Three Tests for Measuring Unjustified Disparate Impacts in Organ Transplantation

the problem of "included variable" bias

Ian Ayres

ABSTRACT Civil rights statutes often prohibit two distinct types of discrimination, referred to as "disparate treatment" and "disparate impact." Disparate treatment is race-contingent decision making. But even decision making that is not affected by people's race may still produce an unjustified disparate impact. For example, a race-neutral transplantation preference for allografts with partial antigen matches might produce an unjustified disparate impact on African Americans with end-stage renal disease. The transplantation preference might make it harder for African Americans to receive a transplant without significantly increasing the chance of transplant survival. Because disparate impact and disparate treatment claims have distinct elements, they require distinct methods of statistical testing. This article analyzes three different ways of testing unjustified disparate impacts in organ transplantation, which I will call the traditional test, the omitted variable test, and the outcome test. Each of these methods of testing for disparate impact are attuned to the problem of "included variable" bias. Controlling statistically for nonracial variables may actually bias the analysis and mask the existence of unjustified disparate impacts.

Townsend Professor, Yale Law School, P.O. Box 208215, 127 Wall Street, New Haven, CT 06520. E-mail: ian.ayres@yale.edu.

This paper draws on Chapters 6 and 11 of the author's book *Pervasive Prejudice* (2002) and on his work as an expert witness in *Cason v. Nissan Motors Acceptance Corporation* (M.D. Tenn. No. 3-98-0223) and similar lawsuits.

Perspectives in Biology and Medicine, volume 48, number 1 supplement (winter 2005):S68–S87

THERE ARE TWO DIFFERENT WAYS to analyze disparities: disparate treatment and disparate impact. *Disparate treatment* involves the conscious decisions of individuals to treat certain people differently, based, in this case, on race (this could also be called race-contingent decision making). Tests of disparate treatment are one way of identifying normatively problematic actions. (There are, however, exceptions. The Equal Protection Clause does countenance disparate treatment by government if it is narrowly tailored to further a compelling government interest [e.g., *Grutter v. Bollinger* (2003), upholding race-conscious admissions programs].) *Disparate impact* studies, by contrast, look at the effect of race-neutral policies on individuals of different races, for example, by disproportionately excluding minority recipients. The fact that there exists a disparate impact does not necessarily raise the same normative concerns as does disparate treatment. An examination of disparities in organ transplantation should use disparate impact to see whether there is an unjustified disparity, since it is unlikely there is systematic disparate treatment of patients based on race.

The law defining the contours of disparate impact liability is a good place to find tests for problematic disparities that may serve as a useful normative benchmark for policymakers. In the employment context, a plaintiff bringing a disparate impact claim need not show that the defendant engaged in race-contingent decision making (or disparate treatment). Instead, the plaintiff need only show that the defendant's policies caused a disparate racial impact. This impact is legally justified only if "the challenged practice is job related for the position in question and consistent with business necessity" (U.S. Code 42 [1994], §2000e-2(k)(1)(A)(ii)), and no alternative employment practice exists that would satisfy the employer's legitimate interests with less of a disparate impact on a protected class (HR-Guide 2001). Because the elements of a disparate impact claim are quite different from the elements of a disparate treatment claim, there need to be distinctive tests for this type of litigation.

DISPARATE TREATMENT TEST

Social scientists have some consensus on the appropriate methods of testing for disparate treatment. When econometricians attempt to test for disparate racial treatment, the goal in a regression analysis is to control for all of the nonracial variables that might have explained a particular set of decisions. The regression asks, for example, whether—after controlling for all potential nonracial variables—a hospital treated a transplant applicant differently because of her race.

In disparate treatment cases, omitted variable bias is often a primary concern. If a disparate treatment regression omits or fails to include a nonracial variable upon which the decision maker actually based her decision, then the regression can erroneously indicate that the decision maker treated minorities differently from whites. For example, if the decision maker has a practice of excluding transplant applicants without a high-school diploma from the transplant list, and if we

further assume that the pool of applicants without diplomas is disproportionately comprised of minorities, then omitting from the regression a control for whether applicants graduated from high school might bias the test of disparate treatment. The regression might superficially indicate that the hospital was less likely to place African Americans on the waiting list, when in fact the hospital was less likely to place applicants without diplomas on the list.

DISPARATE IMPACT TESTS

Tests of disparate impact require a different statistical method. Under disparate impact theory, it is possible for decision-making policies that are facially race-neutral to give rise to liability if they disproportionately burden the plaintiff class. For example, a practice of excluding from the waiting list applicants without a high-school diploma might raise disparate impact concerns if this nonracial criterion disproportionately burdens minorities without cause.

Under disparate impact theory, it is necessary to intentionally omit nonracial variables from a regression to test whether those variables produce a disparate racial impact. In disparate treatment regressions, the idea is to test whether, after controlling for all possible nonracial factors, there is still a racial disparity in decision-making that can therefore be attributed to the decision maker's intentional discrimination. But in a disparate impact case, the idea is to test whether nonracial factors might have caused a racial disparity in the first place. It is inappropriate to control for these nonracial factors in the regression analyzing the impact of a particular set of decisions because we want to see whether these nonracial factors produce racially disparate outcomes.

While econometricians are normally worried about omitted variable bias, when testing for disparate impacts they often need to be more worried about included variable bias.[1] Including controls for nonracial factors that do not represent legitimate business justifications can bias the estimate of whether a decision maker's policies produced an unjustified disparate impact. For example, a recent statistical guide for judges and lawyers emphasizes how including irrelevant variables can bias a regression's estimate of the racial effect:

> Suppose a regression analysis includes a variable for education that, in a race case, is a key determinant of salary differences between black and white employees in a clearly different job group. Regression analysis indicates a high t-statistic on education and an insignificant t-statistic on the race coefficient. Given that in almost all groups, white employees have received more formal education than

[1] The term "included variable bias" is also used by Clogg and Haritou (1997). They point out that adding variables that are correlated with the error term of the regression can bias the estimate of other coefficients of interest: estimating a model with additional controls can cause included variable bias "in spite of the fact that this model may very well lead to reduction in the variance of the prediction. This term is conspicuous by its absence in the literature" (100–101).

