Scholarship Comment

Why Affirmative Action Does Not Cause Black Students To Fail the Bar

Richard H. Sander, *A Systemic Analysis of Affirmative Action in American Law Schools*, 57 STAN. L. REV. 367 (2004).

In a widely discussed empirical study, Richard Sander concludes that affirmative action at U.S. law schools causes blacks to fail the bar.[1] If correct, this conclusion would turn the jurisprudence, policy, and law of affirmative action on its head.[2] But the article misapplies basic principles of causal inference, which enjoy virtually universal acceptance in the scientific community.[3] As a result, the study draws internally inconsistent and empirically invalid conclusions about the effects of affirmative action. Correcting the assumptions and testing the hypothesis directly shows that for similarly qualified black students, attending a higher-tier law school has no detectable effect on bar passage rates.

Part I clarifies the assumptions implicit in Sander's study and explains the inconsistent and indefensible premises on which it rests. Part II presents results from a reanalysis of the data, using alternative methods that correct and reduce the role of these unjustifiable assumptions. The reanalysis

* All analyses presented in this Comment are available at http://www.yalelawjournal.org.

1. Richard H. Sander, *A Systemic Analysis of Affirmative Action in American Law Schools*, 57 STAN. L. REV. 367, 447 (2004).

2. The article has already engendered a host of critical responses. *E.g.*, Ian Ayres & Richard Brooks, *Does Affirmative Action Reduce the Number of Black Lawyers?*, 57 STAN. L. REV. (forthcoming May 2005); David L. Chambers et al., *The Real Impact of Eliminating Affirmative Action in American Law Schools: An Empirical Critique of Richard Sander's Study*, 57 STAN. L. REV. (forthcoming May 2005); Michele Landis Dauber, *The Big Muddy*, 57 STAN. L. REV. (forthcoming May 2005). This Comment is the first, however, to point out the study's inferential flaws of post-treatment bias and extrapolation.

3. *See* Lee Epstein & Gary King, *The Rules of Inference*, 69 U. CHI. L. REV. 1, 34-37 (2002).

suggests that Sander's conclusions are untenable on their own terms.[4] Part III concludes.

I

At the outset, it is important to note that because all the schools in the LSAC Bar Passage Study on which Sander's analysis relies employ some system of affirmative action, no direct conclusion about the effects of affirmative action can be sustained.[5] While researchers in other areas have capitalized on variation in affirmative action rules to identify the effects of affirmative action, such variation does not exist here.[6]

Because there is no information in the data set with which to examine the direct causal effect of affirmative action, Sander is relegated to investigating a different quantity of interest: the causal effect of attending a higher-tier law school. While this is not a causal effect of affirmative action per se, it may be informative in assessing affirmative action's policy impact. For instance, if Sander is correct in claiming that similarly qualified students who go to higher-tier schools (1) are "mismatched" in terms of academic credentials, (2) learn less, and (3) are thus more likely to fail the bar,[7] then affirmative action might appear to hurt those it aims to help.

So how do we investigate the causal effect of attending a higher-tier school? Here, using nontechnical terms, I introduce the assumptions required to interpret Sander's findings as causal effects[8] and show that the study's assumptions are implausible and internally inconsistent.

Two basic tenets underlie any causal inference. The first is that causal inference is inherently counterfactual.[9] If we are interested in how Student *A*'s bar performance would be affected by attending a first-tier versus a second-tier law school, we would ideally observe *A* attending both schools. Yet if *A* attends a first-tier school, we cannot observe her in the counterfactual world where she attends a second-tier school. This

4. For a more extensive and technical presentation of this reanalysis, see Daniel E. Ho, Evaluating Affirmative Action in American Law Schools: Does Attending a Better Law School Cause Black Students To Fail the Bar? (Mar. 9, 2005) (unpublished manuscript), *available at* http://www.yalelawjournal.org *and* http://people.iq.harvard.edu/~dho/research/sander.pdf.

