:[32]

$$P_i = \Sigma b_j X_{ji} + \Sigma d_k D_{ki} + u_i$$

where

P_i = some employment outcome for individual i

X_{ji} = human capital characteristics, e.g. education, work experience

D_{ki} = demographic indicator variables for sex, race, age, etc.

u_i = error term

In the context of an accurately specified (i.e., "good") regression, the level of discrimination may be defined as the magnitude of the coefficient, if statistically significant, representing the protected characteristic (e.g., race or sex).

---

30. Orley Ashenfelter and Ronald Oaxaca, *The Economics of Discrimination: Economists Enter the Courtroom*, 77 Am Econ Rev 321, 321-25 (1987).

31. Thomas F. D'Amico, *The Economics of Discrimination Thirty Years Later: The Conceit of Labor Market Discrimination*, 77 Am Econ Rev 310, 310-15 (1987).

32. Mark R. Killingsworth, *Analyzing Employment Discrimination: From the Seminar Room to the Courtroom*, 83 Am Econ Rev 67 (1993).

C. Different Types of Multivariate Analyses

1. OLS

Every regression method attempts to quantify the impacts of a group of relevant independent variables on the dependent variable. The simplest type of a regression model in the employment discrimination context are linear regressions, also known as "Ordinary Least Squares." In the employment litigation context, this type of regression model is appropriately used to examine continuous data, such as salary differences or the impact of job placement tests on specific groups of applicants. [33]

2. Binary dependent variable (logit/probit)

A logit (logistic) or probit regression model is required to correctly model any binary dependent variable, such as hired/not hired, promoted/not promoted or fired/not fired. [34] Logit and probit models each make assumptions about the underlying functional form of the model that are different from the OLS model (and from each other). Each involves a transformation of the dependent variable, and consequently neither type of regression yields readily-interpretable parameter estimates. Fortunately, with some simple transformations, the regression coefficients can be made understandable in a court of law.

3. Instrumental variables (IV)

A more sophisticated version of the OLS includes a simultaneous analysis of a series of linked regression equations known as instrumental variables (IV) or "latent variables" or "simultaneous equations." IV attempts to confront the measurement error and omitted variables problems that commonly occur in the simple OLS and the logit/probit specifications by "including latent variables to represent the 'error-free' constructs that are measured by the fallible observed variables." [35] This model type will permit both direct and indirect relationships to be studied.

D. Examples of Multivariate Analyses: Illustrations of the State of the Art

Though multivariate analyses are used both for research and investigative/enforcement purposes, the most recent innovations have been in the realm of research. The following examples illustrate the use of instrumental variables, multi-tiered tests, logits and probits for the purposes of separating labor market

---

33. Piette and Sauer, 3 J Legal Econ at 9-10 (cited in note 1).

34. Id at 10.

35. Roger E. Millsap and Ross Taylor, *Latent Variable Models in the Investigation of Salary Discrimination: Theory and Practice*, 22 J Mgmt 653, 656-661 (1996).

discrimination from non-labor market discrimination, examining productivity-based explanations for differential placements, and re-evaluating explanations for occupational segregation and promotions.

1. Instrumental variables: adjusting for the uncertainty of specific employment practices that led to discriminatory outcomes

A study by Malkiel and Malkiel investigating salary differentials illustrate the limitation of the simple regression model and the usefulness of IV in investigating employment discrimination. [36] While a simple linear regression attempts to control for human capital and demographic characteristics, the IV technique incorporates both labor market and non-labor market effects into the coefficient on gender.

In this case, Malkiel and Malkiel recognized that simple linear regressions would not be enough to measure whether wage or salary differentials along gender lines were a result of company practices and procedures. To separate out these factors, the Malkiels adopted an IV approach that added an equation that attempted to consider what factors produced the significant coefficient on gender.[37]

With the simple linear regression model, one would use the equation[38]:

$$CS_i = \Sigma b_{1j} X_{ji} + \Sigma d_{1k} D_{ki} + g_1 JL_i + u_{1i},$$

Where

CS = current salary

JL = job level

X = individual characteristics (i.e. education level, etc.)

D = demographic characteristics (i.e. race, sex, etc.)

u = error term

The instrumental variables approach would lead one to use the following set of equations [39]:

$$CS_i = \Sigma b_{1j} X_{ji} + \Sigma d_{1k} D_{ki} + g_1 JL_i + u_{1i}$$

(primary equation to determine differences in current salary, as above.)

---

36. Killingsworth, 83 Am Econ Rev at 68 (citing Burton G. Malkiel and Judith Malkiel, *Male-Female Pay Differentials in Professional Employment*, 63 Am Econ Rev 693, 693-705 (1973)) (cited in note 32).

37. Id.

38. Id.

39. Id.

$$JL_i = \Sigma b_{2j} X_{ji} + \Sigma d_{2k} D_{ki} + u_{2i}$$

(secondary equation to determine the factors which influence job level.)

The Malkiels found that the coefficient on women was close to zero in the primary equation, implying that women did not receive unequal pay for equal work. However, the large and negative coefficient on women in the second equation indicated that women were much less likely to be in higher job levels than men, possibly a result of discrimination in hiring at the higher levels or in promotion. In summary, the Malkiels' results imply that using the simple OLS on the current salary levels alone could yield biased coefficient estimates for both job level and gender, and deeper explorations into the nature of discrimination are needed. [40]

2. Multiple-tiered test using probit and tobits: adjusting for productivity differences

One of the keys for distinguishing labor-market from non-labor-market discrimination is determining the productivity of individual workers. Kolpin and Singell, in a study that focuses on factors that affect a university department's decision to hire recent male and female Ph.D. recipients, examine productivity-based explanations for the differential placements of women in economics departments.[41] This study contains three related empirical analyses. The first analysis examines employment data from various economics departments to determine whether "good" economics departments are less likely to hire women. The second looks at whether there is a relationship between a department's publication performance and its hiring of female faculty prior to the department's evaluation. The third empirical analysis compares the frequency of publication output—"productivity"—of males and females at comparable institutions. Together, the empirical analyses provide powerful insights into the market processes that affect whether a woman enters and/or progresses in occupations that have been traditionally male-dominated.

---

40. Id.

41. Van W. Kolpin and Larry D. Singell, Jr.. *The Gender Composition and Scholarly Performance of Economics Departments: A Test for Employment Discrimination*, 49 Indust Labor Rel Rev 408, 408-12 (1996). The independent variables used by this study include number of assistant professors, number of female senior professors, binary variable for the three types of institutions (public universities, Ph.D. granting departments, and "pure" economics programs), number of applicants for graduate study, a time trend variable, and publications rankings (which rank departments on the basis of the number of pages published per faculty member in 24 leading journals for the period between 1974 and 1978.) The authors constructed a proxy for departmental quality by classifying schools into five groups. The number of schools increases in successive categories to account for the greater imprecision of the quality measure for 'less scholarly' departments.

After conducting the analyses, the study strongly rejected the productivity-based explanations for the gender differences in placement in economics departments. The results of the first empirical analysis indicate that women were placed in relatively unproductive departments. The second analysis shows that the relative proportion of female assistant professors was a significant predictor of a department's subsequent publications ranking, which suggests that female hires produced more research than males at comparable institutions. The results from the third analysis indicate that female economists produced qualitatively better research than their male counterparts. These findings together supported the authors' hypothesis that there is discrimination against women by economics departments, as women are under-placed despite their greater productivity. [42]

3. Using logits to account for occupational segregation

While many previous studies examine occupational discrimination on the demand for employees, Gill conducted a study that analyzes both the supply and demand sides of the market for jobs.[43] In doing so, Gill not only looks at the probability of being in a queue of applicants for a job, but also analyzes the probability of being selected for a position once in the queue.

