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).

By Daniel E. Ho[1]

In a widely discussed empirical study, Richard Sander concludes that affirmative action at U.S. law schools *causes* black students to fail the bar.[2] If correct, this conclusion would turn the jurisprudence, policy, and law of affirmative action on its head.[3] Yet the article misapplies basic principles of causal inference, which enjoy virtually universal acceptance in the scientific community.[4] 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

[1] All analyses presented in this Comment will be posted at http://people.iq.harvard.edu/~dho/. Thanks to Angela Early, Ian Ayres, Shameem Black, Rick Brooks, John Donohue, Aron Fischer, Jim Greiner, Kosuke Imai, Andy Kennedy, Bill Kidder, Gary King, Rick Lempert, Richard Sander, and Liz Stuart for helpful comments.

[2] Richard H. Sander, *A Systemic Analysis of Affirmative Action in American Law Schools*, 57 STAN. L. REV. 367, 447 (2004) ("[R]acial preferences in law school admissions significantly worsen blacks' individual chances of passing the bar . . . .").

[3] The article has already engendered a host of critical responses. *See, e.g.*, Ian Ayres & Richard Brooks, *Does Affirmative Action Reduce the Number of Black Lawyers?*, 57 STAN. L. REV. (forthcoming 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 2005); Michele Landis Dauber, *The Big Muddy*, 57 STAN. L. REV. (forthcoming 2005). This Comment is the first, however, to point out the study's inferential flaws of posttreatment bias and extrapolation.

[4] *See* Lee Epstein & Gary King, *The Rules of Inference*, 69 U. CHI. L. REV. 1, 19-110 (2002).

for similarly qualified black students, attending a higher-tier law school has no detectable effect on bar passage rates.

Part I of this Comment clarifies the assumptions implicit in the Sander 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 suggests that Sander's conclusions are untenable on their own terms.[5]

I

At the outset, it is important to note that since all of the schools in the LSAC Bar Passage Study on which Sander's analysis relies employ some system of affirmative action, no broad conclusion about the effects of affirmative action can be sustained.[6] 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.[7] Given this basic lack of information, what inferences can the data support?

Since there is no information in the dataset 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 the policy impact of affirmative action. For instance, if Sander is correct in claiming that students who go to a higher-tier school (1) are "mismatched" in terms of academic credentials, (2) learn less, and (3) fail the bar as a result, then a program like affirmative action might appear to hurt those it aims trying to help.

So how do we investigate this causal effect? Here, I introduce the assumptions required to interpret Sander's findings as causal effects.[8] Often stated in technical terms, these assumptions can be explained in substantively meaningful ways without math. Such explanation makes clear that the study's assumptions are implausible and internally inconsistent.

[5] For a more extensive and technical draft presenting 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://people.iq.harvard.edu/~dho/research/sander.pdf.

[6] *See* Paul W. Holland & Donald B. Rubin, *On Lord's Paradox*, *in* PRINCIPALS OF MODERN PSYCHOLOGICAL MEASUREMENT 3, 9-14 (Howard Wainer & Samuel Messick eds., 1983).

[7] *See* Harry Holzer & David Neumark, *Assessing Affirmative Action*, 38 J. ECON. LITERATURE 483, 508 (2000).

[8] *See* Paul W. Holland, *Statistics and Causal Inference (with Discussion)*, 81 J. AM. STAT. ASS'N 945 (1986); Donald B. Rubin, *Estimating Causal Effects of Treatments in Randomized and Non-Randomized Studies*, 66 J. EDUC. PSYCH. 688 (1974).