> black employees, it would appear that education goes a long way towards explaining salary differences between black and white employees. The burden is on the employer, however, to demonstrate separate from the regression, that education was required and affected performance, and hence directly determined salary. To the extent that education is not related to job performance, it is an inappropriate variable to use in a regression. Excluding key variables and including irrelevant variables have the same impact. (Ireland et al. 1998)

The purposeful exclusion of control variables from statistical analysis will accordingly be an essential part of any disparate impact inquiry. Indeed, as the foregoing authority suggests, a variable should be presumptively excluded from the statistical analysis unless the defendant can "demonstrate separate from the regression that [the variable] was required and affected performance."

John Yinger (1998) also succinctly describes the problem of "included variable bias" (what he calls "diverting variable bias"); the need to purposefully exclude certain non-legitimate controls from a regression; and which variables constitute "legitimate" controls:

> Diverting variable bias arises when a variable that is not a legitimate control variable, but that is correlated with race or ethnicity, is included in the regression. The key issue, of course, is how to define what variables are "legitimate." Under most circumstances, economists are taught to err on the side of including too many variables. In this case, however, illegitimate controls may pick up some of the effect of race or ethnicity and lead one to conclude that there is no discrimination when in fact there is. According to the definition of discrimination used here, legitimate controls are those associated with a person's qualifications to rent or buy a house, buy a car or so on—or to use a legal term business necessity. (27)

The problem of included variable bias can also be illustrated by a stylized version of *Griggs v. Duke Power Co.* (1971), the Supreme Court's first disparate impact case. One could imagine running a regression to test whether an employer was less likely to hire African American applicants than white applicants. It would be possible to control in this regression for whether the applicant had received a high-school diploma. Under the facts of *Griggs*, such a control would likely have reduced the racial disparity in the hiring rates. But including in the regression a variable controlling for applicants' education would be inappropriate. The central point of *Griggs* was to determine whether the employer's diploma requirement had a disparate racial impact. The possibility that including a diploma variable would reduce the estimated race effect in the regression would in no way be inconsistent with a theory that the employer's diploma requirement disparately excluded African Americans from employment.

Excluding nonracial factors is inappropriate in disparate treatment tests, but such exclusion is *necessary* in disparate impact tests so as not to bias the coefficient of interest. In disparate impact regressions, it is thus necessary to inten-

tionally exclude even true "causal" variables from the analysis.[2] In a *Griggs* disparate impact regression estimating the probability that particular applicants will be promoted, the high-school diploma variable is excluded, even though it in fact is believed to truly influence whether particular applicants will be accepted. But only by excluding this causal variable can we estimate whether the employer's diploma requirements in fact have a disparate impact. By running the regression both with and without the diploma control, one can estimate how much the diploma requirement contributes to the overall disparate impact of the employer's hiring practices.

Just as it would be inappropriate to include a high-school diploma variable in the *Griggs* context, it would have been inappropriate for me to include an antigen-match variable into my disparate impact tests of kidney transplantation (Ayres, Dooley, and Gaston 1993). The point of that study (about which I'll say more) was not to test whether African Americans with equal antigen matches were less likely to qualify for transplantation, but whether the antigen-matching point system caused an unjustified disparate impact.

So something different from disparate treatment testing is needed. Unfortunately, social scientists do not have a well worked out theory about how to test for unjustified disparate impacts. Many expert witnesses in disparate impact cases still cling to disparate treatment regressions, notwithstanding the radically different elements of a disparate impact test.

The search for authoritative disparate impact tests is made more difficult by the lingering uncertainty about the scope of the legal elements. More than 30 years after *Griggs* and a dozen years after the purposefully vague Civil Rights Act of 1991, there is still not legal clarity on just what constitutes a policy that is "job related for the position in question and consistent with business necessity"; whether a defendant is liable when both the defendant's and plaintiff's actions are but-for causes of the disparate impact; or what variables are necessary in constructing the "qualified pool" for disparate impact analysis. In some sense, a definitive statistical methodology cannot be derived until these legal uncertainties are resolved. A river cannot rise above its source. But since the purpose of this article is to use disparate impact liability as a normative benchmark, I will

[2]At the conference, my commentator James Heckman provided an example in which there could be included variable bias even in disparate treatment regressions. Imagine for example that an employer's hiring decisions are known to be completely determined by four variables: the applicant's race, education, prior work experience, and age. It turns out that a regression that controls for three of these four attributes (race, age, education) may produce more biased estimates about the influence of race than a regression that controls for just two of the attributes (race and age). When a researcher cannot control for all the causal variables, then controlling for a larger subset of causal variables does not necessarily produce less biased results. But this problem of included variable bias is even larger in disparate impact analysis. In the foregoing example, a researcher who had access to all four of the causal variables might still need to intentionally exclude one in order to test whether its exclusion induced a disparate impact in the estimated race coefficient.

proceed to suggest tests that grow out of certain assumptions about the contours of legal liability. In particular, this article briefly lays out three different methods of testing for unjustified disparate impacts, which I will refer to as the traditional (or two-part) test, the omitted variable test, and the outcome test. The goal will be to assess the extent to which these tests can be made to resonate with particular conceptions of the law or with our normative predilections for what constitutes unjustified racial disparities.

