

# ❏ OUTCOME TESTS OF RACIAL DISPARITIES IN POLICE PRACTICES

Ian Ayres  
Yale University Law School

## ❏ Abstract

This article assesses the strengths and weaknesses of using “outcome tests” to assess racial disparities in police practices. An outcome test, for example, might assess whether the probability of finding contraband was higher for whites who are searched than for minorities who are searched.

---

*This article is based on a paper prepared for the National Research Council, Division of Behavioral and Social Sciences and Education, Committee on Law and Justice, Meeting of the Committee to Review Research on Police Policy and Practices, Washington, DC, April 11, 2002. The paper was based on a discussion in Chapter 9 of Ian Ayres, Pervasive Prejudice? Unconventional Evidence of Race and Gender Discrimination (University of Chicago Press, 2001).*

Outcome tests can provide powerful evidence of when a particular kind of decisionmaking has an unjustified disparate impact. Outcome tests can produce a single statistic indicating both traditional elements of a disparate impact case—that decisionmaking disproportionately affects minorities and that this disproportionality is not justified by heightened institutional productivity. Moreover, as discussed below, the outcome tests (while having some limitations of their own) are 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 decisionmaker cares) are systematically different for minorities and nonminorities. If we find that in distributing benefits the decisionmaker 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 decisionmaker. 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 lenders;<sup>1</sup>
- (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;<sup>2</sup>
- (3) editorial acceptance decisions produce higher citation rates for articles written by minorities than by whites, we might infer that acceptance decisions have an unjustified disparate impact in excluding qualified minority articles (see Smart, Shoven, & Waldfogel, 1996 and Ayres & Vars, 2000, p. 427); 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 (see Gwartney & Hayworth, 1974, pp. 873, 876 and Williams, 1998, pp. 461, 509).

Outcome tests can also be effective in analyzing a decisionmaker's allocation of detriments. If we find that in distributing a detriment the decisionmaker 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.

---

<sup>1</sup> Becker suggested that if banks discriminate against minorities we should expect that minorities would have lower default rates. Becker first proposed this approach in a *Business Week* op-ed (see Becker, 1993a), and detailed his suggestion in his Nobel Prize lecture (see Becker, 1993b).

<sup>2</sup>This type of test was the basis of Chapter 7 of my book *Pervasive Prejudice? Unconventional Evidence of Race and Gender Discrimination* (2001).

For example, if we find

- (5) 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 (For an example of this type of testing, see Knowles, Persico, & Todd, 2000).

Because outcome testing is an especially useful tool in assessing allegations of racial profiling by police, the following discussion will focus on “police search” outcome tests as a primary example to illuminate the strengths and weaknesses of this methodology.

### Police Search Outcomes Tests

The ex post probability that a police search will uncover contraband or evidence of illegality is strong evidence of the average level of probable cause that police require before undertaking a search. A finding that minority searches are systematically less productive than white searches is accordingly evidence that police require less probable cause when searching minorities. To be sure, such a finding does not require that we infer that police engaged in disparate treatment—but, at a minimum, it is evidence that whatever criteria the police employed produced an unjustified disparate impact.<sup>3</sup> Such evidence would suggest that if police required the same level of probable cause when searching minorities as when searching whites, there would be fewer minorities searched (or proportionally more whites searched).

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 don’t need to observe and control for all of the variables that police considered in deciding whether to search as long as they can observe the outcome of their decisionmaking. 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 decisionmaker has made an individualized assessment so as to equalize this very probability. Indeed, perversely, the outcome test intentionally

---

<sup>3</sup> Evidence of an unjustified disparate impact can be used as evidence of intentional discrimination (disparate treatment) and under current law unjustified disparate racial impacts of police action can be challenged under federal law. See 28 C.F.R. §42.203(3) implementing 42 U.S.C. §3798d(c).

harnesses omitted variable bias to test whether any excluded (unjustified) determinant of decisionmaking is sufficiently correlated with the included racial characteristics to produce evidence of a statistically significant racial disparity.<sup>4</sup> Any finding that the police searches of individuals with a particular characteristic (such as minority status) induce a systematically lower probability of uncovering illegality suggests that police search criteria unjustifiably subject that class of individuals to the disability of being searched.

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. The decisionmaker in an outcome test by her own decisions defines what she thinks the qualified pool is, and the outcome test then directly assesses whether the minorities and nonminorities so chosen are in fact equally qualified. A finding that chosen minorities produce better outcomes than chosen whites suggests that the decisionmaker unfairly excluded some qualified minorities from benefits (or subjected them to unjustified detriments). As applied to police searches, a finding that the search success rate (i.e., the probability of finding evidence of illegality) is systematically lower for searched minorities than for searched whites suggests that minorities less deserved (that is, were less “qualified”) to be searched. A defense that police searching decisions were driven by the underlying criminality of those searched—and that minorities make up a larger proportion of those deserving to be searched—would be contradicted by systematically lower success rates when such searches were in fact completed.