5. *See* Paul W. Holland & Donald B. Rubin, *On Lord's Paradox, in* PRINCIPALS OF MODERN PSYCHOLOGICAL MEASUREMENT: A FESTSCHRIFT FOR FREDERIC M. LORD 3, 9-14 (Howard Wainer & Samuel Messick eds., 1983) (discussing how the failure to define a control group resolves Lord's paradox).

6. *See* Harry Holzer & David Neumark, *Assessing Affirmative Action*, 38 J. ECON. LITERATURE (NO. 3) 483, 508 (2000).

7. Sander, *supra* note 1, at 449-54.

8. *See* Paul W. Holland, *Statistics and Causal Inference (with Discussion)*, 81 J. AM. STAT. ASS'N 945 (1986) (discussing the formal statistical model of causal inference).

9. *See* Epstein & King, *supra* note 3, at 34-37; Holland, *supra* note 8, at 945.

"fundamental problem of causal inference"[10] plagues even controlled, randomized laboratory experiments: If a unit is exposed to the treatment, we do not observe it under control.

The second tenet of causal inference is that we must at least be able to imagine conducting an experiment that manipulates a "treatment," or causal factor of interest. Laboratory scientists assess causal effects by actually conducting such experiments. For Sander's study, this would require randomly assigning a subset of students to tiers (the treatment) and observing differences in bar passage rates (the outcome). Randomization and a sufficiently large sample ensure that the students we are comparing across tiers are similar, such that different rates of bar passage can be attributed to the treatment. To estimate the average causal effect we can then simply calculate the difference in bar passage rates across tiers.

The problem for legal scholars and social scientists is that laboratory experiments are often infeasible, expensive, or unethical. Instead, to investigate causal effects researchers must resort to analyzing data in which there is no treatment randomization (so-called "observational data"). The hypothetical experiment discussed above nonetheless elucidates the key assumptions in standard methods (e.g., regression) used to infer causal effects from observational data. The goal of such methods is simply to get as close as possible to the hypothetical experiment by holding constant all other factors that affect the outcome but are present prior to the treatment.

I focus here on the key result in Sander's study of the causal effect of law school tier on bar passage.[11] The study attempts to explain the outcome of bar passage with a regression model that includes law school GPA, LSAT score, undergraduate GPA, gender, and race. Finding that law school grades have a stronger association with bar passage than law school tier does,[12] the study concludes that there exists a "trade-off between 'more eliteness' and 'higher performance.' . . . If one is at risk of not doing well academically at a particular school, one is better off attending a less elite school and getting decent grades."[13] The central claim is that going to a higher-tier law school causes less qualified students to learn less and earn lower grades, decreasing bar performance by more than higher school quality increases it. If this is so, our hypothetical experiment should reveal that students randomized into a higher tier have lower bar passage rates. The intuitive idea behind Sander's analysis is that if we hold constant all factors (variables) that a law school admissions committee observes (i.e.,

10. Holland, *supra* note 8, at 947.

11. Sander, *supra* note 1, at 444 tbl.6.1.

12. To be precise, Sander concludes this from the higher standardized coefficient on law school grades than on law school tier.

13. Sander, *supra* note 1, at 445.

that might affect bar passage but are present prior to admission to law school), then we can attribute the difference in bar passage to the difference in law school tier, thereby approximating our hypothetical experiment. So what are some of the key assumptions required for this to be true?

The first crucial assumption is that the variables we control for (undergraduate and law school GPA, LSAT score, gender, and race) are not themselves affected by the treatment of law school tier (i.e., they are "pretreatment variables"). Why? If we hold constant something that is itself affected by the treatment, then we are removing precisely one of the main effects we are trying to study.[14] So for example, when assessing the effect of smoking on death, we do not control for lung health, because this would remove one of the primary ways that smoking affects death.

Recall that Sander controls for law school grades. But Sander himself argues, correctly, that law school tier strongly affects law school grades.[15] Therefore, controlling for law school grades will never produce the right estimates of the effect of law school tier.[16] This is the first basic flaw in the Sander analysis. The "mismatch hypothesis" posits that admitting black students to a higher-tier school causes lower grades and reduces bar passage. Yet in estimating the causal effect on bar passage, the study, by its own account, should not control for law school grades. As Part II explains, removing this bias by excluding law school grades reveals that the aggregate impact of law school tier on bar passage within Sander's original framework is undetectable. Sander's ostensible decomposition of the law school tier effect into performance and eliteness effects thereby derives from a basic misinterpretation of the regression analysis.