The study used data from the 1976 and 1981 waves of the National Longitudinal Surveys of young men.[44] To examine the role of discrimination in determining racial differences in occupational structure, the study used logit techniques. Logits were used to analyze the factors that affect the probability that an individual will choose an occupation and the probability that an individual will be hired for a desired job by an employer.

The study used three different dependent variables. The first is the probability that an individual acquires a job in a certain occupation; the second is the probability that an individual wants to be employed in that occupation. The third is the conditional probability that an individual is able to acquire a job in a certain occupation given that he or she wanted it. Three different logit analyses were conducted on each of these three dependent variables while controlling for human capital, personal characteristics, regional controls, regional unemployment rates, and family background. The study found that whites were more likely to end up in professional, managerial, sales and clerical, and craft occupations than are African-Americans.[45]

This study illustrates a useful technique for investigating discriminatory occupational structures, but one problem may be its assumption that preferences for occupations were stable over time, a characteristic for which it is difficult to determine and control. This is a common problem among multivariate analyses and will be discussed in a later section.

---

42. Id at 421-422.

43. Andrew M. Gill, *The Role of Discrimination in Determining Occupational Structure*, 42 Indust Labor Rel Rev 610, 610-11 (1989).

44. Id at 614.

45. Id at 621-22.

4. Analyzing promotions using probits

Another use for regressions is detecting and measuring employment discrimination in promotions. A study by Paulin and Mellor tests the proposition that the gender and racial composition of an employee's occupation significantly affects the likelihood of promotion. [46] For the study, the authors collected three years of data on employee grade levels and supervisor evaluations from the personnel files of a medium sized financial services firm.

Paulin and Mellor dispute the human capital theory that workers possess a great degree of control over their promotion process and operate their study under the assumption that a promotion depends not only on having the requisite human capital, but also on the probability of an opening for promotion and quality of the competition. Each of these criteria upon which a promotion is based are affected by the rules that a firm has regarding the way jobs are classified and defined in relation to each other, as well as the rules determining hiring, layoff, transfer and promotion of workers. [47]

In addition to using simple correlations of "Spearman coefficients" to determine the relationship between grade level of occupation and concentration of minorities or women in the occupations, Paulin and Mellor use the probit technique to examine whether occupations with high concentrations of females and/or minorities are structured in a way that impedes the advancement of those individuals to higher ranking positions.[48]

To conduct their probit analysis of promotions, Paulin and Mellor use, as a dependent variable, "probability of promotion" and among independent variables, an individual's human capital accumulation, the competition, the race/gender of an individual, characteristics of the occupation within which the individual works, and proxies for "ambition." Though "ambition" may be difficult to quantify, omitting the variable could bias the results of the probit because it affects promotion and is correlated with other covariates.

The study was conducted using the probit in three different ways: one model examines all employees, another model is specific to gender/race subgroups, and the third model distinguishes between exempt and non-exempt[49] categories. The authors found that the gender/race composition of an occupation did indeed affect the likelihood of promotion in several cases. "Except for white males," the study concluded, "an individual's job performance and the degree of competition matter."[50]

---

46. Elizabeth A. Paulin and Jennifer M. Mellor, *Gender, Race and Promotions within a Private-Sector Firm*, 35 Indust Rel 276 (1996).

47. Id at 278.

48. Spearman coefficients are based on the ranking of each occupation according to average grade level, percent female and percent minority. Id at 284.

49. Non-exempts are employees paid on an hourly basis and are compensated for overtime. Exempts are paid an annual salary and are not compensated for overtime. Id at 288.

50. Id at 294.

E. THE VERDICT ON MULTIVARIATE ANALYSES

1. Strengths

As analytical techniques, multivariate studies offer the following advantages:

Conducting these studies can be relatively inexpensive, if one makes use of existing evidence; however, costs could increase if data is not readily available and extensive information gathering is necessary.

Regressions allow for quantification of relationships between a group of variables and a single dependent variable. This is particularly important because there is no other methodology used to measure discrimination that can answer questions regarding the magnitude of the impact of certain characteristics on the employment outcome.

One can separate out the various effects that influence the outcome being measured. Unlike disparity studies, one would not have to worry about how certain characteristics of a benchmark influence the comparison between the benchmark and the "affected" pool because such traits can be "held constant," or controlled for, in the regression analysis.

Regression analyses are not subject to experimenter bias. They rely on data that illuminate the trends in the relevant population.

This methodology is useful for examining employment discrimination at any level. Unlike the matched-team testing methodology, regressions can be used to examine discrimination at hire, promotion, and termination (through the use of logits or probits) as well as discriminatory wage outcomes.

There are no ethical concerns. Unlike testing, ethically questionable techniques of examining information are not inherent in the regression methodology.

2. Weaknesses

This methodology also has weaknesses.

There are problems with interpretations of the wage gap. First, the wage gap misses some of the earnings losses subsumed by the productivity gap. Second, it arbitrarily includes some of the impact of non-labor market discrimination, because of chronic misspecification problems, where group differences in productivity-related characteristics attributable to non-labor market discrimination are sometimes missed by the productivity gap estimates and mistakenly subsumed, instead, by the residual wage gap.[51]

The results of regression studies may be vulnerable to charges of omitted variable bias. If there are important independent variables that are not controlled for, which is often the case when information is not available or incomplete, the effects of the omitted variable are then subsumed in residual. If these omitted variables differentially affect people with different demographic characteristics, then the error

51. Francine D. Blau and Marianne A. Ferber, *Discrimination: Empirical Evidence from the United States*, 77 Am Econ Rev 316, 316-20 (1987).

term will be correlated with demographic variables and distort the values of the coefficient on demographic indicators.[52]

Included variables may themselves be affected by discrimination, which means that differences among people in different demographic groups who have the same individual characteristics measure only the "incremental" discrimination which doesn't account for the discrimination induced differences in these characteristics. In addition, if included variables are affected by discrimination, then the coefficients on the individual characteristics are endogenous, rendering the estimates biased.[53]

Regressions measure correlation, not causation. With regressions, it is difficult to determine what specific employment practices led to discriminatory outcomes, because these models do not answer questions about intent or provide causal explanations for patterns found in data. A significant and negative coefficient on sex as an indicator variable (if the dependent variable is wage) doesn't explain why women have lower wages. Was it unequal pay for equal work? Unequal access to better paid work? Different starting salaries? Different rates of salary increases? Different rates of promotions for men than women? Regressions do not provide any answers.

Sampling error may lead to an indication of discriminatory outcomes. A regression result that indicates discriminatory outcomes may be a result of sampling error, even if all factors that influence pay are accounted for in the individual characteristics X and the unobservables are uncorrelated with the X's and distributed identically.[54]

IV. TESTING

While empirical studies attempt to measure disparate impact, matched-team testing aims to investigate unequal treatment. Sometimes also called "auditing," "audit-testing," or "matched-pair testing," this method, in the employment context, relies on the use of perfectly "matched" teams whereby two or more testers—identical in age, education, appearance and qualifications, but differ only in race, gender or whatever form of discrimination is being investigated—are sent out to apply for the same jobs. Because this experimental research design attempts to control for all visible and personal characteristics of the job applicants, testers who encounter different treatment by employers are presumed to have encountered discrimination. Not only have the results from testing studies helped educate the public and policymakers about the extent of discrimination in the workplace, testers have also produced critical evidence for anti-discrimination law enforcement. [55]

---

52. Killingsworth, 83 Am Econ Rev at 67 (cited in note 32).

53. Id.

54. Ashenfelter and Oaxaca, 77 Am Econ Rev 321, 323 (cited in note 30).

55. For many years, the Federal Government, through the Justice Department and the Department of Housing and Urban Development, has used testing to investigate discrimination in housing and enforce fair housing laws. Last year, the Government began looking into ways to

A. OVERVIEW OF METHODOLOGY

Testers were first successfully used in the 1950's by civil rights activists to investigate discrimination by public transit authorities.[56] This technique was more widely adopted in the 1960s to enforce fair-housing laws, and it proved to be extremely effective in generating evidence of housing discrimination.[57] In the 1970's, this methodology was extended to investigate discriminatory conduct by lending institutions. By the late-1980s and early 1990s, researchers began using matched-team testing to investigate employment discrimination.