But while the outcome test methodology has important strengths, it has limitations as well. First and foremost, the test is primarily a test of whether decisionmaking 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 overinclusive as a test of disparate treatment.

Outcome tests can be underinclusive 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 extols outcome tests as being the “direct” approach to measuring discrimination. His definition of “discrimination,” however, does not capture all race-contingent decisionmaking. Analyzing bank lending, Becker concludes: “If banks discriminate against minority applicants, they should earn *greater*

---

<sup>4</sup> Stephen Ross and John Yinger (1999, pp. 107, 112) have noted that the default approach attempts to identify mortgage discrimination by purposely omitting variables from the regression.

profits on the loans actually made to them than on those to whites” (Becker, 1993b, pp. 385, 389) But this is only true if the discrimination is caused by associational animus and is not necessarily true if instead the discrimination is caused by statistical inference.

As applied to police searching criteria, an outcome test would likely capture efforts by police to arbitrarily target and harass a minority population. But it might not capture express racial profiling that was based on valid statistical inference. For example, if police were correct in inferring among some group of otherwise observationally equivalent suspects that minority suspects had a higher likelihood than whites of possessing contraband and used the race expressly as a part of their criteria for searching, then in equilibrium we might not observe lower search success rates for minorities than for whites. Even though the police were engaging in a type of disparate racial treatment—express racial profiling—the outcome test might show no racially disparate outcomes.

If underinclusion were the only problem, an outcome test might still provide a valuable “one-tailed” test of the existence of disparate treatment. But certain forms of the outcome test may also be overinclusive as a test of disparate treatment—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 treatment 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 want to ask whether the least qualified whites to whom banks were willing to lend had a higher default rate than the least qualified minorities to whom banks were willing to lend. If lenders dislike lending to minorities, then the least qualified minority to whom they would be willing to lend (the marginal minority borrower) should have a lower expected default rate than the least qualified nonminority to whom 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 & Megbolugbe, 1993, pp. 277, 309; see also Galster, 1993, p. 141). Lenders might still discriminate against minority lenders—in the straightforward sense that the lending threshold for minorities might be more stringent than for nonminorities—but we might still 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 police search decisions. As discussed above, a finding that police have systematically lower success rates when searching minorities than when searching whites might raise concerns that police were using race-contingent search thresholds. But observing that the average search success rate for minorities was lower than for whites does not necessarily prove that the threshold (or marginal) expected success rate was lower for minorities than for whites. Disparate treatment tests are normally tests of decisionmaking on the margin, but real world data at times only allows researchers to assess infra-marginal effects.<sup>5</sup>

This problem of infra-marginality does not, however, equally undermine all outcome tests. If either the decision or the outcome is nondichotomous, it may become easier for the researcher to identify the marginal effects. For example, in the bail bond-setting context, the fact that judges were setting continuous (nondichotomous) bail amounts allowed us to directly test the marginal impact of their decisions. 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 nondichotomous, it may be easier to identify whether the threshold decisionmaking is discriminatory. Thus, for example, in the citation studies mentioned above, researchers in measuring the number of citations given to articles written by minorities and nonminorities can assess not just the average level of success (as with a dichotomous outcome variable, such as non-default on a loan) but can estimate the entire distribution of success. By analyzing this distribution, it may be possible to identify whether the editors in making acceptance decisions systematically demand more or fewer expected citations in accepting the marginal (least likely cited) articles of minority authors.

---

<sup>5</sup> It should be emphasized that this infra-marginality problem can also cause outcome analysis to be underinclusive as a test of disparate treatment. For example, in the policing context, we might as a theoretical matter observe minority search success rates to be higher on average even though police require less probable cause on the margin to search when searching minorities. As argued before, however, underinclusiveness would not undermine the test's use as a one-sided test for disparate treatment.

The most daunting problem for researching the use of outcome tests concerns those contexts in which both the decision and outcome are dichotomous. This might well describe both mortgage lending (where the bank decides lend/not lend, and the outcome is default/nondefault) and police search (where the police decide search/not search and the outcome is contraband found/not found). In such circumstances, the basic structure of the data does not allow the researcher to go beyond observing averages. But even in these contexts, I believe that outcome tests can still be of use for two reasons.<sup>6</sup>

First, in some contexts evidence of racial disparities in the average outcome is strong evidence of disparities on the margin. Here it is useful to contrast mortgage lending and police searches. In mortgage context, evidence about average defaults may not provide strong evidence of marginal expected default rate. Because white borrowers, as an empirical matter, are likely to control more wealth, it should not be surprising if whites are disproportionately infra-marginal borrowers (with low default rate). Observing lower average white default rates should accordingly not give us confidence that the expected default rate of the marginal white is less than that of the marginal minority borrower.