Two basic tenets underlie any causal inference. The first tenet is that causal inference is inherently *counterfactual*.[9] If we are interested in how Student *A* would perform on the bar by going to a top-tier versus a second-tier law school, we would ideally be able to observe *A* attend both schools. Yet if *A* attends a first-tier school we cannot observe *A* in the counterfactual world where she attends a second-tier school. This "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 where a researcher could manipulate a "treatment," or causal factor of interest. Laboratory scientists assess causal effects by actually conducting that experiment. For Sander's study, this would require *randomly assigning* a subset of students to tiers (the treatment) and observing differences in bar passage (the outcome). Randomization and a sufficiently large sample ensure that the subset of students we are comparing across tiers are similar, such that differences in 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 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 this 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 the Sander 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 grades have a stronger association with bar passage than law school tier, 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."[12] The central claim is that going to a higher-tier law school

[9] *See* Epstein & King, *supra* note 4, at 34-37; Holland, *supra* note 8, at 945.
[10] Holland, *supra* note 8, at 947.
[11] Sander, *supra* note 2, at 444 tbl.6.1.
[12] *Id.* at 445.

causes students to learn less and earn lower grades, decreasing bar performance by more than 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 this analysis is that if we hold constant all factors ("variables") that a law school admissions committee observes (i.e., 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, gender, and race) are not themselves affected by the treatment of law school tier (i.e., they are "pretreatment variables"). Why is that so? If we hold constant something that is itself affected by the treatment, then we would be removing precisely the effect that we are trying to study. So for example, when assessing the effect of smoking on death, we do not control for lung health, as this would remove precisely one of the primary ways in which smoking affects death.

Recall that Sander controls for law school grades. But Sander himself argues that law school tier affects law school grades.[13] Therefore, controlling for law school grades will never produce the right estimates of the effect of law school tier. To see just how pathological this "posttreatment bias" can be, take a hypothetical example of the causal effect of smoking.[14] If smoking causes both low birth weight and increases the infant mortality rate, incorrectly controlling for birth weight may lead to a completely wrong conclusion as to the causal effect. Suppose we have data on 110 smokers and 110 non-smokers for which the overall infant mortality rate is higher for parents who are smokers (61/110) than for nonsmokers (39/110). By controlling for birth weight, however, the infant mortality rate can actually be lower for *both* high-birth-weight and low-birth-weight smokers than for nonsmokers (1/10 vs. 30/100 and 60/100 vs. 9/10, respectively).[15] By wrongly controlling for birth weight (which is itself a consequence of smoking), we may estimate that smoking has

[13] *Id.* at 373 ("[R]acial preferences have the effect of . . . sharply lowering [black students'] average grades.").

[14] For more formal treatment, 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 656 (1984). The example presented here is a stylized version of the Wilcox-Russel hypothesis. *See* Allen J. Wilcox, *Birthweight and Perinatal Mortality: The Effect of Maternal Smoking*, 137 AM. J. EPIDEMIOLOGY 1098 (1993).

[15] This is the reverse of what statisticians call "Simpson's paradox," namely that 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 238 (1951).

exactly the opposite effect than what is true. To get a sense of how problematic this is, even if smoking had been *randomly assigned* in a classic experiment, controlling for birth weight induces posttreatment bias, yielding the wrong conclusion.

This is the first basic flaw in the Sander analysis. The mismatch hypothesis posits that admitting black students to a higher-tier school will cause lower grades and decreased 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 law school grades reveals that the aggregate impact of law school tier on bar passage within Sander's original framework is indetectable and that the bar passage difference between black and white students remains stark. Sander's ostensible decomposition of the law school tier effect into performance and eliteness effects thereby derives from a basic misreading 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*. If true, this assumption permits the researcher to attribute any remaining differences to law school tier. But the assumption is likely violated if top-tier students have better letters of recommendation or have graduated from more prestigious undergraduate institutions, which Sander does not "control" for.[16] A focus of the economics literature has been precisely on how these so-called "unobserved" factors might affect outcomes, and for good reason.[17] If there are unobservable differences in students across tiers, the original analysis as well as the approach presented here fail. Differences in bar passage rates could be due to any number of unobserved factors, not law school. Hence, the approach presented in Part II also requires assuming the absence of unobserved factors. An extension of the Sander analysis controlling for a wider range of variables, thereby making this assumption more believable, further indicates that there is no evidence for the Sander hypothesis.[18]

Lastly, as part of the regression analysis Sander makes assumptions about how the pretreatment variables affect the probability of bar passage.[19] The regression analysis assumes, for example, that LSAT and GPA linearly and additively affect a transformation of the probability of passing the bar.