Testing, which can be conducted in person, over the telephone, or through correspondence (where "equally matched" resumes are mailed to prospective employers), can detect and measure discriminatory treatment by employers in numerous ways, usually including the following: whether a tester-applicant is given an opportunity to interview; whether he or she is offered a job or given a job referral; and whether there are differences in the offered starting wage or other non-cash compensation.

This methodology can also measure more subtle, but nonetheless possibly discriminatory differences in treatment. Tester-applicants can document whether they are offered an employment application and how long they are made to wait before an employer responds to a request for an interview. If a tester is called back for interview, he or she can record the length of interview and the quality of interactions with the employer (for example, whether encouraging or discouraging comments were made). Testing can uncover whether there was "steering" by the employer; i.e. a tester may be offered a job, but a dead-end or less desirable job than another tester. Testers can also report whether they have been given preferential treatment or employment opportunities, such as if the applicant is automatically considered for an unadvertised and better job with the same employer.

There are two general types of testing for employment discrimination. Research testing can provide valuable documentation of the existence and extent of discrimination as well as track changes in the nature and levels of discrimination over time. The results can be used not only to inform the public about discrimination in their community but also to alert policymakers about problem areas to

---

use this methodology to investigate employment discrimination, and the Department of Labor, through the Office of Federal Contractor Compliance Programs, launched a pilot testing program in the Northeast. For the last several years, the Fair Employment Council of Greater Washington, a non-government organization, has also been using the testing technique to investigate the extent of discrimination in hiring and to sue offending employers.

56. See *Evers v Dwyer*, 358 US 202 (1958) (holding that tester has the right to challenge segregated seating on a bus).

57. James J. Heckman and Peter Siegelman, *The Urban Institute Audit Studies: Their Methods and Findings*, in Michael Fix and Raymond J. Struyk, eds, *Clear and Convincing Evidence: Measurement of Discrimination in America* 187, 190 (The Urban Institute Press, 1993).

target for future anti-discrimination efforts.[58] Research testing may also be used by the government to evaluate the effectiveness of programs or efforts intended to combat employment discrimination. A typical research audit for investigating discrimination in the labor market might include hundreds of tests using matched pairs of testers sent out to employers that were randomly selected from a pool of job openings advertised in a newspaper.

Enforcement testing can be used to investigate specific employers' hiring practices, to monitor compliance with injunctive remedies, and to gather evidence for litigation.[59] Employers, public or private, can also self-test for discriminatory practices to ensure compliance with laws and protect themselves against future litigation. An enforcement test might be prompted by an allegation of discriminatory conduct on the part of an employer. The complaint is then followed up by testers visiting the employer. In order to rule out random events and establish a pattern of unequal treatment, repeated audits with different teams of testers are sent to the same organization. Testing results are then collected and documented to serve as evidence in trial. [60]

B. COMPONENTS OF THE STATE-OF-THE-ART METHODOLOGY

All good testing, whether for research or enforcement purposes, share similar principles and techniques: exact matching and extensive tester training. Recently, researchers have also begun adopting the technique of "sandwich testing."

1. General principles

a. Match the testers as much as possible

Good testing requires that testers are matched in all observable personal characteristics such as gender, age, height, weight, education, experience, grooming, demeanor and eloquence.[61] More recently, researchers have begun trying to con-

---

58. Michael Fix, George C. Galster and Raymond J. Struyk. *An Overview of Auditing for Discrimination*, in Michael Fix and Raymond J. Struyk, eds, *Clear and Convincing Evidence: Measurement of Discrimination in America* 11-12 (The Urban Institute Press, 1993).

59. Fair Employment Council of Greater Washington, Inc. *Employment Testing*, 2-3. See also Fix, Galster and Struyk, *Overview* at 1-2 (cited in note 59).

60. Title VII of the Civil Rights Act gives individuals the right to sue employers, the Equal Employment Opportunity Commission the authority to punish public agencies or private employers, and the Justice Department the responsibility to bring suits against state and local governments charged with employment discrimination. Enforcement testing seeks to provide the necessary evidence for such proceedings. Ronald Mincey, *The Urban Institute Audit Studies: Their Research and Policy Context*, in Michael Fix and Raymond J. Struyk, eds, *Clear and Convincing Evidence: Measurement of Discrimination in America* 179 (The Urban Institute Press, 1993).

61. Interview with Kennington Wall, Senior Project Coordinator of Fair Employment Council of Greater Washington (May 13, 1997). Interview with Michael Fix, Director of Program on Immigration, The Urban Institute (May 14, 1997). Peter A. Riach and Judith Rich, *An Investigation of Gender Discrimination in Labor Hiring*, 21 Eastern Econ J 343, 346-47 (1995).

trol for "unobservable" traits and behavior by screening testers with psychology and behavior tests so that teams can be matched in temperament and personality.[62] To ensure the best matches possible, testers are selected only after extensive comparison with others and a review by a panel of researchers.[63] One way that tester-selection may be improved in the future, to further ensure identical matches and minimize experimenter bias, may be to have an independent and blind review panel screen the matched pairs for similarities and identify differences.[64]

b. Training

The most sophisticated tests ensure that testers receive extensive, systematic training, whereby testers learn to follow a standardized "script" of words, actions and reactions when conducting the testing. The most recent studies pay testers flat-rate wages so as to not distort incentives to act in ways that differ from their script; for example, testers are not rewarded for finding discrimination nor are they rewarded for getting a job offer. The designers of enforcement tests are especially careful to make sure that testers do not personally benefit from discovering discrimination.[65] To minimize experimenter effects, testers are vigorously monitored, record their experiences independently, and are not permitted to discuss the results of their test experiences with each other.[66]

c. "Sandwich-testing"

Another recent methodological innovation is the use of "sandwich testing," the essence of which is using three or more, rather than just a pair, of testers.[67] When only a pair of testers is used, and the majority applicant received more favorable treatment than the minority applicant, the differential treatment may be vulnerable to criticism that the chance or problematic "matching" of test teams played a role in the outcome.[68] But if there is a team of several "matched" testers—for example, two white testers who consistently received preferential treatment over the identically matched black tester(s), then discrimination provides the more plausible explanation for differences in outcomes. "Sandwich testing" may yield a more conservative documentation of discrimination, but the

---

62. Interview with Michael Fix (cited in note 62).

63. Id.

64. Interview with Professor Thomas Kane, Associate Professor of Public Policy, Harvard University (April 15, 1997).

65. For example, the Fair Employment Council of Greater Washington, Inc., a non-profit organization that conducts employment discrimination testing, ensures that their testers agree in advance that if they become plaintiffs in litigation, any damages awarded must be turned over to the FEC. Roderic V.O. Boggs, Joseph M. Sellers, and Marc Bendick, Jr, *Use of Testing in Civil Rights Enforcement*, in Michael Fix and Raymond J. Struyk, eds, *Clear and Convincing Evidence: Measurement of Discrimination in America* 345, 352 (The Urban Institute Press, 1993).

