U.C.L.A. Law Review

How Not to Lie About Affirmative Action

Sherod Thaxton

ABSTRACT

As challenges to race-conscious admissions policies are, once again, advancing through the federal courts, research proclaiming to identify the wideranging effects of affirmative action across a variety of educational settings is influencing this litigation through amici and expert testimony. It is crucial, then, that empirical research used to support claims by parties on either side of the affirmative action debate adhere to the fundamental precepts of causal inference. Yet the literature on causal inference is both vast and dense, and as a result, many judges, lawyers, legislators, and laypersons interested in understanding both the intended and unintended consequences of affirmative action are ill-equipped to understand the debate—especially when quantitative social scientists on both sides of the issue appear to draw conflicting (though not necessarily equally credible) inferences from the same data. The purpose of this Article is to lay bare the core requirements of credible causal inference to the uninitiated, highlighting how inattention to (and sometimes outright disregard for) these rules has muddied the debate over the effect of affirmative action in law schools and in college admissions more generally. The Article empirically examines the six primary deficiencies impacting extant research on affirmative action in law schools: (1) posttreatment bias, (2) nonresponse bias, (3) omitted variable bias, (4) interpolation bias, (5) extrapolation bias, and (6) measurement error bias. I conclude the Article by describing what a scientifically defensible examination of the effect of affirmative action in legal education with currently available data would entail. While no approach to causal inference is infallible, the careful analyst can attempt to ameliorate the impact of these biases.

AUTHOR

Sherod Thaxton is Professor of Law at the UCLA School of Law. He also holds courtesy appointments in the Department of African American Studies and the Department of Sociology.

For comments on or conversations about this Article, many thanks to Sameer Ashar, LaToya Baldwin Clark, Devon Carbado, Jennifer Chacón, Kimberlé Crenshaw, Laura Gómez, Cheryl Harris, Richard Lempert, Jerry López, Sunita Patel, and Noah Zatz. Linda Karr O'Connor, Director of Scholarly Support and Research Assistance at the Hugh & Hazel Darling Law Library at UCLA School of Law, offered valuable research assistance. The editors and staff of the *UCLA Law Review* provided outstanding editorial contributions. Naturally, all remaining errors are my own.

TABLE OF CONTENTS


THE UNIVERSITY OF CALIFORNIA
UCLA
SCHOOL OF LAW

"Causal claims are important for society and it is crucial to know when scientists can make them."[1]

"The first step to not fooling others [with statistics] is to not fool ourselves."[2]

INTRODUCTION

Over thirty years ago, in his pioneering article, "How Not to Lie with Statistics," prominent political methodologist Gary King remarked, "One of the most glaring problems with much quantitative political science is its uneven sophistication and quality. . . . [T]he *same* mistakes are being made or 'invented' over and over. . . . Too often, we *learn* each others' mistakes rather than *learning from* each others' mistakes."[3] In *Mismatch: How Affirmative Action Hurts Students It's Intended to Help and Why Universities Won't Admit It* (hereinafter *Mismatch*), Richard Sander and his coauthor, Stuart Taylor, Jr., offer a critique of race-conscious affirmative action policies in college/university admissions and claim, *inter alia*, that the "evidence points *overwhelmingly* toward a large law school mismatch problem, one that affects Hispanics as well as blacks" and that "all of the other tests also support mismatch when they are modeled in a reasonable way."[4] On closer inspection, however, the boldly stated conclusions presented in *Mismatch* withstand neither conceptual nor empirical scrutiny,[5] and are vulnerable to the same