[16] Sander only assessed sensitivity beyond the reported variable set to part-time status, family income, and parents' education. Sander, *supra* note 2, at 445 n.213. Clearly, admissions committees and students observe much more information than this.

[17] *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).

[18] *See* Ho, *supra* note 5 (matching on 180 variables to reassess the bar-passage hypothesis).

[19] *See* PETER MCCULLAGH & J.A. NELDER, GENERALIZED LINEAR MODELS 107-10 (2d ed. 1989).

These assumptions are unjustified and largely untestable from the data.[20] The key to keep in mind here is that Sander wants to use his model to predict the counterfactual of 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 to reduce cholesterol levels from 240 to 200 in a trial study, that does not mean that it would reduce levels from 100 to 60. To ensure against such extrapolation, an analysis should check that sufficient first and second tier students, for example, exist in the range of pretreatment variables. This is clearly an issue here: First-tier students on average scored 5 points higher on their LSAT scores (t-stat=48.3) and had an average undergraduate GPA of 3.5, compared to 3.2 for non-first-tier students (t-stat=33.6). Would first tier students have fared worse on the bar by attending a second tier school? We can only predict this by looking at students who are actually similar in these respects.

II

In a reanalysis of the data, I (1) correct for posttreatment bias by omitting law school GPA and (2) relax the role of unwarranted assumptions that extrapolate from the data by matching exactly on all variables.[21] Matching is a technique that is particularly suitable for drawing a causal inference with minimal assumptions.[22] It is also intuitive. Rather than relying on model assumptions regarding the structure of causal relationships (e.g., that LSAT scores linearly affects a deterministic function of the latent probability of passing the bar), we simply find all students that are the same on all observable variables (LSAT, undergraduate GPA, race, and gender) *except for law school tier*.[23] 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 achieve the purpose that treatment and control groups are similar in all

[20] *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. 12, 2004) (unpublished manuscript), *available at* http://gking.harvard.edu/files/matchp.pdf.

[21] The approach follows Ho, Imai, King & Stuart, *supra* note 20.

[22] *See, e.g.*, Lee Epstein, Daniel E. Ho, Gary King & Jeffrey A. Segal, *The Effect of War on the Supreme Court*, 80 N.Y.U. L. Rev. (forthcoming 2005) (matching Supreme Court cases decided during war and during peace to assess the impact of war on civil rights and liberties decisions).

[23] I use freely available MatchIt software. Daniel E. Ho, Kosuke Imai, Gary King & Elizabeth A. Stuart, MatchIt: Nonparametric Preprocessing for Parametric Causal Inference (Jan. 10, 2005), *available at* http://gking.harvard.edu/matchit/.

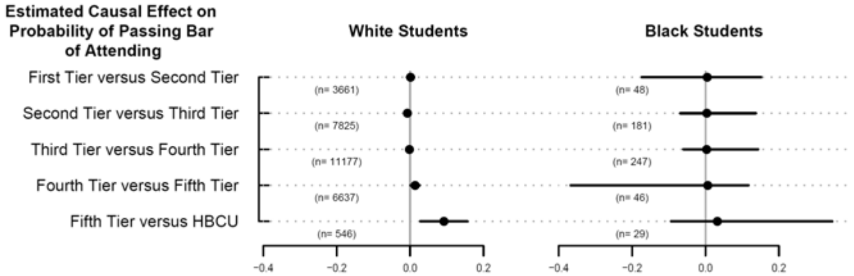

Figure 1: Estimated Causal Effects of Attending School Tiers on White Students (left panel) and Black Students (right panel) with 95% confidence intervals. Estimates are derived by exact matching all students on race, gender, LSAT and undergraduate GPA, and simulating asymptotic posterior effects from a logistic model of bar passage on all covariates and tier indicator. *n* represents matched sample sizes. This Figure shows that attending a higher-tier law school has no detectable effect on black students. The only substantial effect is for white students for whom attendance at a fifth-tier school as opposed to an historically black college or university (HBCU) causes a gain of roughly ten percentage points. This ordinal scaling is due to the original Sander analysis.

variables.) The general guideline for how to choose a matching model is to identify students as similar as possible to create "balance" across tiers. Since matched students here are *identical in every pretreatment respect* for which Sander controlled, better balance cannot be achieved within the confines of the original analysis. Once similar students from different tiers have been matched, it is straightforward to assess the difference in bar passage rates between students who attended different tiers.[24]

Even before matching, one telling sign is that accounting for posttreatment bias by simply omitting law school grades reveals a sharp *negative* association with being black, reflecting the black-white bar passage gap.[25] (Incorrectly including law school grades reduces this test score gap, leading Sander to wrongly conclude the black-white bar passage gap is solely 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 similarly qualified students.