66. Interview with Kennington Wall (cited in note 62).

67. Interview with Michael Fix (cited in note 62).

68. Heckman and Siegel, *Methods* at 222-23, 224-25 (cited in note 58).

findings are also considered by the research community to be more robust, and very likely in a court of law, the evidence of discrimination more compelling.[69]

2. Specifics for research testing

The key to good research testing is ensuring that the results obtained are valid, i.e., the data obtained present a good picture of what is going on in a particular labor market. Validity can be facilitated through random sampling of employers and ensuring that pairs are indeed well matched. Once the results are obtained, the data can also be analyzed using regressions to identify conditions or factors correlated with higher or lower levels or discrimination.

a. Random sampling

To ensure validity of results for a research test, employers chosen to be tested should be selected at random from a local, regional or national pool of organizations. In some studies, employers are selected from a pool of those who advertise job openings in newspapers. The sample size depends on the level of statistical variance in the population, but experience has shown that a typical sample size for a metropolitan area is about 500 tests.[70] For a nationwide study, the sample size may be over 1,000 tests.[71]

b. Ensuring that all the teams used are well matched

To ensure that the testers are indeed identically matched, some researchers have urged the use of a Fischer test to assess homogeneity across different pairs.[72] When a study relies on results obtained from multiple pairs, researchers not only need to ensure that testers are matched one-on-one in their pairs, but also that the pairs are not too different from each other. If there is heterogeneity across pairs, there could be wide discrepancies in treatment, but the net results of an audit may show little discrimination. Also, researchers would not be able to reject the hypothesis that systematic differences in treatment are due to differences between the testers.[73]

---

69. Interview with Michael Fix (cited in note 62).

70. Id.

71. See Mark Bendick, Jr. *Employment Discrimination Against Older Workers: An Experimental Study of Hiring Practices* (Fair Employment Council of Greater Washington, 1993).

72. The Fischer test, also known as the F-test, can test the equality of two variances and is equivalent to the ratio of the explained variance over the unexplained variance. Other things being equal, we would expect that if there were a strong statistical relationship between the two variances, the F-test statistic would be large. Robert S. Pindyck and Daniel Rubinfeld, *Econometric Models & Economic Forecasts* 64-65 (McGraw-Hill, Inc., 1991).

73. A University of Colorado study of discrimination against black and Hispanic testers in Denver, Colo. was faulted for the heterogeneity of its data. For example, in one of the pairs the black tester was greatly favored over the white, while in a second pair the white tester was greatly favored over his black colleague. The net results of those two audit pairs alone—if the discrimination in favor of the black was subtracted from the discrimination in favor of the

c. Analyzing results with regressions

The data yielded by research testing can also be analyzed using regressions. To identify "trouble spots," researchers can apply econometric models to determine how differences in treatment vary under different conditions. A study by Bendick, Jackson and Reinoso collected results from six studies using testers and analyzed whether various factors, such as gender, location (urban or suburban areas), job advertising medium (whether job was advertised in urban paper or suburban paper), or firm size, had any effect on discrimination.[74]

3. Specifics for enforcement testing

Testing for enforcement purposes does not require random sampling, for employers that are suspected of discrimination can be targeted. Nor do the small number of tests required—only two or three tests are necessary, generally—lend themselves to regression analysis. What enforcement testing requires is thoughtful planning, every step of the way, for possible future litigation. Above all, enforcement tests need two things:

a. Articulate testers with no skeletons in the closet

Sophisticated organizers of enforcement audits should know that the credibility of testers as witnesses is very important. They must hire testers who are articulate enough to present testimony clearly in court. Testers also need to have backgrounds without personal or legal problems that can weaken their credibility as plaintiffs or witnesses. Moreover, testers need to be available for trial if necessary—they must be willing to remain in contact with the testing organization and appear in court in case their testimony is needed.

b. Documenting anecdotes and stories

While it is important for both research and enforcement testing to document quantifiable data (for example, whether the tester was given an employment application, the length of interview, whether a job was offered and starting wage), it is particularly important for enforcement testing to require record-keeping by testers that are essay-like, so anecdotes can be gleaned and the narratives can later be used as testimony in court.[75] Testimony by victims of discrimination can play an important role in how the court decides a case,[76] and having the discriminatory experience already documented on paper can make the case easier to present and the evidence more convincing to a trier of fact.

---

white—would suggest little overall discrimination, when in fact there was strong discrimination going both ways, leaving one to wonder whether the testers were that well "matched" to begin with. Heckman and Siegelman, *Methods* at 220-222 (cited in note 58).

74. Marc Bendick, Jr., Charles W. Jackson and Victor A. Reinoso, *Measuring Employment Discrimination Through Controlled Experiments*, 23 Rev Black Pol Econ 25 (1994).

75. Boggs, Sellers, and Bendick, *Use of Testing* at 353-54 (cited in note 66).

76. See *Sears*, 839 F2d 302 (cited in note 19).

C. EXAMPLES OF TESTING

The following are four examples of how testing has been used to investigate employment discrimination. These studies illustrate the depth and breadth of the state-of-the-art testing techniques and enforcement efforts. Although these studies share the same underlying methodologies, they vary in their goals and measurement techniques.

1. Examples of research testing

a. In-person auditing: Studies by the Urban Institute investigating racial discrimination

The Urban Institute (UI) has been widely credited as the first to design and carry out studies using matched testers, standardized procedures and random samples of employers to investigate hiring discrimination. In 1989 UI conducted a study using Anglo and Hispanic testers in Chicago and San Diego, and in 1990 tests using black and white testers were carried out in Chicago and Washington, D.C.[77]

Commissioned by the United States General Accounting Office, the Anglo/Hispanic study aimed to investigate whether the 1986 Immigration Reform and Control Act, which imposed sanctions against employers hiring illegal immigrants, resulted in increased discrimination against legal immigrant job applicants. The researchers used Hispanic testers who were "foreign"-looking and -sounding—the testers spoke fluent English with noticeable Spanish accents, and six out of eight of the Hispanic testers had moustaches. The study used four pairs of testers in each city and completed 360 tests. The black-white study, sponsored by the Rockefeller Foundation, was conducted with five pairs of testers in each city and completed a total of 476 tests.[78]

For each study, the testers were trained—the Hispanic/Anglo testers for two-and-a-half days and the black/white testers for five days. Testers alternated being first to contact the employer to ensure that no tester had first-contact advantage, and they were closely monitored throughout the testing process.

While the findings of the black/white test have not been the subject of much dispute, the Hispanic/Anglo study has been criticized, mainly for its use of testers who were not perfectly matched. Critics point out, quite correctly, that the tests were insufficient for demonstrating a bias against Hispanics; what the tests did show was perhaps a bias against job applicants with foreign accents and facial hair.[79] The criticisms are not entirely fair, however, considering that the original purpose of the research was to investigate whether there was a bias

---

77. Mincey, *Research* at 171 (cited in note 61).

78. Wendy Zimmerman. *Summary of the Urban Institute's and the University of Colorado's Hiring Audits*, in Michael Fix and Raymond J. Struyk, eds, *Clear and Convincing Evidence: Measurement of Discrimination in America* 407, 407-08 (The Urban Institute Press, 1993).