The second crucial assumption is that there is no difference in students after holding constant undergraduate GPA, LSAT score, gender, and race, except for law school tier.[17] If true, this assumption permits the researcher to attribute any remaining differences in bar passage rates to law school

14. For a more formal discussion of this issue, see Paul R. Rosenbaum, *The Consequences of Adjustment for a Concomitant Variable That Has Been Affected by the Treatment*, 147 J. ROYAL STAT. SOC'Y SERIES A (GEN.) 656 (1984).

15. Sander, *supra* note 1, at 373 ("[R]acial preferences have the effect of . . . sharply lowering [black students'] average grades . . . .").

16. To see just how pathological this "post-treatment bias" can be, take the example of the causal effect of smoking on infant mortality. If smoking both causes low birth weight and increases the infant mortality rate, incorrectly controlling for birth weight may lead to an estimate that smoking actually saves lives. *See* Allen J. Wilcox, *Birthweight and Perinatal Mortality: The Effect of Maternal Smoking*, 137 AM. J. EPIDEMIOLOGY 1098 (1993) (illustrating post-treatment bias). This is the reverse of what statisticians call "Simpson's paradox," in which controlling for some variable can reverse aggregate proportions. *See* E.H. Simpson, *The Interpretation of Interaction in Contingency Tables*, 13 J. ROYAL STAT. SOC'Y SERIES B (METHODOLOGICAL) 238 (1951). Even if smoking had been randomly assigned in an experiment, controlling for birth weight would induce post-treatment bias, yielding the wrong conclusion.

17. To be more precise, this assumption means that we have controlled for all variables that are causally prior to tier, associated with tier, and affect bar passage conditional on tier.

tier. But the assumption is likely violated if, for example, students at top-tier schools have better letters of recommendation or have graduated from more prestigious undergraduate institutions, factors which Sander does not control for.[18] A key focus of the economics literature has been how these so-called "unobserved" factors might affect outcomes,[19] and for good reason. If there are unobservable differences in students across tiers, Sander's analysis and the approach presented below both fail: Differences in bar passage rates could be due to any number of unobserved factors, not just law school tier. My approach here also requires assuming the absence of unobserved factors. An extension of my analysis that controls for a wider range of variables, thereby making this assumption more believable, further indicates that there is no evidence for the Sander hypothesis.[20]

Sander's final key assumption concerns *how* the pretreatment variables affect the probability of bar passage.[21] The analysis assumes, for example, (1) that LSAT score and GPA linearly and additively affect a transformation of the probability of passing the bar and (2) that the effect of going from a sixth- to a fifth-tier school is roughly the same as going from a second- to a first-tier school. These assumptions are unjustified and unnecessary.[22] The analysis attempts to use such a model to predict how particular students would have performed in a counterfactual law school tier. Yet certain students are simply incomparable across tiers, and so predicting how they would have performed in a different tier is subject to highly questionable assumptions—what statisticians call extrapolating from the data. For example, if a new drug is found in a trial study to reduce cholesterol levels from 240 to 200, that does not mean that it would reduce levels from 100 to 60. To insure against such extrapolation, an analysis should check that first- and second-tier students are roughly comparable in the range of observed variables, such as LSAT score and undergraduate GPA. This is clearly an issue here: First-tier students on average scored five points higher on the LSAT (*t-stat* = 48.3) and earned an undergraduate GPA of 3.5, compared to 3.2 for non-first-tier students (*t-stat* = 33.6).

18. Beyond the reported variable set, Sander only assesses sensitivity to part-time status, family income, and parents' education. Sander, *supra* note 1, at 445 n.213. Clearly, admissions committees and students observe much more information than this.

19. *See* Stacy Berg Dale & Alan B. Krueger, *Estimating the Payoff to Attending a More Selective College: An Application of Selecting on Observables and Unobservables*, 117 Q.J. ECON. 1491 (2002).