In contrast, in the police context, it is harder to articulate why the average search success rates would not be a credible proxy for the marginal success rate. If researchers found that the average white search success rate was systematically higher than that for minorities, it would be difficult to explain why whites were more likely to be infra-marginal searchees. To argue that this finding was not evidence of discrimination, the police would need to say that they searched every one with a minimum probable cause but that of those meeting this standard whites for some reason had a systematically higher chance of possessing evidence of illegality. The difficulty is articulating a particular reason why the infra-marginal white would have a higher probability of possessing contraband even though the marginal white did not.<sup>7</sup> The core issue is whether evidence of average (infra-marginal) racial outcome differences is probative of marginal (or

---

<sup>6</sup> John Knowles, Nicola Persico and Petra Todd (2000) have recently suggested a third reason. The strategic actions of those subject to a decision may systematically move the average success rate toward the marginal success rate, thus making the average a better proxy.

<sup>7</sup> Disproportionate minority recidivism might begin to provide such a theory. If recidivists are more skilled in secreting contraband and if recidivists are disproportionately minorities, then we might think of nonrecidivists as being the infra-marginal searchees (whose search would produce a higher probability of uncovering contraband). If nonrecidivists are disproportionately white, then the average search success rates of whites might be higher than that of minorities.

threshold) outcome differences. When there are not compelling reasons to suspect the inframarginal effects to differ from the marginal—as I have suggested is much more the case with police searches than with mortgage lending—the outcome tests can still provide valuable information about the probable existence of discrimination.

Second, 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 average white search uncovers contraband 15% of the time, while the average minority search uncovers contraband only 10% of the time. The police could raise infra-marginality as a defense to the claim that this finding proves disparate treatment: for example, they might argue that they stop all people who display at least a 5% probability of having contraband (and of this group, it just so happens that 15% of whites have contraband while only 10% of minorities do). In essence, the police would be arguing that they apply a uniform (5%) threshold to all suspects regardless of race—so that at the margin there is no disparate treatment. But this would not be a defense to the claim that police search criteria impose a disparate impact on minorities. The finding of an average racial disparity must mean that there exists some higher uniform probable cause threshold (between 5% and 15%) that would have subjected disproportionately fewer minorities to search. Or, in other words, the finding that white searches are systematically more successful than minority searches suggest that choosing a low uniform threshold had a disparate impact of disproportionately exposing minorities to unsuccessful searches. 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. Under the previous hypothetical, as long as the uniform probable cause threshold (5%) was not unreasonably low, the police could argue that their searching criteria had a justified disparate impact. 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 outcomes is evidence of racial differences in the threshold (or marginal) decisionmaking.

### ☒ 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 decisionmaker conditioning her decisions on this characteristic generally 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 decisionmaker'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. And finally imagine that if the police stopped all people wearing such a baseball cap, that 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 minority 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.

But 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 decisionmaking 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 colorblind criterion would be that it would produce 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 of 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—its unwillingness to engage in racial profiling. This cap hypothetical provides a cautionary tale for overdefining what constitutes racial profiling. If, given our nation's painful history of minority oppression, we are more concerned with the possibility of invidious discrimination by police against minorities (than in favor of minorities), then we should more stringently scrutinize race-contingent decisionmaking in which minority status makes it *more* likely that a search will occur, than decisionmaking in which minority status makes it *less* likely that a search will occur. This is merely an implication of what legal scholars sometimes refer to as the “anti-subordination principle”—which argues that courts should more strictly scrutinize government actions

that burden a traditionally subordinated group than ones that burden a nonsubordinated group.<sup>8</sup>

The outcome test still can provide strong evidence (putting aside for the moment the infra-marginality problem) 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). As applied to police searches, a finding that the minority search success rate was systematically lower than that of whites would at a minimum indicate that the search criteria were less valid when applied to minorities than to whites.

However, such a showing (that particular decisionmaking 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.<sup>9</sup> In the foregoing cap hypothetical, police might succeed in arguing that the search criteria imposed at worst a *justified* disparate impact because 75% of all cap searches uncovered illicit contraband.

---

<sup>8</sup> Unfortunately, this principle has been explicitly rejected by the Supreme Court. See *Adarand Constructors, Inc. v. Peña*, 515 U.S. 200, 224 (1995) (explaining that “consistency” requires the same strict standard of review apply no matter “the race of those burdened or benefited by a particular classification” [quoting *City of Richmond v. J.A. Croson Co.*, 488 U.S. 469, 494 (1989) (plurality opinion)]).