79. Heckman and Siegelman, *Methods* at 217-218 (cited in note 58).

against employees who seemed like immigrants. Also, given that a significant proportion of the Hispanic population might have facial hair and/or speak English with an accent, the findings of this study may indeed be indicative of the levels of discrimination against Hispanics in the cities studied.

b. Correspondence testing: Investigating age discrimination

The technique of correspondence testing is illustrated in a 1994 study of age discrimination commissioned by the American Association of Retired People. Conducted by the Fair Employment Council of Greater Washington, Inc., a non-profit organization in the District of Columbia, the study sought to investigate whether employers treated job applicants differently because of age. The FEC created a pair of resumes, identical in all qualifications listed, such as education and work experience but differed in applicant age—one resume was that of a 57 year old, and the other was that of a 32 year old. The resumes, along with comparable cover letters, were mailed to a random, nationwide sample of 775 large companies and employment agencies.[80]

Employer responses indicated that there was disparate treatment of the applicants. The younger applicant was favored 43 percent of the time, while the older applicant was favored 16.5 percent of the time. The FEC subtracted the amount of favorable treatment of older workers from the amount of unfavorable treatment, which resulted in a "net" difference of 26.5 percent. The study also found that the level of discrimination differed dramatically depending on geographic region and industry. For example, there was more discrimination against older workers in the South and West (at 25.6 percent and 42.2 percent, respectively) than in the Northeast and Midwest (at 8.3 percent and 8.4 percent, respectively). In the finance/real estate sector, the net favorable treatment of the younger worker was 6.6 percent of the time, while in manufacturing the younger worker was preferred almost 100 percent of the time.[81]

Perhaps even more revealing for public education purposes is the second part of the study, which compared the levels of age discrimination across more successful and less successful firms. Using a composite of business and industry indicators, the FEC compiled an index for company success. The FEC found a strong correlation between non-age-discrimination and company success. The lower-ranked (less successful) firms tended to discriminate more (at 32.2 percent), while the most successful firms hardly discriminated at all (at 2.2 percent). The study concluded that "it's not good business to discriminate" against older workers.[82]

---

80. Bendick, *Employment Discrimination*, (referring to Marc Bendick, Jr., Charles W. Jackson and J. Horacio Romero, *Employment Discrimination Against Older Workers: An Experimental Study of Hiring Practices*, 8 J Aging Soc Pol 1 (1996)) (cited in note 72).

81. Id at 2-3.

82. Id at 3.

2. Examples of enforcement testing

a. Gathering evidence for litigation: A sex discrimination case

Testing can be used to investigate gender discrimination, and the methodology has also been increasingly used in recent years as a means to gather evidence for employment discrimination litigation. A recent victory was won in *Molovinsky v Fair Employment Council of Greater Washington, Inc. (FEC)*, through which the District of Columbia Court of Appeals upheld a jury award of $79,000 in damages to the plaintiffs, two of whom were testers, in a case involving sex discrimination by the owner of an employment agency.[83] The case first arose when a woman applied for a job with an employment agency and found that the owner acted in an extremely disrespectful and vulgar manner to her—at one point suggesting to her that he wanted her to work for him as a prostitute or that he would pay her for sex. Appalled by this treatment, the woman contacted the Washington Lawyers' Committee for Civil Rights, which then forwarded her case to the FEC.[84]

The FEC sent four testers, two men and two women, to the employment agency. The women testers were treated in ways similarly to the first woman—they were subjected to crude language and offers to work as prostitutes. The male testers, however, were not treated in any way that was out of the ordinary. The jury awarded damages to the original plaintiff, the two female testers, and the FEC. On appeal, the Court of Appeals of the District of Columbia upheld the verdict.[85]

b. Testing for compliance: A pilot program by the OFCCP

Testing can be used to monitor whether organizations are complying with anti-discrimination laws. In 1995, the Office of Federal Contract Compliance Programs (OFCCP), which enforces laws that prohibit federal contractors from employment discrimination, launched a pilot program in the Washington, D.C. area to explore using this methodology. While other federal agencies in the past have used testers to investigate discrimination in housing, this pilot program was the first time the federal government used this technique to investigate employment discrimination.[86]

The test targeted large employers in the banking industry that had not had a compliance review for some time and that were also "flagged" by the Equal Employment Data System. Trained black and white test teams, matched with the same qualifications and characteristics such as demeanor, personality, level of enthusiasm and self-confidence, mailed out resumes and visited employers in

83. 683 A2d 142 (D C 1996).

84. Id at 144-45.

85. Id at 149.

86. Interview with Joseph DuBray, Director of Region III, Office of Federal Contract Compliance Program (May 19, 1997). Vogel, Kelly, Knutson, Weir, Bye and Hunke, Ltd., *Use of 'Testers' in Employment Discrimination Investigation Begun,* North Dakota Employment Law Letter (Dec 1996).

person.[87] The testing revealed that white testers were given greater access to unadvertised job opportunities, were steered to more prestigious and higher-paying jobs, and were interviewed for a longer period of time.[88]

While the data are revealing, some problems with the test design and execution made the OFCCP cautious about interpreting and using the data. The sample size of 14 banks was too small to provide statistically analyzable and generalizable results. There were some problems in selecting employers—it turned out that many companies were not hiring at the time, so both black and white testers were often rejected, which made discrimination more difficult to assess. There was turnover among testers. The study intended, but was not able to conduct multiple tests with the same employer; repeat testing would have allowed the results to be more likely due to discrimination rather than chance. [89]

D. THE VERDICT ON TESTING

1. Strengths

As a research or enforcement methodology, testing overcomes numerous deficiencies of empirical studies and offers the following advantages:

As a research method, testing generates data that are more robust. Regressions can control for easily quantifiable characteristics (e.g., age, the number of years of education, and number of years of work experience.) Testing can do all that and control for qualitative characteristics such as physical appearance (e.g., height or attractiveness), personality, the quality of work experience, and quality of educational background. When "sandwich-testing" is used, or repeated audits of the same employers are made, a testing study can also better distinguish between random events and systemic discrimination.

Testing allows researchers to gather richer data on discriminatory practices, overt and subtle, by employers. It can examine the entire job application and interview process for disparate treatment, while econometrics generally can analyze only a single out-

---

87. Employment Standards Administration, Office of Federal Contract Compliance Programs, *Testers Pilot Program Executive Initiative* 4 (1996) <http://www.dol.gov/dol/esa/public/media/reports/ofccp/testers.htm>.

88. In one test, when a test team applied for a job at the same bank branch, both were told that the branch was not hiring at the time, but the white tester was given the telephone number of the manager at another branch and was told to say that he was referred by the Human Resources Department of the first branch, while the black tester-applicant was told simply that there were no job openings. After making follow-up calls, the black tester was told that his application was out of date and told to submit another application. In another test, the test team applied for bank teller jobs. The white tester was told he was being considered for a "trust processor" position, which paid more than the teller position, and then he was told his resume was submitted for a "portfolio assistant" position, which was another step up from the "trust processor" position. The black tester, in the meantime, was told that only part-time teller positions were available. In yet another test, the black tester was interviewed for 15 minutes, while the white tester was interviewed for 2 hours 10 minutes. Id at 6-8.

89. Interview with Joseph DuBray (cited in note 87).

come (for example, average wages, or hire rate) as a proxy to determine discriminatory behavior.[90] Some real-life examples of employment discrimination that have been uncovered by testing that would have been almost impossible for other research methodologies to detect include:

> *1. Discrimination in opportunity to interview:* For example, two equally qualified and matched testers, one black and the other white, responded to a newspaper advertisement seeking a restaurant manager. The black applicant was not interviewed, while the white applicant was interviewed and offered the job. Even after the white applicant turned down the job offer and the black applicant made repeated calls requesting an opportunity to interview, the restaurant did not respond to those requests.[91]
>
> *2. Discrimination in starting wages:* Two matched testers, one black, one white applied for a sales job in the women's clothing section at a department store. Both were offered jobs, but the black woman was offered $6.50 an hour while the white woman was offered $7.50 an hour.[92]
>
> *3. Discriminatory "steering":* Two matched testers, one black, one white, applied for a sales job at an auto dealership advertised in the local newspaper. The black applicant was told that the entry-level position for a sales job is to be a porter/car washer. The white applicant, with the same qualifications and history, was interviewed immediately for the sales position, with no mention of the porter/car washer entry-level position.[93]

Testing can provide a clear picture of the extent of discrimination in a labor market. Information provided by victimization studies and self-reports tends to be anecdotal and difficult to standardize and quantify—their results cannot be generalized and therefore are not very useful for providing a clear picture of the extent of systematic discrimination in the labor market. Econometric studies can yield findings of discrimination, but the conclusions are more tentatively received—discrimination cannot be proven but only inferred from the disparities in outcomes and an unexplained residual, presumed to be discrimination.