20. *See* Ho, *supra* note 4, at 7-10 (matching on 180 variables to reassess mismatch).

21. *See* P. MCCULLAGH & J.A. NELDER, GENERALIZED LINEAR MODELS 107-10 (2d ed. 1989) (defining the generalized linear model with a logistic link).

22. *See* Daniel E. Ho, Kosuke Imai, Gary King & Elizabeth A. Stuart, Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference 13-16 (Oct. 13, 2004) (unpublished manuscript), *available at* http://gking.harvard.edu/files/matchp.pdf.

Would top-tier students have fared better on the bar had they attended a second-tier school? We can only estimate this with students who are actually similar in these respects.

II

In a reanalysis of the data, I (1) remove post-treatment bias by excluding law school GPA and (2) relax the role of unwarranted assumptions that extrapolate from the data by matching exactly on all variables in the original Sander analysis. Matching is a technique that is particularly suitable for drawing a causal inference with minimal assumptions.[23] It is also intuitive. Rather than relying on model assumptions regarding relationships of variables (e.g., that LSAT scores linearly affect a deterministic function of the latent probability of passing the bar), we simply find all students who are the same on all observable variables (LSAT score, undergraduate GPA, race, and gender) *except for* law school tier.[24] These represent the subset of students whom we might have randomly assigned to a tier in an experiment. (Recall that randomization is used precisely to ensure that treatment and control groups are similar in all variables.) The general guideline for choosing a matching model is to identify students who are as similar as possible to generate "balance" across tiers. Because matched students here are identical in every pretreatment respect for which Sander controls, better balance cannot be achieved within the confines of the original analysis. Once similar students from different tiers have been matched, assessing the difference in bar passage rates between students who attended different-tier schools is straightforward.[25]

Even before matching, eliminating post-treatment bias by removing law school grades from Sander's regression model reveals a sharp negative association between bar passage and being black, reflecting the black/white

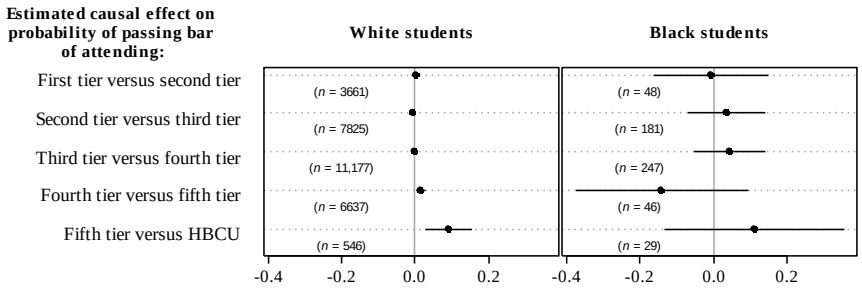
23. *See, e.g.*, Lee Epstein, Daniel E. Ho, Gary King & Jeffrey A. Segal, *The Supreme Court During Crisis: How War Affects Only Non-War Cases*, 80 N.Y.U. L. REV. 1 (2005) (matching cases decided during war and peace to assess how war affects civil rights and civil liberties).

24. I use freely available MatchIt software. Daniel E. Ho, Kosuke Imai, Gary King & Elizabeth A. Stuart, MatchIt: Nonparametric Preprocessing for Parametric Causal Inference, http://gking.harvard.edu/matchit (last visited Apr. 4, 2005).

25. To be precise, I follow the suggestions of Ho, Imai, King & Stuart, *supra* note 22: Effects are estimated by matching all students exactly on race, gender, LSAT score, and undergraduate GPA and simulating asymptotic posterior effects from a logistic model of bar passage on all covariates and tier indicator. Because all matches are exact, results are robust to the type of adjustment employed (e.g., logistic regression, subclass-weighted difference-in-means). GPA is discretized into tenths of a point, and LSAT score is discrete on a 10-to-48-point scale, so exact matches work particularly well in this application. I consider only exact matches between proximate tiers to reduce bias on unobservables. *See* Dale & Krueger, *supra* note 19, at 1492-93.

bar passage gap.[26] Including law school grades in the model reduces this test score gap, leading Sander to wrongly conclude that the black/white bar passage gap is attributable to affirmative action. More importantly, reducing the role of unwarranted assumptions by matching reveals no evidence that attending a higher-tier law school affects bar passage rates for similar students. Figure 1 summarizes estimated causal effects of attending a specified tier compared to the tier immediately beneath it.