1. John Antonakis, Samuel Bendahan, Philippe Jacquart & Rafael Lalive, *On Making Causal Claims: A Review and Recommendations*, 21 LEADERSHIP Q. 1086, 1086 (2010).
2. Andrew Gelman, *Ethics in Statistical Practice and Communication: Five Recommendations*, SIGNIFICANCE, Oct. 2018, at 40, 43.
3. Gary King, *How Not to Lie With Statistics: Avoiding Common Mistakes in Quantitative Political Science*, 30 AM. J. POL. SCI. 666, 666, 684 (1986).
4. RICHARD SANDER & STUART TAYLOR, JR., MISMATCH: HOW AFFIRMATIVE ACTION HURTS STUDENTS IT'S INTENDED TO HELP, AND WHY UNIVERSITIES WON'T ADMIT IT 86 (2012) (emphasis added); *see also* Richard H. Sander, *Replication of Mismatch Research: Ayres, Brooks and Ho*, 58 INT'L REV. L. & ECON. 75, 82 (2019) [hereinafter Sander, *Replication of Mismatch*] ("In short, we think [our] results are very powerful confirmation of mismatch."); *accord* Richard H. Sander, *Replication of Mismatch Research: The Case of Ayres and Brooks* (UCLA Summer Works-in-Progress Workshop, Series 17, 2018) [hereinafter Sander, Whitepaper]. Sander, *Replication of Mismatch*, *supra*, serves, in part, as the methodological appendix for the analyses presented in *Mismatch* and related work.
5. Conceptual and empirical analysis are distinct inquiries. "There is, in short, a clear and decisive difference between *knowing how to test* a battery of hypotheses and *knowing the theory* from which to derive hypotheses to be tested." Robert K. Merton, *Sociological Theory*, 50 AM. J. SOCIO. 462, 463 (1945). Conceptual analysis consists of studying the constituent parts of a theory in order to gain a clearer understanding of the particular issue in which the theory is involved. *See* Guillermina Jasso, *Principles of Theoretical Analysis*, 6 SOCIO. THEORY 1, 11 (1988); *id.* at 11–19; *see also,* especially, *infra* Part IV. Empirical analysis,

core criticisms that have plagued research ostensibly demonstrating mismatch effects in law schools for nearly fifteen years. While *Mismatch* purports to identify the deleterious effects of affirmative action on its intended beneficiaries across a variety of education settings,[6] Sander is primarily known for his research in the law school context, and *Mismatch* focuses heavily on earlier critiques of Sander's work in this area. In particular, *Mismatch* responds to several articles and *amicus curiae* briefs written by critics of Sander's earlier research which proclaimed to show, empirically, that affirmative action causes black students to perform poorly in law school, drop out of law school, fail the bar exam, and earn less as lawyers.[7] *Mismatch* claims to identify and correctly resolve problems with those critiques, and after doing so, uncover even stronger support for mismatch theory. According to proponents of mismatch theory, beneficiaries of affirmative action in law schools learn less than they would have in the absence of such policies, and as a result, suffer the negative consequences of being outmatched by their peers in law school.[8] Critics of mismatch theory have countered:

---

on the other hand, involves an assessment of a theory's truthfulness or accuracy—meaning how well the theory is supported by empirical research evidence. *See* KARL POPPER, THE LOGIC OF SCIENTIFIC DISCOVERY 91, 281 (2d ed. 1968); *see also,* especially, *infra* Part II.

6. *See infra* note 379 (summarizing the research literature on mismatch effects at the undergraduate level and concluding that, based on the most comprehensive and methodologically sound studies of elite private and public universities, there is no evidence of mismatch effects).

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

   Technically speaking, Richard Sander does not measure the impact of affirmative action; rather, he claims to measure the effect of attending a more selective (and academically competitive) law school. Daniel E. Ho, Comment, *Why Affirmative Action Does Not Cause Black Students to Fail the Bar*, 114 YALE L.J. 1997, 1998 (2005) ("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."). But as Peter Arcidiacono and Michael Lovenheim explain, "[t]aken at face value, the estimates in [Sander, *Systemic Analysis*, *supra*] suggest that attending a more elite law school lowers one's chances of passing the bar *regardless of one's entering credentials or race*. . . . A problematic conclusion one could draw from Sander's results is that everyone is harmed by going to a more elite law school, as the negative effect on [law school] GPA swamps the positive direct effect of school quality." Peter Arcidiacono & Michael Lovenheim, *Affirmative Action and the Quality-Fit Trade-Off*, 54 J. ECON. LITERATURE 3, 17 (2016) (emphasis added and omitted).