The Traditional Test

The traditional test of disparate impact separately tests whether a particular policy disparately burdens minorities, and whether the policy furthers a legitimate interest of the decision maker. For example, in an employment setting, it would be traditional to assess the relative exclusion rate of white and black applicants caused by an employment policy to establish a prima facie case of disparate impact liability and then separately analyze whether that policy is valid. The first part establishes whether or not the test produces a disparate impact; the second establishes whether or not the impact is justified.

At first, there would seem to be a fair amount of consensus about how to test for a prima facie impact. Indeed, in the employment context, Web sites will automatically calculate a disparity ratio and tests of statistical significance for the casual user (HR-Software 1998). It is essentially this methodology that Laura G. Dooley, Robert S. Gaston and I applied to kidney transplantation (Ayres 2002; Ayres, Dooley, and Gaston 1993; Gaston et al. 1993, 1994, 1995). We used different analyses to establish that HLA matching made it more difficult for African Americans to qualify for transplantation and, separately, that HLA matching (especially giving points for partial matches) did not increase expected allograft survival.

The first part of our analysis documented how the antigen-matching preferences caused a disparate impact against potential African American recipients. Antigen matching restricts the availability of cadaveric kidneys for black patients for the simple reason that most donors are white, and white kidneys tend to have different antigens from black kidneys. For example, a study in Illinois calculated how well 352 cadaveric kidneys matched 604 patients on the local United Network for Organ Sharing (UNOS) waiting list. The study revealed that while only 52% of the overall waiting list was white, whites dominated the class of recipients having four or more antigens matching—with 71.8% of these well-matched kidneys (Lazda and Blaesing 1989).[3]

[3]Potential black recipients, who made up 39.9% of the overall waiting list, comprised only 16.2% of the four or more antigen matches. These discrepancies are further exacerbated if the analysis is restricted to matching the cadaveric kidneys from white donors. In that case, white patients would receive 75.2% of the four or more antigen matches, and black patients would receive only 14%. These latter figures may be more relevant on a nationwide level because the Illinois study contained a relatively elevated proportion of cadaveric kidneys from black donors (13.9%).

The second (and separate) part of our analysis showed that the point system's strong historic preference for partial antigen matches was not medically justified by longer survival rates. The point system placed heavy weight on the quality of the antigen match, making 10 out of approximately 17 possible points contingent on the number of antigens matched. By contrast, the system gave only one point to the patient who had waited for the longest period; those who had not waited as long got fractions of a point (UNOS 1992). The net result was almost complete emphasis on antigen matching in determining allocation, with time on the waiting list serving largely as a tie-breaker. Thus, in vying for a particular kidney, a patient with only one antigen matched could conceivably be awarded a kidney over someone who had waited up to two years longer.

While six-antigen matching—and possibly zero-antigen mismatching—has been shown to significantly enhance kidney transplant survival, we showed that there is a much weaker correlation between the quality of matching and transplant survival when one or more antigens are mismatched. For example, in a 1989 single-center report, while whites and blacks had different survival rates, matching for one or more antigens did not make a statistically significant difference in patient or graft survival at one, two, or three years for either white or black recipients when compared to transplants with no matched antigens (Greenstein et al. 1989).

The combination of these two types of analysis provided strong and straightforward evidence that points for partial antigen matching created a medically unjustified disparate impact. It forced African Americans to wait longer for transplants without increasing the expected transplant survival rate. Fortunately, UNOS rules have subsequently been modified along the lines that we suggested to ameliorate this problem (Ayres 2002, 483).

But there are important flash points of disagreement about how to conduct the traditional test. The disparity prong of the test can proceed only if one first constructs what is sometimes called a "qualified pool" that essentially determines the universe of data to be analyzed.[4] For example, in the employment context, when testing whether an employer's hiring practices have a disparate racial impact, it is appropriate to limit comparison to the group of candidates who are qualified in the sense that they meet minimum characteristics for employment. Thus, in the airline industry, it would be appropriate to limit the qualified pool of pilot applicants to those applicants who were licensed to fly.

Defendants are increasingly convincing courts to consider limiting the qualified pool to a smaller universe of individuals. In effect, the courts are converting the back-end justification inquiry into a front-end qualification inquiry. If the court determines that a high-school diploma is one of the minimum job re-

[4]I have previously noted that "the Supreme Court in *Hazelwood* and subsequent opinions required plaintiffs to calculate the racial composition of 'the qualified . . . population in the relevant labor market' or the 'otherwise-qualified applicants'" (Ayres and Siegelman 1996, 1492).

quirements, this determination is tantamount to a finding that the use of this requirement is justified and precludes the plaintiff from testing whether the diploma requirement created a disparate impact. Limiting the qualified pool to diploma holders would have killed the plaintiff's case in *Griggs*.

But courts have not been uniform in their qualified pool determinations. While courts have sometimes limited the qualified pool in hiring cases to subsets of applicants, they tend not to limit the qualified pool in firing cases to subsets of current employees (Ayres and Siegelman 1996). But of course it would be possible to argue that not all current employees are equally "qualified" for firing. The qualified pool in a layoff case as a theoretical matter might be limited to less productive and/or absentee employees.

In the transplant context, it is the norm to implicitly adopt a qualified pool of either people needing transplants or people who are placed on recipient waiting lists. But here too it would be possible to limit the universe of analysis to a more refined subset—say, people who had well-matched kidneys.

Raising the qualification bar, in my mind, tends to muddy the waters. While the justification prong of the traditional test is crucial to the test's normative relevancy, the double counting created by front-end and back-end justifications is, if anything, a move away from an overarching theory of justification. Of course, the even bigger difficulty in implementing the traditional test concerns the scope of this back-end justification. The Civil Rights Act of 1991 unhelpfully countenanced disparate impacts if they were "job related for the position in question and consistent with business necessity" (U.S. Code 42 [1994], §2000e-2(k)(1)(A)(ii)). What this means in the employment context no one is exactly sure, making it all the harder to translate this standard for use in non-employment settings.