Testing can be used to target a specific city, industry or even employer for policymaking or enforcement purposes. This allows for more efficient allocation of scarce research resources.

As a research or enforcement technique, testing inspires confidence. The procedures are easily understandable and resonate well with the public and courts. Testing results are politically and legally persuasive.

Increasingly, testing is becoming a powerful tool for providing evidence for litigation. The EEOC has proclaimed that testers for employment discrimination have standing to sue, and the District of Columbia Court of Appeals has also recently recog-

---

90. Fix, Galster and Stryuk, *Overview* at 12 (cited in note 59); Bendick, Jackson and Reinoso, 23 Rev Black Pol Econ at 28 (cited in note 75).

91. Bendick, Jackson and Reinoso, 23 Rev Black Pol Econ at 33 (cited in note 75).

92. Id.

93. Id.

nized that testers have legal standing, upholding a verdict awarding damages for testers who encountered sex discrimination.[94] Because the evidence obtained is so persuasive, the use of testing for law enforcement may lower litigation costs. Studies of housing discrimination litigation suggest that cases that use testing as evidence are more likely to settle before the trial and result in a more favorable settlement for the plaintiff.[95]

Testing can have deterrent effects on employers. If a testing program is widely publicized in a region, employers may be more careful not to discriminate. Although this may result in lower levels of discrimination detected, it would bring more employers into compliance with the law and increase equal employment opportunities.

2. Weaknesses

As a technique, testing is not without its limitations:

The technique's investigative powers are largely limited to entry-level jobs or dealings with employment agencies, where a lot of interviewing or other complex interactions with employers are not required. Testing is not feasible for investigating discrimination in promotions, or firing. Nor is it useful for investigating hiring in higher-level positions that require in-depth personal interviews and background checks.

Because fabricated resumes are used, testing becomes immediately ineffective if the employer attempts to check tester-applicants' credentials. This could also distort testing results. For example, an employer who conducts thorough background checks and discovers falsification may refuse to interview or hire both testers. Testing organizations may be able to prevent such situations by instituting mechanisms for substantiating resumes.

Testing is vulnerable to charges of experimenter bias. Although testing methodology is becoming increasingly sophisticated to control for unobservable differences between testers, the technique may still be vulnerable to charges that expectations of the people who conduct the test and testers, who know the purpose of the study, can consciously or subconsciously influence the results. The phenomenon of "self-fulfilling prophesies" have been widely documented and observed in psychological studies, where the mere expectations of researchers changed the behavior of unwitting test subjects. Generally speaking, one of the best ways to protect against experimenter bias is to "double-blind" the study, i.e. the people conducting the experiment do not know the purpose or expected outcome of the study. For testing, however, it would be extremely difficult, if not impossible, to blind testers, for testers must be trained to act out a "script" and taught how to gather and record data. Current efforts at reducing experimenter effects include paying testers flat wages to remove monetary incentives to uncover discrimination.[96] Another way to reduce bias, or perception by others that there may have been a bias, in the selection and matching of testers may be having an

94. *Molovinsky*, 683 A2d 142 (cited in note 84).

95. Fix, Galster and Stryuk, *Overview* at 15 (cited in note 59).

96. Id at 31.

independent and "blind" panel select and match testers to ensure the best matches possible

The sampling frame used by many research tests may underestimate the level of discrimination in the workplace. Many studies have sampled employers that advertise job openings in newspapers. But few jobs are actually obtained this way. More common routes to obtaining employment include word-of-mouth, specialized databases, and personnel placement agencies. Moreover, research has shown that employers that advertise openings in newspapers are less likely to discriminate than employers that rely on personal contacts or referrals.[97]

As a research and evidence-gathering technique, testing has raised some ethical concerns. Testing uses scarce employer resources, such as time and administrative costs. It also involves faking resumes and lying during interviews. Policymakers and commentators have criticized the legitimacy of testing, likening the technique to entrapment and a waste of employer resources,[98] which could lead to negative public sentiments about the practice.

Testing can be costly. For the research results to be considered valid, this type of testing requires large number of separate runs, which could be very costly.[99] A full-scale research audit involving planning and preparation, training of testers, and hundreds of separate tests can cost hundreds of thousands of dollars.[100] The cost of an enforcement audit can range from a few hundred to thousands of dollars, depending on the complexity of the discrimination being investigated. The price includes wages for testers and organizers, overhead expenses, and costs of analyzing results.[101] The cost generally does not include legal expenses or costs imposed on employers, such as administrative costs of processing testers' applications.

V. "VICTIMIZATION STUDIES"

Another way in which discrimination in employment can be examined is through polls and surveys of people who have been victimized by discrimination.

---

97. Heckman and Siegelman, *Methods* at 213 (cited in note 58); Fix, Galster and Struyk, *Overview* at 32 (cited in note 59).

98. See The Future Direction of the EEOC, Hearing before the Subcommittee on Employer-Employee Relations of the House of Representatives Committee on Education and the Workforce, 105th Cong, 2d Sess 62 (March 3, 1998) (statement of Harris W. Fawell, Subcommittee Chairman, arguing against an EEOC initiative to use testers to detect employment discrimination). See also id at 7 (statement of Newt Gingrich, Speaker of the House, arguing that the use of testers wastes employer resources and amounts to entrapment). See also Susan J. Wells, *The Hunt for Bias In Hiring,* NY Times 3:12 (March 8, 1998) (quoting the National Federation of Independent Business calling testing "reprehensible" and "misleading," and quoting Barry Lawrence, a spokesman for the Society for Human Resource Management as saying that testers tend to conduct "covert operations" and that testing "doesn't seem to be a very ethical practice.")

99. Fix, Galster and Struyk, *Overview* at 1 (cited in note 59).

100. Interview with Michael Fix (cited in note 62).

101. Interview with Kennington Wall (cited in note 62).

Although these studies are often widely reported by the news media, the results are generally not recognized as meaningful statistical measures for discrimination.[102]

Take, for example, the study commissioned in 1994 by the Women's Bureau of the Department of Labor. To conduct the survey, questionnaires were distributed nationwide through newspapers, magazines, businesses, community organizations, and even labor unions. The bureau received 250,000 back over a four-month period, and to supplement that data, it also conducted a telephone survey using the same questions on a "scientific," nationally representative sample of 1,200 women. Among the findings of the "scientific sample" was that 14 percent of white women and 26 percent of minority women surveyed reported losing a promotion or a job because of their gender or race.[103]

Another example of such a study is the survey of 3,000 women by Roper Starch Worldwide, sponsored by cigarette-maker Philip Morris USA. Of the women polled, 84 percent agreed with the statement that "... regardless of changes that may have occurred, women still face more restrictions in life than men do," and 77 percent reported that sexual discrimination in the workplace "remains a serious problem."[104]

The primary value of these studies lies in the information they provide to the public and policymakers about people's perceptions, values and priorities as witnesses or victims of employment discrimination. But the very essence of such studies, which is an invitation for victims to step forward, identify themselves and tell about their experiences, can also produce misleading results. These studies are often flawed by "response bias"—people who have had experiences with discrimination or strong opinions on the subject are more likely to take the time to respond than those who do not know or care at all, thus distorting the results.[105]

VI. SELF-REPORTS

A less common but still sometimes useful methodology for investigating discrimination is through so-called self-reports, or surveys of employers on the extent of their discriminatory practices. These studies can be administered by mail, in-person or over the phone,[106] and can provide information to the public and policymakers about discrimination among employers. As a data-gathering technique, however, these survey do not inspire a lot of confidence because em-

---

102. Fix, Galster and Struyk, *Overview* at 13 (cited in note 59).

103. The study received 250,000 questionnaires that were distributed through more than 1,000 businesses, community organizations, labor unions, newspapers, and magazines nationwide. Tamar Lewin, *Working Women Say Bias Persists*, NY Times 1:9 (Oct 15, 1994).