<sup>9</sup> Vicki Schultz has informed me that the 1966 and 1970 EEOC guidelines required evidence of subgroup racial validity (so called “differential validation”) requiring employers to conduct separate validation studies for different racial groups. See *U.S. v. City of Chicago*, 549 F.2d 415, 433 (7th Cir.) (requiring differential validation). The Supreme Court even endorsed it in *Albermarle Paper Co. v. Moody*, 422 U.S. 405, 435 (1975). But the Uniform Guidelines eliminated the requirement for differential validation and replaced it with something called “unfairness studies.” See 29 CFR 1607.14B(8). Subgroup validation is still required with language; see Mark Kelman, (1991). And as Christine Jolls has recently noted (see Jolls, 2000), 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. *Bradley v. Pizzaco of Nebraska*, 7 F.3d 795, 799 (1993) (“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 decisionmakers may have a duty to remedy racial disparate impacts by resorting to express racial disparate treatment. However, such a duty may run afoul of the 1991 Civil Rights Act’s ban on race-norming. See 42 U.S.C. 2000e-2(l) (Supp. IV 1992).

The foregoing analysis suggests then that the outcome analysis tests relative subgroup validity of decision criteria and not whether the criteria are valid with respect to the full sample of people being searched. While this is an important theoretical concern, as applied to criminal context (where in many jurisdictions minorities comprise a majority of those who are searched) it is highly unlikely that police search criteria could be valid overall without being valid to minorities. For example, as applied to the outcome test of bail bond setting,<sup>10</sup> the finding that bail-setting criteria were not valid with regard to minorities strongly indicated that these criteria were not valid overall—for the simple reason that minorities made up over three quarters of those for whom bail was set. It is difficult to believe that criteria that are not valid with regard to the overwhelming majority of observations would nonetheless be valid with regard to the entire population.<sup>11</sup>

## ☒ Summary

“Outcome tests” have important strengths in comparison to traditional auditing tests of disparate treatment. Most importantly, these new tests avoid the recurrent “omitted variable bias” or “qualified pool” problems that plague attempts to show disparate treatment on the basis of traditional audits or with disparate impact evidence. But these new tests also have limitations. The outcome tests may be overinclusive because of problems of infra-marginality or subgroup validity. However, because there are in particular contexts adequate responses to each of these problems, this relatively new type of test deserves to be part of the accepted arsenal of civil rights empiricism. Outcome tests can provide credible evidence especially when combined with other (more traditional) types of evidence that decisionmaking subjects minorities to an unjustified disparate impact.

---

<sup>10</sup> See Chapter 7, Ayres (2001).

<sup>11</sup> However, in other contexts where minorities comprise only a small proportion of those subject to a particular type of decision, there remains a stronger possibility that criteria which were subgroup invalid might still be valid overall.

## References

- Adarand Constructors, Inc. v. Peña, 515 U.S. (1995).
- Ayres, I. (2001). *Pervasive prejudice? Unconventional evidence of race and gender discrimination*. Chicago: University of Chicago Press.
- Ayres, I., & Vars, F. E. (2000). Determinants of citations to articles in elite law review. *Journal of Legal Studies* 29, 427.
- Becker, G. S. (April 19, 1993a). The evidence against banks doesn't prove bias. *Business Week*. (Available online at [www.businessweek.com](http://www.businessweek.com))
- Becker, G. S. (1993b). Nobel lecture: The economic way of looking at behavior. *Journal of Political Economy*, 101, 385, 389.
- Carr, J. H., & Megbolugbe, I. F. (1993). The Federal Reserve Bank of Boston study on mortgage lending revisited. *Journal of Housing Research*, 4, 277, 309.
- Galster, G. C. (1993). The facts of lending discrimination cannot be argued away by examining default rates. *Housing Policy Debate*, 4, 141.
- Gwartney, J., & Hayworth, C. (1974). Employer costs and discrimination: The case of baseball. *Journal of Political Economy* 82, 873, 876.
- Knowles, J., Persico, N., & Todd, P. (2001) Racial bias in motor vehicle searches: Theory and evidence. *Journal of Political Economy*, 109, 1, 203-229.
- Ross, S., & Yinger, J. (1999). The default approach to studying mortgage discrimination: A rebuttal. In M. A. Turner & F. Skidmore (Eds.), *Mortgage lending discrimination: A review of existing evidence*.
- Smart, S., Shoven, J., & Waldfogel, J. (1996). A citation-based test for discrimination at economic and finance journals. Cambridge, MA: National Bureau of Economic Research Working Paper 5460.
- Williams, J. F. (1998). Title VII and the reserve clause: A statistical analysis of salary discrimination in major league baseball. *University of Miami Law Review*, 52, 461, 509.