In the transplantation setting, the natural justificatory benchmark would seem to concern medical success. But what measure of medical success? Should allograft survival (five-year survival rate or half-life) or quality of life be the measure? And these measures by themselves are notoriously noneconomic. Shouldn't the expected costs of producing the allograft survival be taken into account in some way? My prior work on kidney transplantation relied on an empirically contingent case that attempted to avoid these difficult issues. The partial-antigen-matching rules disparately excluded African Americans without increasing transplant survival (or any measured evidence of quality of life). The number of transplants at issue was constrained by the number of cadaveric donations, and hence our suggested de-emphasis of HLA points was a wash as far as transplantation costs were concerned.[5]

The tougher problem concerns race-neutral policies that dramatically exclude African Americans but mildly enhance medical outcomes. Should medical efficacy ever be sacrificed in the name of equity? The kidney matching rules have

[5]However, further analysis might indicate that transplantation to African American recipients entailed higher costs of post-transplantation treatment (Gaston et al. 1993).

long done this on other grounds—favoring blood type O and pre-sensitized and longer-waiting recipients regardless of poorer antigen matches.[6]

Without answers to these questions, the best that an empiricist can do is to try to evaluate whether a tradeoff is necessary and, if so, what the terms of trade are. To ameliorate the disparate impact of a particular policy one must first determine how much needs to be sacrificed in terms of survivability.

The Omitted Variable Test

The bifurcated structure of the traditional test is slightly unsatisfying. In this and the next section, I will sketch the strengths and weaknesses of two alternative tests that provide a more unified analysis of whether a policy is producing an unjustified disparate impact.

Tests of disparate impact, unlike tests of disparate treatment, must to some degree embrace omitted variable bias. Even in the first prong of the traditional test, it is not possible to determine whether a particular transplant requirement produces a disparate impact, if that requirement becomes part of what the qualified pool controls for.

The omitted variable test is an attempt to provide a more thorough view of how to exclude variables from a regression. The basic idea is to include in a regression those variables that would reflect a valid justification for the policy in question. The regression would allow the justified regressor to absorb the disparate racial impact that might have been found in a simple difference-of-means test. If, after including these "justification" variables in the regression, the racial disparity is eliminated (or becomes statistically insignificant), then the regression indicates that the disparate impact is justified. But as before, it is essential to exclude unjustified variables so that these will not absorb what would otherwise be a racial disparity.

It would also be necessary to partially omit even justified variables by including them in the regression but limiting the maximum or minimum size that the coefficient could take. For example, in the lending context, imagine that it was justifiable to charge borrowers in a higher-risk credit tier a 2 percentage point higher interest rate. Then, in a regression to test for unjustified disparate impacts in the lending terms, it would be appropriate to add a control variable for borrowers in the high-risk credit tier. The predictable effect of adding this credit tier dummy would be to reduce the size of the African American borrower coefficient. However, if the regression revealed that the lenders were charging customers with this attribute 5 percentage points more, then it would be necessary to omit the unjustifiable portion (here, 3 percentage points) of the credit tier effect.

---

[6]Lloyd Cohen and Melisa Michelsen (1996), however, have made an ingenious argument as to why these preferences might serve a more dynamic conception of efficiency (see also Epstein 1997). As an empirical matter, however, these preferences almost certainly involve a sacrifice in medical efficacy (Ayres 2002, 218).

In the transplantation context, my prior empirical work suggests that, if you were going to run an omitted variable test of disparate impact in probability of matching particular types of donated organs, then it would be important to exclude controls for certain partial antigen matches that are not empirically justified by leading to better transplant survivability. Furthermore, for fuller (five or six) antigen matches it would be appropriate to include controls, as this degree of matching is associated with higher survivability. Depending on whether the law or one's private norms require a tradeoff or "accommodation" of equity with efficiency, it might also be necessary to cap the maximum amount that the coefficients on these variables could take.

This omitted variable approach also can be used to test whether individual parts of a decision-making procedure produce unjustified disparate impacts. For example, imagine that it is determined that partial antigen matching is not a justified transplant criterion. It would be possible to run two probit regressions—both with and without controls for whether there was a partial antigen match—attempting to predict the probability that a recipient will qualify for transplantation. If excluding the (invalid) partial antigen control produced a statistical increase in the race coefficient, this would be evidence that this component of the decision making caused an unjustified disparate racial impact.

Indeed, under a *Connecticut v. Teal* (1982) type of analysis, it might be possible to use these dual regressions to show that a specific component of the decision-making process produces an unjustified disparate impact against minorities, even if minorities overall are not burdened by the process. For example, in the foregoing hypothetical probit regressions estimating transplantation probabilities, one could imagine that the coefficient on the minority indicator variable was estimated to be positive in both the regressions including and excluding controls for partial antigen matching, thus indicating that minority applicants had a heightened probability of qualifying for transplantation. Nonetheless, if the regressions indicated that excluding the partial antigen control caused a statistically significant drop in the minority coefficient (but still left the coefficient positive), this would be evidence that the partial antigen matching preference had an unjustified disparate impact on minority applicants.

While the omitted variable test at first blush provides a more unitary test of whether a decision maker's policies produce an unjustified disparate impact, it still relies on a separate determination of which factors are justified and how much of a justification they provide. Because it is only appropriate to partially include (and hence control for) variables that would provide a valid business justification, it is essential to have an independent theory of what types of factors might constitute a valid justification. This will require an analysis quite similar to prong two of the traditional test.

The Outcome Test

Outcome tests can provide powerful evidence of when a particular kind of decision making has an unjustified disparate impact. Outcome tests can produce a single statistic indicating both traditional elements of a disparate impact case—that decision making disproportionately affects minorities and that this disproportionality is not justified by heightened institutional productivity. Moreover, as discussed below, the outcome tests (while having some important limitations of their own) are again not susceptible to the traditional omitted variable bias concern.