104. Judith H. Dobrzynski, *Women Less Optimistic About Work, Poll Says*, NY Times D5 (Sept 12, 1995).

105. Oversight Hearing on the Impact of *Adarand v Peña* at 44 (statement of George R. LaNoue) (cited in note 5).

106. Fix, Galster and Stryuk, *Overview* at 12-13 (cited in note 59).

ployers can be deceptive about, or not even aware of, their own discriminatory practices.

An example is the study conducted by Kirschenman and Neckerman in 1989 on employers in Chicago.[107] For this study, the researchers conducted interviews with a sample of 185 Chicago-area businesses. Employers were asked both closed and open-ended questions on their attitudes and hiring practices. Responses were then coded, categorized and analyzed.

Obviously, this method yields valuable data only when employers are completely honest about themselves in their experiences, perceptions and practices in their responses. For this study, however, the authors believed that the responses were frank. "[W]e were overwhelmed by the degree to which Chicago employers felt comfortable talking with us—in a situation where the temptation would be to conceal rather than reveal—in a negative manner about blacks," Kirschenman and Neckerman wrote.[108] When asked if there were differences in the work ethic of whites, blacks and Hispanics, only half of the employers surveyed said they believed there were no differences between the races. More than a third ranked blacks last, and 7.6 percent ranked Hispanics last. Employers also expressed worry about racial tensions and believed that a homogenous work force creates better relations among employees.[109] There were also views among black employers that poor blacks were more likely to be dishonest than other groups because of the economic pressures they face.[110]

The study found that employers considered race and class to be important factors in hiring decisions. Inner-city workers, especially African-American males, were believed by employers to be unstable, dishonest, involved with drugs and gangs, lacking a work ethic or lazy, and having no personal charm. The results of their study suggested that employers used group membership, such as race, as a proxy in assessing an individual's labor productivity.[111]

While the results of this study may not be generalizable to the population of employers at large, it did paint an impressionistic picture of employer attitudes and yielded an important insight—that a good deal of discrimination may be due to employers' misperception of worker productivity. Earlier studies often assumed that discrimination was irrational firm behavior: as firms seek to hire the best workers for the job, discrimination would lead to higher costs, and competitive market forces will drive the firms that discriminate out of business. It could be that employers are acting fully rationally based on wrong signals and information.

---

107. Joleen Kirschenman and Kathryn M. Neckerman. *"We'd Love to Hire Them, But. . .": The Meaning of Race for Employers*, in Christopher Jencks and Paul E. Peterson, eds, *The Urban Underclass* 203 (The Brookings Institute, 1991).

108. Id at 207.

109. Id at 211.

110. Id at 213.

111. Id at 204.

VII. IMPLICATIONS FOR THE FUTURE

Many lessons may be learned from the analyses of the different types of methodologies used to measure discrimination. A few include the following:

Different methodologies are complementary, and should be used as such. Statistical evidence, whenever possible, should be bolstered by anecdotal evidence, or victim testimony that can be provided by testers. Courts may have held that if the statistical evidence is highly compelling, then no anecdotal evidence may be needed.[112] But weaker statistical evidence, unaccompanied by any anecdotal evidence, would not provide sufficient proof of discrimination.[113] Especially where the evidence is weak, anecdotal evidence may be the critical factor for courts in deciding whether there was indeed discrimination.

Comprehensiveness is key. Each measurement type should be pushed to the limit. In the context of regressions, a more thorough analysis would include a better specification of the factors that influence employment outcomes. Instrumental variables and multi-tiered tests are more comprehensive ways to approach problems that have traditionally escaped the abilities of simpler forms of multivariate analysis

More information regarding discriminatory influences can be revealed by tracking people over time. Disparities exist not just in hiring but can also emerge in "tracking" over time, or the existence of the proverbial glass ceiling. Longitudinal studies may serve as better measures of tracking discriminatory factors that influence occupational mobility of women and minorities. Researching and analyzing career experiences of minority employees compared to whites over time may be a better way to measure progress, educate the public and policymakers about the problem areas and needs for future anti-discrimination efforts.

In addition, policy makers can play an active role in facilitating better analyses of discrimination measurement. Among the issues that policymakers should consider are:

How could better data be collected? All measurements of discrimination require good data for accurate analyses, but often the information is incomplete or not available. Requiring better record-keeping by employers and public access to information may be one solution, as is the case of public contracting, where the federal government requires that contractors keep detailed employment records for data analysis purposes.

---

112. See *Hazelwood School District*, 433 US at 307-308 (cited in note 24). See also *Segar*, 738 F2d at 1278 (noting that "when a plaintiff's statistical methodology focuses on the appropriate labor pool and generates evidence of discrimination at a statistically significant level, no sound policy reason exists for subjecting the plaintiff to the additional requirement of either providing anecdotal evidence or showing gross disparities") (cited in note 10).

113. See *Sears*, 839 F2d at 311 (cited in note 19) (noting that "examples of individual discrimination are not always required, but we think that the lack of such proof reinforces the doubt arising from the questions about validity of the statistical evidence," citing *Griffin v Board of Regents*, 795 F2d 1281, 1292 (7th Cir 1986)).

How should one equitably balance the costs incurred by the errors that could be made? Further discussion needs to be conducted regarding this issue. This issue is particularly a problem in disparity studies, where the selection of the significance level determines the balance between minimizing false accusations of innocent firms and minimizing false acquittals of guilty ones. A common complaint of the significant level used by the courts is that the 5 percent level used by social scientists is "arbitrary"; however, there has been no serious discussion in the legal or policy community on what a better criterion might be.

APPENDIX A: LINGERING ISSUES ON TESTING

A. HOW DISCRIMINATION SHOULD BE MEASURED

There have been disagreements among researchers of the testing methodology on how discrimination should be measured. Part of the disagreement arises from different views of what constitutes discrimination. The gross approach, which resembles the legal notion of discrimination, posits that any unfavorable treatment of minorities ought to be considered discrimination, while the net approach, which is more aligned with an economic world view, subtracts the amount of unfavorable treatment of majorities from the amount of unfavorable treatment of minorities for a grand total "measure" of discrimination. Another part of the disagreement arises over the how random events should be distinguished from systematic discriminatory behavior; unequal treatment could be due to discrimination or error.[114]

Differences over how to account for random errors and how to define discrimination can have important implications on how data are interpreted. Fix, Galster and Stryuk identify four different ways to analyze "disparate treatment" that starkly illustrate these differences in measurement and outcomes: (1) "Discriminatory Inclination," or when agents systematically penalize an applicant because of race, whether or not the ultimate outcome is discriminatory; (2) "Gross Unfavorable Treatment," or when agents systematically give the majority applicants greater access, whether or not the treatment was due to discrimination or chance; (3) "Systematic Unfavorable Treatment," or when agents systematically give the majority applicants greater access because of penalities based on race, but penalties that do not result in unequal access or incidences due to chance are not considered discriminatory; and (4) "Net Market Effects," or the comparison of majority-favored actions with minority-favored actions, creating a "net" effect of discriminatory outcome in the market.[115]

---

114. John Yinger, *Audit Methodology: Comments*, in Michael Fix and Raymond J. Struyk, *Clear and Convincing Evidence: Measurement of Discrimination in America* 259, 261 (The Urban Institute Press, 1993).