FIGURE 1. ESTIMATED CAUSAL EFFECTS ON PROBABILITY OF BAR PASSAGE OF ATTENDING SCHOOL TIERS ON WHITE AND BLACK STUDENTS



Note: Horizontal lines indicate 95% confidence intervals; *n* represents total number of matched students. For estimation details, see *supra* note 25.

The left and right panels present effects on white and black students, respectively.[27] The horizontal axis represents the effect on the probability of passing the bar. The dots represent the average causal effect on bar passage, and the horizontal bars plot the 95% confidence interval, signifying the uncertainty of the estimate.[28] The vertical gray lines bisecting the horizontal axes indicate no detectible effect. If the confidence interval intersects this line, the difference in bar performance is statistically indistinguishable from 0. For example, the top left row indicates that after matching 3661 white students on all variables except for tier, the effect of attending a first-tier as opposed to a second-tier law school is statistically indistinguishable from 0. As Figure 1 shows, all but one of the estimates is close to 0, indicating no substantive impact of the marginal decision to attend a higher- or lower-tier school. Students with the same LSAT score, undergraduate GPA, and gender perform similarly on the bar irrespective of law school tier.

26. This is closest to the effect reported in Ayres & Brooks, *supra* note 2 (manuscript at 11-14), which, however, does not account for extrapolation across tiers and which redefines tier related to the white median within a range of index scores. Chambers and his coauthors subclassify on index score alone. Chambers et al., *supra* note 2 (manuscript at 22).

27. Sample sizes were insufficient to estimate comparable effects for Asians, Latinos, and other minorities, so these groups are excluded from this analysis.

28. The confidence intervals are wider for black students due to smaller sample sizes.

The one statistically significant result is that white students on average have a ten percent increase in bar passage probability if they attend a fifth-tier school rather than a historically black college or university (HBCU) (this ordinal scaling is solely due to Sander's coding). This result could be an indication that white students actually do better in homogenous environments, or it could be the result of a so-called "stereotype lift," but it is also possible that it is due to a failure to observe enough information. Students in the two tiers could be at different income levels, for example, in which case different bar passage rates might not be an effect of the school but perhaps of a higher rate of school-time employment at HBCUs.[29] That said, with enough tests we also expect statistically significant relationships to occur at a certain frequency even if the relationship is random (classic "Type 1" error). Regardless of the explanation, white students' stronger performance at fifth-tier schools than at HBCUs says nothing about Sander's hypothesis that black students fare worse at higher-tier schools.

The direct test of Sander's hypothesis is whether black students who are similar in qualifications but attend higher-tier schools fare worse on the bar. This is evidently not the case. While it is true that similarly qualified black students get lower grades as a result of going to a higher-tier school, they perform equally well on the bar irrespective of law school tier. Moreover, the lack of statistically significant differences does not appear to be simply a function of sample size: The point estimates are largely centered on 0. In short, whichever way one cuts it, there is no evidence for the hypothesis that law school tier causes black students to fail the bar.

III

As the reception of Sander's article demonstrates, empirical legal studies is an important and burgeoning research area. Yet just as scholars have realized the potential for empirical techniques that have energized research frontiers in the social sciences, we must also become aware of the assumptions, limitations, and credibility of those techniques. "The blind use of complicated statistical procedures . . . is doomed to lead to absurd conclusions."[30] Our ability to draw causal inferences is limited by the quality of the data and the credibility of the assumptions maintained. Once we understand those manipulable policies about which our data can actually be informative, empirical research may enrich and elucidate policy debates.

—Daniel E. Ho

29. This also suggests that ordering the tiers as in the original analysis, with mostly HBCUs comprising the bottom tier, may be questionable.

30. Holland & Rubin, *supra* note 5, at 18.