The basic idea of the outcome test is to analyze whether the outcomes (about which the decision maker cares) are systematically different for minorities and nonminorities. If we find that in distributing benefits the decision maker effectively demands better outcomes from minorities than from whites, we may infer that there was a class of minorities that might have received benefits and produced the same quality of outcomes for the decision maker. Thus if we find that:

1. Lending decisions produce higher profits on loans to minorities than to whites, we might infer that the lending decisions have an unjustified disparate impact in excluding qualified minority borrowers;[7]
2. Bail bond setting decisions produce higher appearance rates for minorities than for whites, we might infer that bond setting decisions have an unjustified disparate impact on minority defendants (Ayres 2002, chap. 7);
3. Editorial acceptance decisions produce higher citation rates for articles written by minorities than by whites, we might infer that acceptance decision have an unjustified disparate impact in excluding qualified minority articles (Ayres and Vars 2000; Smart and Waldfogel 1996); and
4. Hiring decisions produce higher productivity for minority workers than for white workers, we might infer that hiring decisions have an unjustified disparate impact in excluding qualified minority workers (Gwartney and Haworth 1974, 876; Williams and Chambless 1998, 509).

Outcome tests can also be effective in analyzing a decision maker's allocation of detriments. If we find that in distributing a detriment the decision maker effectively accepts poorer outcomes from minorities than from whites, we may infer that there was a class of minorities that might have avoided the detriment. For example, if we find that police search decisions are systematically less productive with regard to minorities than with regard to whites, we might infer that search decisions have an unjustified disparate impact in subjecting undeserving minorities to being searched (see Knowles, Persico, and Todd 2001).

[7]In a *Business Week* op-ed (1993a) and his Nobel Prize lecture (1993b), Becker suggested that if banks discriminate against minorities we should expect that minorities would have lower default rates.

As applied to transplantation, the natural outcome test would be to assess whether transplants made to minority recipients survived longer than those that were transplanted to non-minority recipients. As a first cut, evidence that minorities given transplants survived longer would raise concerns that the allocation rules at a minimum had an unjustified disparate impact in excluding minority recipients. But as we shall see, making this type of inference turns out to raise a number of difficult issues.

A major advantage of these outcome tests is that they are not susceptible to the omitted variable bias critique that has plagued traditional regression-based tests of disparate treatment. Researchers do not need to observe and control for all the variables that the transplant officials considered in deciding whether to transplant as long as they can observe the relevant outcome. The outcome tests are not embarrassed by omitted variable bias, because under the null hypothesis there should be no observable variables that systematically affect the probability of success once the decision maker has made an individualized assessment so as to equalize this very probability. Indeed, perversely, the outcome test intentionally harnesses omitted variable bias to test whether any excluded (unjustified) determinant of decision making is sufficiently correlated with the included racial characteristics to produce evidence of a statistically significant racial disparity.[8] Any finding that transplant recipients with a particular characteristic (such as minority status) induce a systematically higher probability of transplant survivability suggests that transplant criteria unjustifiably subject that class of individuals to the disability of being excluded from transplantation.

This omitted variable point can be restated in more legalistic terms. The outcome test is not susceptible to the "qualified pool" problem that plagues both traditional disparate impact and disparate treatment issues of proof. In an outcome test, the decision maker herself defines what she thinks the qualified pool is, and the outcome test then directly assesses whether the minorities and non-minorities so chosen are in fact equally qualified. A finding that chosen minorities produce better outcomes than chosen whites suggests that the decision maker unfairly excluded some qualified minorities from benefits (or subjected them to unjustified detriments). As applied to transplantation, a finding that the transplantation survivability is systematically higher for minority recipients than for white recipients suggests that some number of minorities deserved more (that is, were better "qualified") to receive transplants. A defense that transplant decisions were driven by the underlying science of survivability—and that minorities were excluded from transplantation because they tend to have lower expectations of survivability—would be contradicted by systematically higher success rates when minority transplants were in fact completed.

But while the outcome test methodology has important strengths, it has lim-

[8]Stephen Ross and John Yinger (1999, 112) have noted that the default approach attempts to identify mortgage discrimination by purposely omitting variables from the regression.

itations as well. First and foremost it is a test of whether decision making criteria have an unjustified disparate impact. While such evidence can be quite probative of disparate treatment, there are ways that the outcome test can be both under- and over-inclusive as a test of disparate treatment.

Outcome tests can be under-inclusive as tests of disparate treatment because they are not well structured to capture disparate racial treatment motivated by rational statistical inference—so-called statistical discrimination. In his Nobel Prize lecture, Gary Becker (1993b) rather bizarrely extols outcome tests as being the "direct" approach to measuring discrimination. His definition of "discrimination," however, does not capture all race-contingent decision making. Analyzing bank lending, Becker concludes, "If banks discriminate against minority applicants, they should earn greater profits on the loans actually made to them than on those to whites" (389). But this is only true if the discrimination is caused by associational animus. The outcome test may produce no racial difference even in the shadow of disparate treatment, if instead the discrimination is caused by statistical inference.

More importantly, for these purposes, the outcome test may produce racial differences even if there is no disparate treatment but instead merely a race-neutral policy that produces an unjustified disparate racial impact. Instead of thinking of the outcome test as a direct test of disparate treatment, it is better to think of it as an indirect test of disparate impact.

Certain forms of the outcome test may also be over-inclusive as a test of disparate impact, particularly with regard to what I will call problems of "infra-marginality" and "subgroup validity."