8. Sander refers to the negative learning consequences resulting from affirmative action policies as "learning mismatch." *See* Richard Sander & Aaron Danielson, *Thinking Hard About "Race-Neutral" Admissions*, 47 U. MICH. J.L. REFORM 967, 984–85 (2014). He also hypothesizes that affirmative action policies have two additional negative consequences: "competition mismatch" and "social mismatch." *Id.* at 984–88; *see also* Richard Sander,

> [T]he fact that students at more elite schools generally have higher incoming credentials. . . . in turn, creates a tougher pool within which to compete for grades but also creates a pool more likely to achieve success on the bar exam and in employment.[9]

Mismatch, as well as prior[10] and subsequent[11] work by Sander, focuses significant attention on an article coauthored by Ian Ayres and Richard Brooks,[12] in which Ayres and Brooks offer a modified test of mismatch theory using the same data as Sander and fail to find any support for Sander's conclusions. According to Sander and Taylor, Ayres and Brooks's paper was the "single-most widely anticipated response," "[b]oth [Ayres and Brooks] were sophisticated empiricists with doctorates in economics, and Ayres was one of the most famous social scientists at any law school,"[13] and Ayres and Brooks's article offered "a compelling, independent test of mismatch."[14] But as I explain below, Ayres and Brooks's analysis was not the most compelling test of mismatch because it was based on two highly questionable assumptions that Ayres and Brooks explicitly acknowledged could not be verified by the

The Stylized Critique of Mismatch, 92 TEX. L. REV. 1637, 1642–43 (2014) [hereinafter Sander, *Stylized Critique*]. Competition mismatch is akin to the economic concept of substitution effects, and occurs when a student, because of receiving lower grades due to affirmative action, modifies her behavior in response and, for example, changes her original education or career plans. *Id.* Social mismatch supposes that students self-segregate because of academic credentials. *Id.* at 1643. According to Sander, these three phenomena—which he refers to as "first-order mismatch effects"—provide a direct theoretical link between disparity in academic credentials and various outcomes. *Id.* at 1657. Sander hypothesizes that these first order effects may lead to second order effects, such as lower graduation rates, failing the bar examination, or lower wages for students experiencing mismatch. *Id.* at 1643–45; *accord* Sander & Danielson, *supra*, at 984. *But see* CAMILLE Z. CHARLES, MARY J. FISCHER, MARGARITA A. MOONEY & DOUGLAS S. MASSEY, TAMING THE RIVER: NEGOTIATING THE ACADEMIC, FINANCIAL, AND SOCIAL CURRENTS IN SELECTIVE COLLEGES AND UNIVERSITIES 200 (2009) (analyzing academic performance of college students at twenty-eight colleges and universities and noting that there is "no evidence that affirmative action exacerbates the basic process of disidentification for individual minority students [and] affirmative action . . . does not bring about a reduction of objective or subjective work effort on the part of students either by itself or through interaction with stereotype internalization").

9. Gregory Camilli & Kevin G. Welner, *Is There a Mismatch Effect in Law School, Why Might it Arise, and What Would it Mean?*, 37 J.C. & U.L. 491, 504–05 (2011).
10. *E.g.*, Richard H. Sander, Reply, *A Reply to Critics*, 57 STAN. L. REV. 1963 (2005) [hereinafter Sander, *Reply to Critics*].
11. *E.g.*, Sander, *Replication of Mismatch*, *supra* note 4.
12. Ian Ayres & Richard R. W. Brooks, Response, *Does Affirmative Action Reduce the Number of Black Lawyers?*, 57 STAN. L. REV. 1807 (2005).
13. SANDER & TAYLOR, *supra* note 4, at 77.
14. *Id.*; *see also* Sander, *Reply to Critics*, *supra* note 10, at 1986 ("I was delighted when I heard Ian Ayres and Richard Brooks were writing one of the responses in this issue; I know and respect both men, and I felt both of them would give the issues a fair and fresh examination.").

data[15]—as well as other unstated assumptions that were capable of being evaluated by the data[16]—rendering their results merely suggestive rather than conclusive. Ayres and Brooks explain, "Unfortunately, the nature of the data underlying this analysis prevents us from reaching definitive conclusions—just as Sander cannot definitely assert the converse, again, given the available data."[17] These shortcomings notwithstanding, Ayres and Brooks's analysis identified important problems with Sander's analytical framework and their results undermined Sander's claim that black students were harmed by affirmative action.[18]