Figure 1 summarizes estimated causal effects of attending a specified tier compared to the tier immediately beneath it. The left panel presents

[24] To be precise, I follow the suggestions of Ho, Imai, King & Stuart, *supra* note 21, and exact match on race, gender, LSAT, and undergraduate GPA, followed by logistic regression adjustment with a treatment indicator and gender, LSAT, and undergraduate GPA. Because all matches are exact here, results are robust to the type of adjustment employed (e.g., logistic regression, subclass-weighted difference-in-means). Because GPA is discretized into tenths of a point and LSAT is discrete on a ten to forty-eight-point scale, exact matches work particularly well in this application. I consider only exact matches between proximate tiers: first and second tier, second and third tier, etc., to reduce bias on unobservables. *See* Dale & Krueger, *supra* note 17.

[25] This is closest to the effect reported in Ayres & Brooks, *supra* note 3, which, however, does not account for extrapolation across tiers and redefines tier relate to the white median within a range of index scores. Chambers et al., *supra* note 3, subclassify on index score alone.

effects on white students, and the right panel presents effects on black students.[26] 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.[27] The vertical grey line intersecting the middle of the horizontal axis indicates no effect. If the confidence interval intersects this line, the difference in bar performance is statistically indistinguishable from zero. For example, the top left row indicates that matching 3661 white students exactly on all variables except for tier, the effect of attending a first-tier as opposed to second-tier law school is statistically indistinguishable from 0. As Figure 1 shows, all but one of the estimates for white students are close to zero, indicating no substantive impact of the marginal decision to attend a higher- or lower-tier school. In other words, students similar in LSAT, undergraduate GPA, and gender will perform similarly on the bar whether they attend a higher-tier school or not.

The one statistically significant result is the effect of attending a fifth-tier school versus a historically black college or university (HBCU) (note that this ordinal scaling is taken from the original Sander analysis): White students on average have a ten percent increase in bar passage probability if they attend a fifth-tier (non-HBCU school) instead of an HBCU. This could be an indication that white students actually do better at homogenous environments, but it is also possible that this is due to a failure to observe enough information. The small number of students could be different in income levels, for example, in which case it is not an effect of the school but perhaps a higher rate of school time employment at HBCUs.[28] That said, with enough tests we also expect a statistically significant relationship to occur at a certain frequency even if the relationship is random (classic "type 1" error). Regardless of the explanation, white students' superior 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 that black students who are similar in qualifications but attend higher-tier schools should 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 just as well on the bar irrespective of law school tier. Moreover, the lack of statistically significant differences does not appear to

[26] Sample sizes were insufficient to estimate comparable effects for Asian, Latino, and other minorities, so these are excluded from these results.

[27] The confidence intervals are wider for black students due to smaller sample size.

[28] This also suggests that ordering the tiers as in the original analysis, with mostly historically black colleges and universities as the bottom tier, may be questionable.

be simply a function of sample size: Virtually all of the point estimates are centered at zero. 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, the field of empirical legal studies is an important and burgeoning research area in the law. Yet just as scholars have realized the potential for empirical techniques that have energized research frontiers in the social sciences, scholars must simultaneously become aware of the assumptions, limitations, and credibility of empirical techniques. "The blind use of complicated statistical procedures . . . is doomed to lead to absurd conclusions."[29] Our ability to draw make causal inferences is limited by the quality of the data collected and the credibility of the assumptions maintained. And once we understand those *manipulable* policies about which our data can actually be informative, empirical research can serve to inform, enrich, and elucidate those policies that are actually close to being considered for adoption.

[29] Holland & Rubin, *supra* note 6, at 18.