The infra-marginality problem. A potential problem with outcome assessments as tests of disparate impact arises if researchers are only able to measure the average outcome and not the outcomes associated with the marginal decision. In the mortgage context, a test of disparate treatment would ask whether the least qualified whites to which banks were willing to lend had a higher default rate than the least qualified minorities to which banks were willing to lend. If lenders dislike lending to minorities, then the least qualified minority to which they would be willing to lend (the marginal minority borrower) should have a lower expected default rate than the least qualified nonminority to which they are willing to lend (the marginal nonminority borrower). Unfortunately, marginal default rates are unobservable, and researchers are often only able to estimate the average default rates conditional on being above this marginal lending threshold (Carr and Megbolugbe 1993, 309; Galster 1993). Lenders might still discriminate against minority borrowers—in the straightforward sense that the lending threshold for minorities might be more stringent than for nonminorities—but we might see that the average rate of minority default (conditional on being above the minority lending cutoff) is higher than the average rate of nonminority default (conditional on being above the nonminority lending cutoff). As long as infra-marginal nonminority borrowers have lower expected default rates than

infra-marginal minority borrowers, a comparison of average defaults may mask disparate treatment by lenders in setting the minimum thresholds for granting loans.

A similar infra-marginality problem could also limit the use of outcome analysis as a measure of disparate racial treatment in transplant decisions. As discussed above, a finding that minority transplant recipients have a systematically higher success rate raises concerns that a transplant system unjustifiably excludes potential minority recipients from the transplant process. But observing that the average search success rate for minorities was higher than for whites does not necessarily prove that the threshold (or marginal) expected success rate was higher for minorities than for whites. Disparate treatment tests are normally tests of decision making on the margin, but real-world data at times only allows researchers to assess infra-marginal effects.

This problem of infra-marginality does not, however, undermine all outcome tests equally. If either the decision or the outcome is non-dichotomous, it may become easier for the researcher to identify the marginal effects. For example, in a bail bond setting context, the fact that judges were setting continuous (non-dichotomous) bail amounts allowed Joel Waldfogel and me to directly test the marginal impact of their decisions (Ayres 2002, chap. 5; Ayres and Waldfogel 1994). The judges' ability to individually vary the bail amount in a sense makes every defendant marginal—and thus avoids the infra-marginal problem that has plagued the application of outcome tests to the mortgage context (where lenders make a much more dichotomous decision about whether to lend or not).

Similarly, if the outcome itself is non-dichotomous, it may be easier to identify whether the threshold decision making is discriminatory. Thus, for example, in the citation studies mentioned above, researchers, by measuring the number of citations given to articles written by minorities and nonminorities, can assess not just the average level of success with regard to a dichotomous outcome variable (such as nondefault on a loan) but also the entire distribution of success. By analyzing this distribution, it may be possible to identify whether the editors systematically demand more or fewer expected citations in accepting the marginal (least likely cited) articles of minority authors. This last point is especially relevant for the transplantation context, where the key outcomes of success related to survivability are non-dichotomous. It may therefore be possible to ascertain whether the distribution of expected survivability is censored from below at a different point for minority and nonminority recipients.

While the infra-marginality problem can limit the usefulness of outcome analysis as a test of disparate treatment, infra-marginality is not as much of a problem when interpreting the outcome analysis merely as a test of unjustified disparate impact. For example, imagine researchers find that the transplant of the average white recipient survives for three years, while the transplant of the average minority recipient survives for five years. Transplant officials could raise infra-marginality as an alternative to the hypothesis that this finding proves dis-

parate treatment: for example, they might argue that they transplant into all people who have at least a 2.5-year expected survivability (and of this group, it just so happens that minority recipients have a longer average survivability). In essence, the transplant officials would be arguing that they apply a uniform (2.5-year) threshold to all potential recipients, regardless of race, so that at the margin there is no disparate treatment.

But this would not be a defense to the claim that transplant criteria impose a disparate impact on minorities. The finding of an average racial disparity in survivability means that there exists some higher uniform survivability threshold (between 2.5 and 5 years) that would have included a higher proportion of minorities in the transplant pool. Or, in other words, a finding that white transplants are systematically less successful than minority transplants suggests that choosing a low uniform threshold had a disparate impact on the proportion of minorities receiving transplants.

But while a finding of disparity in the average search success rates would be evidence of a disparate impact, it might—taking into account the infra-marginality—no longer imply evidence of an unjustified disparate impact. In the previous example, as long as there were sufficient organs to transplant all of the high-survival recipients, it might have been justified to transplant the additional organs to lower-survival recipients (who turn out to be disproportionately white). Because the number of organs transplanted is determined by the supply, it might be justifiable to choose the survivability threshold that meets the supply. So ultimately, outcome analysis can provide strong evidence of a disparate racial impact, but whether the impact is justified or not may turn on whether evidence of racial disparities in the average outcome is evidence of racial differences in the threshold (or marginal) decision making.

The subgroup validity problem. A second limitation on the use of outcome tests as evidence of disparate racial treatment concerns what I term the subgroup validity problem. Put simply, when a particular observable characteristic is valid for some races but not for others, it is possible that a decision maker conditioning her decisions on this characteristic might induce racially disparate outcomes. To put the matter provocatively, when a particular observable characteristic is only a valid proxy of desert for some races, then a decision maker's *unwillingness* to engage in disparate racial treatment may induce just the racial disparities in outcomes that are generally a concern.

For example, imagine that wearing a particular type of baseball cap is strong evidence of drug possession when done by whites but not when done by minorities. In the extreme, imagine that 100% of whites wearing this cap possess drugs, and 0% of minorities wearing this cap possess drugs. Finally, imagine that if the police stopped all people wearing such a baseball cap, 75% of those stopped would be white (possessing illicit drugs) and 25% would be minorities (not possessing illicit drugs). These stylized examples suggest that the baseball cap is a valid indicator of illicit activity for whites, but it is not valid for the minor-

ity subgroup. Moreover, because 75% of the baseball cap wearers are white, we might claim that the characteristic is valid overall for the entire population—after all there is a 75% chance that a cap search will uncover illicit drugs.