In response to Ayres and Brooks's article, Sander claims that when "correctly done, [Ayres and Brooks's test] produced a *stunning confirmation* of the [mismatch] theory";[19] but, as Ayres and Brooks underscore in their analysis, the data from which Sander's conclusions are derived are incapable of providing confirmation or disconfirmation of mismatch effects, and simply show that Sander's conclusions concerning mismatch effects are unreliable and highly dependent on his modeling assumptions.[20] In fact, Ayres and Brooks even carefully qualify their findings of a reverse mismatch effect—they reported that affirmative action *increases* the number of black lawyers.[21] Other critics of

15. Ayres & Brooks, *supra* note 12, at 1832 (explaining that their examination of mismatch is "not an implausible test, [but] it is important to describe its limitations").
16. *See, e.g.*, *infra* Subparts II.D–II.E (discussing, respectively, interpolation bias and extrapolation bias).
17. Ayres & Brooks, *supra* note 12, at 1853; *see also infra* note 26 and accompanying text (emphasizing the deficiencies present in Ian Ayres and Richard Brooks's analytical framework and the superiority of an alternative approach presented around the same time as Ayres and Brooks's analysis by political methodologist Daniel Ho); *infra* Subpart II.F.1.b.
18. Ayres & Brooks, *supra* note 12, at 1853 ("Sander's conclusion that these [racial] disparities [in bar passage rates] are dominantly or solely caused by affirmative action does not withstand closer analysis. . . . [W]e found no compelling evidence that the system of affirmative action in place in 1991 reduced the number of black lawyers.").
19. SANDER & TAYLOR, *supra* note 4, at 83 (emphasis added); *see also* Sander, *Reply to Critics*, *supra* note 10, at 1995 ("Ayres and Brooks's findings provide as much support for the mismatch theory as was conceivable going in, given the restrictive way they constructed their tests.").
20. Ayres & Brooks, *supra* note 12, at 1833. Ayres and Brooks explain:
    > [Our analytical] approach—with all its limitations—still appears to be among the strongest and most direct available evidence on mismatch in the law school setting. We make this point not as an indication of the quality of the analysis, but as a strong statement about the weaknesses of the data (for this question) on which we and Sander rely. We leave it to the responsible reader to make of it what she will.

    Id.
21. Ayres & Brooks, *supra* note 12, at 1816 ("[W]e find that elimination of affirmative action would result in a decrease of black lawyers. . . . We do not wish to claim that our predictions . . . are foolproof. All of these predictions (Sander's and ours) are sensitive

mismatch theory have identified several fundamental, and persistent, problems with Sander's analytical framework, but rather than responding to those perceived shortcomings in a comprehensive manner, Sander's responses to these critiques, including those offered in *Mismatch*, narrowly focus on a smaller set of claims and either completely ignore or inadequately address most of the alleged errors that invited such strong criticism to begin with.[22]

Interestingly, the most compelling and methodologically sound independent test of law school mismatch was not conducted by Ayres and Brooks; rather, it was presented by political methodologist, and then–law student, Daniel Ho, in the form of two short Comments in the *Yale Law Journal* and a methodological appendix, all totaling a mere twenty-eight pages.[23] Sander's reply

---

to the assumptions behind them."); *accord* Jesse Rothstein & Albert Yoon, *Mismatch in Law School* (Nat'l Bureau of Econ. Rsch., Working Paper No. 14275, 2008) (finding no evidence of mismatch effects and some evidence of reverse mismatch effects); *see also infra* note 379 and accompanying text (summarizing research from several national studies of student performance at highly selective private and public colleges and universities that report reverse mismatch effects for black and Latinx students).