115. Fix, Galster and Struyk, *Overview* at 26 (cited in note 59).

The four different "measures" of discrimination can be restated as follows:[116]

F = the proportion of agents audited who had discriminatory inclination (who systematically penalized an applicant because of race)

P_{maj} = proportion of cases favoring the majority

P_{min} = proportion of cases favoring the minority

P_{rmaj} = proportion of cases favoring the majority because of random factor

Definition of Discrimination	Existence Measure
(1) Discriminatory Inclination	$f > 0$
(2) Gross Unfavorable Treatment	$P_{maj} > 0$
(3) Systematic Unfavorable Treatment	$P_{maj} - P_{rmaj} > 0$
(4) Net Market Effects	$P_{maj} - P_{min} > 0$

Depending on which measure is used, the same data could lead to completely different conclusions about the type and extent of discrimination. Each of these definitions also has its advantages and disadvantages. Notice that definition (1) does not distinguish between discriminatory proclivities and discriminatory outcomes. Definition (2) does not distinguish random from systematic events, which makes it elegant and easy to use but could distort the true extent of systematic, nonrandom discrimination. Randomness is factored out in definition (3), which requires that random events be distinguished from systematic discrimination. However, whether or not an event was caused by chance can also be difficult to ascertain, an additional consideration that could make it difficult for courts to find discrimination. Definition (4) could be looked at in two ways. First, discrimination in favor of minorities (P_{min}) could be thought of as a proxy for the extent of random events that favor the majority (P_{rmaj}). But that would presume symmetry in random factors, which may not be the case. Second, if there is significant systematic discrimination against minorities and favoritism toward minorities (reverse-discrimination), then using the "net" approach could paint a misleading picture that there is little or no discrimination when in fact there could be rampant discrimination going both ways.[117]

Currently, many within the research community appear to be converging on using the "net" definition of discrimination,[118] but this definition, as just men-

116. Id at 28-29.

117. Id at 26-8.

118. Interview with Michael Fix (cited in note 62).

tioned may disguise the true nature and extent of discrimination. Some refinements have been made for estimating "gross" discrimination. To better account for random events, three- or four-person test teams could be used. Conducting tests with more than two testers, such as two Anglos and one Hispanic in a team, would allow researchers to better assess the effect of random factors on outcomes.[119]

B. WHETHER TESTERS HAVE "STANDING" TO SUE

Another issue that has not been fully resolved is the legal issue of "standing" for testers of employment discrimination. A party is said to have standing to sue another in a court of law if he or she has a personal, tangible and legally protectable interest at stake. For many years, the Supreme Court has recognized the rights of testers to sue in cases of public transportation, housing, and lending discrimination.[120] But as of yet, the highest court has not determined whether testers in employment discrimination have the right to sue.

Given the current legal landscape and statutory framework, it is highly likely that testers will be found to have standing in employment discrimination cases. Although in 1994, the U.S. Court of Appeals for the District of Columbia Circuit held that testers investigating a company for employment discrimination have no constitutional standing to sue,[121] it is recognized in legal circles that that the decision has little persuasive power on subsequent tester standing issues. This is because the case was filed before the 1991 Civil Rights Act was amended, legislation which now provides monetary damages as a remedy for discrimination against any individual, including testers.[122] In 1996, the District of Columbia Court of Appeals held that testers did have standing to sue under the D.C. Human Rights Act. [123] The D.C.H.R.A. contains a section almost identical to the provision in the amended Civil Rights Act, and both give any individual unlawfully discriminated against by an employer the right to sue for damages.

In a report issued last year, the EEOC stated that testers who are discriminated against have standing to sue and may be entitled to compensatory and/or

---

119. Yinger, *Audit Methodology* at 263 (cited in note 115).

120. See *Evers*, 358 US 202 (cited in note 57) (holding that tester investigating segregation in seating on a bus had right to sue). See also *Havens Realty Corporation v Coleman*, 455 US 363 (1982) (holding that testers may bring claims under the Fair Housing Act).

121. *Fair Employment Council of Greater Washington, Inc. v BMC Marketing Corporation*, 28 F3d 1268 (D C Cir 1994).

122. The *BMC* court pointed out that at the time of the alleged discrimination, "only equitable remedies were available under Title VII," and testers could not show future injury, since they had no intention of accepting jobs. Id at 1272-73. See also Interview with Claudia Withers, Executive Director of Fair Employment Council of Greater Washington, DC and Council's attorney in *BMC* case (May 13, 1997); *Molovinsky*, 683 A2d at 146 (cited in note 84).

123. *Molovinsky*, 683 A2d at 146 (cited in note 84).

punitive damages under Title VII of the Civil Rights Act.[124] Testers encountering discrimination may also be entitled to injunctive relief and attorney's fees, but would not be entitled to reinstatement or back pay, because they did not intend to accept the jobs.[125]

124. U.S. Equal Employment Opportunity Commission Office of Legal Counsel, *Enforcement Guidance: Whether "Testers" Can File Charges and Litigate Claims of Employment Discrimination* 2, 6 (Document No. N-915.002) (1996) <http://www.eeoc.Gov/docs/testers.txt>.

125. Id at 8.

Appendix B: Summary of the Purposes and Uses of Employment Discrimination Measuring Techniques

	Multivariate Analyses	Disparity Studies	Testing	Victimization Studies	Self-Reports
What the methodology can measure	Discrimination in: ♦ Hiring ♦ Wages ♦ Promotion ♦ Termination	Discrimination in: ♦ Hiring ♦ Wages ♦ Promotion ♦ Termination	Discrimination in: ♦ Opportunity to interview ♦ Length of wait before interview ♦ Length/quality of interview ♦ Offer of employment ♦ Starting wage and benefits ♦ Type of job offered (steering) ♦ Offer of additional opportunities	Perceptions of victimization	Employer attitudes and perceptions
Possible uses	♦ Educate public and policymakers ♦ Used as evidence for employment discrimination litigation cases ♦ Identify and determine magnitude of discrimination	♦ Educate public and policymakers ♦ Used as evidence for employment discrimination litigation cases ♦ Infer discriminatory exclusion	♦ Educate public & policymakers ♦ Enforce laws — gather evidence — monitor compliance ♦ Identify problem areas ♦ Give feedback on anti-discrimination efforts	Educate public and policymakers	Educate public and policymakers

Appendix C: Strengths and Weaknesses of Employment Discrimination Measuring Techniques

	Multivariate Analyses	Disparity Studies	Testing	Victimization Studies	Self-Reports
Primary Strengths	◊ Inexpensive ◊ Can measure magnitude of impact of certain characteristics on outcomes ◊ Can separate out factors influencing employment outcomes being measured ◊ Not subject to experimenter bias ◊ Can examine employment discrimination at any level ◊ No ethical concerns	◊ Relatively inexpensive ◊ Not subject to experimenter bias ◊ Can examine employment discrimination at any level	◊ Ability to control for applicant differences generates robust data ◊ Yields data on overt and subtle discrimination ◊ Provides quantitative *and* anecdotal evidence ◊ Can target a region, industry, or employer ◊ Can deter future discrimination	Offer insights into victim perceptions, especially differences in results among groups	Offer insights into employer attitudes and perceptions
Problems And limitations	◊ Omitted variables could bias results ◊ If included variables are affected by discrimination, the regression will measure only "incremental" discrimination ◊ Difficult to determine specific employment practices that led to discrimination	◊ Appropriate data may not exist ◊ Arbitrary use of 5% rule for finding guilt ◊ Chances may mistakenly be excluded as a cause for discriminatory outcomes	◊ Limited to hiring for low-level positions ◊ Cannot investigate job mobility or firing ◊ Becomes ineffective if employer conducts background checks ◊ Vulnerable to experimenter bias ◊ Raises ethical concerns	◊ Respondents may not be truthful ◊ Respondents may not be aware of own experiences with discrimination ◊ Possible response bias	◊ Respondents may not be truthful ◊ Respondents may not be aware of own discriminatory practices ◊ Possible response bias