Under these stylized facts, what is a police department likely to choose as its search criteria? In today's politically charged environment, the department might want to avoid just searching whites wearing the cap, fearing that such decision making would constitute illegal racial profiling. As an alternative, it might choose to stop all those who wear the cap (minorities and nonminorities alike). However, the result of such a color-blind criterion would be systematically poorer outcomes for minority searches than for white searches. While I argued above that lower search success rates for minorities might be indicative of the most blatant type of police attempts at racial harassment, in this hypothetical the systematically lower minority search success rate is caused by the police department's unwillingness to engage in disparate racial treatment, that is, its unwillingness to engage in racial profiling. This cap hypothetical provides a cautionary tale for over-defining what constitutes racial profiling. The outcome test still can provide strong evidence that the criteria for minority searches are less valid than the criteria for nonminority searches—and hence might still show that police demand less probable cause when searching minorities than whites. But the cap hypothetical vividly illustrates that an unwillingness to engage in disparate treatment can itself have a disparate impact that is unjustified (when judged from the perspective of subgroup validity).

However, a showing that particular decision-making criteria are systematically less valid for minorities might not be sufficient to make out a case that the disparate impact was unjustified. It is far from clear whether disparate impact law does (or should) require a showing that particular criteria are valid for racial subgroups.[9] In the foregoing baseball cap example, police might succeed in arguing that the search criteria imposed at worst a *justified* disparate impact because 75%

[9]The 1966 and 1970 EEOC guidelines required evidence of subgroup racial validity (so-called "differential validation"), thus requiring employers to conduct separate validation studies for different racial groups (e.g., *United States v. City of Chicago* 1977, 433). The Supreme Court even endorsed this in *Albermarle Paper Co. v. Moody* (1975, 435). But the Uniform Guidelines eliminated the requirement of differential validation and replaced it with something called "unfairness studies" (Code of Federal Regulations 1978, title 29, sec. 1607.14B(8)). Subgroup validation is still required with language (Kelman 1991, 1192). As Christine Jolls (2000) has recently noted, a disparate racial impact decision invalidating an employer's "no beard" policy (as having an unjustified disparate impact on African Americans) has expressly endorsed race-contingent remedies: in *Bradley v. Pizzaco of Nebraska* (1993, 799), the federal court ordered that "[t]he injunction shall be carefully tailored to place Domino's under the minimal burden of recognizing a limited exception to its no-beard policy for African American males who suffer from PFB and as a result of this medical condition are unable to shave." Such decisions suggest that decision makers may have a duty to remedy racially disparate impacts by resorting to express racially disparate treatment. However, such a duty may run afoul of the 1991 Civil Rights Act's ban on race norming (U.S. Code 42 [1992], §2000e-2(1)).

of all cap searches uncovered illicit contraband. The foregoing analysis suggests then that the outcome test assesses the relative validity of decision criteria with respect to each subgroup and not with respect to the full sample of people being searched.

Interestingly, however, in the transplant context, there may be more willingness to engage in certain types of express disparate treatment. The clear evidence of racially different antigen distributions and, more generally, racially different immunosuppression systems, provides a genetic basis for at least racially different treatments. Indeed, my earlier research showed that African American kidney recipients fared far better under a particular "quad" therapy of immunosuppressant drugs than white recipients (Ayres 2002, 198). The relative acceptability of "separate but equal" treatment approaches may help resolve the disparate treatment/disparate impact dilemma posed by the foregoing baseball cap example.

Finally, it is important to emphasize an important tension between the traditional test and the outcome test that is implicitly raised by my previous work on kidney transplantation. My earlier application of the traditional test provided strong evidence that the antigen-matching point system produced an unjustified disparate racial impact. But an outcome analysis of the kidney recipients would probably have suggested just the opposite. Counter to the outcome hypotheticals posed above, the transplants of African American kidney recipients do not survive longer than those of white recipients (Ayres 2002, 191). How is it possible that these two tests of disparate impact could point in different directions?

One reason for the difference is that the outcome test by necessity tests the transplantation system in toto, while the traditional test can carve out and analyze the potential disparate impacts of particular policies. Therefore, if the unjustified disparate racial impact of antigen matching is counterbalanced by other aspects of the system, then the net effect may be no (or in this case, an inverted) disparate impact. The outcome test therefore empirically overrules the Supreme Court's finding in *Connecticut v. Teal* (1982) that unjustified disparate impacts in any part of the decision-making process are actionable even if the overall result does not produce unjustified racial disparities. The outcome test cannot isolate the impact of individual components of the decision-making process.

Conclusion

The take-home lesson of this article is that distinct empirical methods must be used to test for disparate treatment and unjustified disparate impact. In particular, econometricians must overcome their long-bred embarrassment about omitted variable bias when running disparate impact regressions. What would be a biased coefficient in a disparate treatment regression is often precisely the effect that is being tested. Adding right-hand side variables willy-nilly needs to be avoided. When testing for disparate impacts, one often needs to be more concerned about "included variable" bias than "omitted variable" bias.

This article has looked briefly at three different types of disparate treatment tests—all of which in one way or another intentionally omit variables. The omitted variable and the outcome tests both have important strengths in comparison to traditional auditing tests of disparate impact. Most importantly, these new tests avoid the included variable bias concern. But these new tests also have limitations. The omitted variable test still requires an independent analysis of what variables are justified. And the outcome tests may be over-inclusive because of problems of infra-marginality or subgroup validity. Because there are in particular contexts adequate responses to each of these problems, these two new types of tests should be part of the accepted arsenal of civil rights empiricism. These new tests can provide credible evidence especially when combined with other (more traditional) types of evidence that decision making subjects minorities to an unjustified disparate impact. The search for definitive tests will, however, ultimately depend on particularized theories of what constitutes a justified disparity.

Finally, a complete empirical analysis of disparate impact would also need to consider whether less restrictive policies exist—policies that accomplish legitimate interests while producing a less disparate racial impact. This final type of analysis might be undertaken as a Paretian exercise, in which one searches for alternative policies that produce smaller racial disparities with absolutely no lessening of the decision maker's nonracial objectives. Or it might be done as more of an accomodationist exercise where researchers would investigate how much of a reduction in disparity could be accomplished by marginally sacrificing what would otherwise be a legitimate interest of the decision maker.

REFERENCES

Albermarle Paper Co. v. Moody, 422 U.S. 405 (1975).

Ayres, I. 2002. *Pervasive prejudice? Non-traditional evidence of race and gender discrimination*. Chicago: Univ. of Chicago Press.

Ayres, I., L. G. Dooley, and R. S. Gaston. 1993. Unequal racial access to kidney transplantation. *Vand Law Rev* 46:805–61.

Ayres, I., and P. Siegelman. 1996. The q-word as red herring: Why disparate impact liability does not induce hiring quotas. *Texas Law Rev* 74:1487–1526.

Ayres, I., and F. E. Vars. 2000. Determinants of citations to articles in elite law reviews. *J Legal Stud* 29:427–50.

Ayres, I., and J. Waldfogel. 1994. A market test for race discrimination in bail setting. *Stanford Law Rev* 46:987–1046.

Becker, G. S. 1993a. The evidence against banks does not prove bias. *Business Week*, April 19.

Becker, G. S. 1993b. The economic way of looking at behavior. *J Pol Econ* 101(3):385–409.

Bradley v. Pizzaco of Nebraska, Inc., 7 F.3d 795 (8th Cir. 1993)

Carr, J. H., and I. F. Megbolugbe. 1993. The Federal Reserve Bank of Boston study on mortgage lending revisited. *J Housing Res* 4(2):277–313.

Clogg, C. C., and A. Haritou. 1997. The regression method of causal inference and the dilemma confronting this method. In *Causality in crisis? Statistical methods and the search for causal knowledge in the social sciences*, ed. V. R. McKim and S. P. Turner, 83–112. Notre Dame: Univ. of Notre Dame Press.

Code of Federal Regulations. 1978. Uniform guidelines on employee selection procedures. Title 29, sec. 1607.

Cohen, L. R., and M. Michelsen. 1996. The efficiency/equity puzzle and the race issue in kidney allocation: A reply to Ayres, et al. and UNOS. *Ann Rev Law Ethics* 4:137–85.

Connecticut v. Teal, 457 U.S. 440 (1982).

Epstein, R. A. 1997. *Mortal peril: Our inalienable right to health care?* New York: Perseus Press.

Galster, G. C. 1993. The facts of lending discrimination cannot be argued away by examining default rates. *Housing Policy Debate* 4(1):141–46.

Gaston, R. S., et al. 1993. Racial equity in renal transplantation: The disparate impact of HLA-based allocation. *JAMA* 270(11):1352–56.

Gaston, R. S., et al. 1994. Race and allocation of kidneys for transplantation. *JAMA* 271 (4):269–71.

Gaston, R. S., et al. 1995. HLA matching in renal transplantation. *N Eng J Med* 332(11): 752–53.

Greenstein, S. M., et al. 1989. Does kidney distribution based upon HLA matching discriminate against blacks? *Transplant Proc* 21(6):3874–75.

Griggs v. Duke Power Co., 401 U.S. 424 (1971).

Grutter v. Bollinger, 123 S. Ct. 2325 (2003).

Gwartney, J., and C. Haworth. 1974. Employer costs and discrimination: The case of baseball. *J Pol Econ* 82(4):873–81.

HR-Guide. 2001. HR guide to the Internet. EEO: Disparate impact. http://www.hr-guide.com/data/G702.htm.

HR-Software. 1998. Disparate impact analysis (an on-line Internet-based application). http://www.hr-software.net/EmploymentStatistics/DisparateImpact.htm.

Ireland, T. R. et al. 1998. *Expert economic testimony: Reference guide for judges and attorneys*. Tucson: Lawyers and Judges Publishing.

Jolls, C. 2000. Accommodation mandates. Unpublished manuscript.

Kelman, M. 1991. Concepts of discrimination in general ability job testing. *Harvard Law Rev* 104:1157–1247.

Knowles, J., N. Persico, and P. Todd. 2001. Racial bias in motor vehicle searches: Theory and evidence. *J Pol Econ* 109(1):203–29.

Lazda, V. A., and M. E. Blaesing. 1989. Is allocation of kidneys on basis of HLA match equitable in multiracial populations? *Transplant Proc* 21(1):1415–16.

Ross, S. L., and J. Yinger. 1999. The default approach to studying mortgage discrimination: A rebuttal. In *Mortgage lending discrimination: A review of existing evidence*, ed. M. A. Turner and F. Skidmore, 107–36. Washington, D.C.: Urban Institute.

Smart, S., and J. Waldfogel. 1996. A citation-based test for discrimination at economic and finance journals. NBER Working Paper 5460.

U.S. Code 42 (1994), §2000e-2.

United Network for Organ Sharing (UNOS). 1992. Policy §3.5.2.

United States v. City of Chicago, 549 F.2d 415 (7th Cir. 1977).

Williams, J. F., and J. A. Chambless. 1998. Title VII and the Reserve Clause: A statistical analysis of salary discrimination in major league baseball. *U Miami Law Rev* 52:461–527.

Yinger, J. 1998. Evidence on discrimination in consumer markets. *J Econ Perspect* 12(2): 23–40.