22. *See* William C. Kidder & Richard O. Lempert, *The Mismatch Myth in U.S. Higher Education: A Synthesis of Empirical Evidence at the Law School and Undergraduate Levels*, *in* AFFIRMATIVE ACTION AND RACIAL EQUITY: CONSIDERING THE *FISHER CASE* TO FORGE THE PATH AHEAD 105, 108 (Uma M. Jayakumar & Liliana M. Garces eds., 2015) [hereinafter Kidder & Lempert, *Mismatch Myth*] ("With the exception of Sander's erstwhile coauthor Doug Williams, every social scientist we know of who has independently analyzed the data Sander used has reported results that dispute his conclusions."). A longer version of William Kidder and Richard Lempert's paper is available online. WILLIAM C. KIDDER & RICHARD O. LEMPERT, THE MISMATCH MYTH IN AMERICAN HIGHER EDUCATION: A SYNTHESIS OF EMPIRICAL EVIDENCE AT THE LAW SCHOOL AND UNDERGRADUATE LEVELS (2014) [hereinafter KIDDER & LEMPERT, WHITE PAPER], http://www.law.uh.edu/ihelg/monograph/13-12.pdf [https://perma.cc/2G6X-TLPS].

In his defense of mismatch theory, Williams notes, "With the exception of [Sander, *Systemic Analysis*, *supra* note 7], previous research using the BPS [Bar Passage Study] has concluded that the balance of evidence is against the mismatch hypothesis." Doug Williams, *Do Racial Preferences Affect Minority Learning in Law Schools?*, 10 J. EMPIRICAL LEGAL STUD. 171, 172 (2013). Williams claims his study makes several methodological improvements over prior work analyzing the BPS data and he reports strong support for mismatch theory. But Williams fails to address many of the core concerns raised by the initial critics of Sander's work, and he inadequately addresses others. These fundamental flaws in his research have been identified by several prominent empirical scholars, including a list of who's who among econometricians. *See infra* note 30 and accompanying text.

23. *See* Ho, *supra* note 7 (8 pages); Daniel E. Ho, Reply, *Affirmative Action's Affirmative Actions: A Reply to Sander*, 114 YALE L.J. 2011 (2005) [hereinafter Ho, *Reply to Sander*] (6 pages); 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) [hereinafter Ho, Evaluating Affirmative Action] (14 pages); *see also* David Bjerk, Comment, *Replication of Mismatch Research: Ayres, Brooks, and Ho*, 58 INT'L REV. L. & ECON. 3 (2019) ("Arguably the most well-known of these more

to Ho's critical assessment of his work consisted of proffering two different analyses that simply reproduced the errors Ho initially highlighted and failed to squarely address the core shortcomings that Ho identified.[24] Perhaps this is why *Mismatch* mentions Ho's critique only once, summarily dismissing it in a single sentence as containing "an obvious design flaw" and claiming that, when the alleged flaw is removed, Ho's analysis provides support for mismatch.[25] It would have been more valuable for *Mismatch* to focus on Ho's analysis, which was published around the same time of Ayres and Brooks's evaluation, because, until now, Ho provided the most insightful and penetrating critique of Sander's research on mismatch effects.[26] Ho explains that his assessment of Sander's analyses, unlike previous work (including Ayres and Brooks's article), "formally

rigorous critiques [of mismatch] are Ayres and Brooks (2004) and Ho (2005), both of whom argue that their results largely negate the conclusions Sander draws in his 2004 paper.").

24. Richard H. Sander, Response, *Mismeasuring the Mismatch: A Response to Ho*, 114 YALE L.J. 2005 (2005) [hereinafter Sander, *Mismeasuring*]; *see also infra* Subpart II.A.2; *infra* Appendices A–C.

25. SANDER & TAYLOR, *supra* note 4, at 85. Sander's response to Ho raised several issues that, according to Sander, undermined Ho's critique. *See* Sander, *Mismeasuring*, *supra* note 24, at 2009. Sander, however, is unable to identify anything in the vast causal inference literature to support his assertions. In fact, Sander only makes a vague reference to one social science text on research design in education settings to support his analysis: a very dated, classic work by Donald Campbell and Julian Stanley, DONALD T. CAMPBELL & JULIAN C. STANLEY, EXPERIMENTAL AND QUASI-EXPERIMENTAL DESIGNS FOR RESEARCH (1963). *See* Sander, *Mismeasuring*, *supra* note 24, at 2006 n.10 (citing CAMPBELL & STANLEY, *supra*). Yet Sander's cite to his sole reference, which is replete with cautions about drawing inferences from treatment and control groups absent random assignment, is misguided. For example, Campbell and Stanley note that "randomization is conceived to be a process occurring at a specific time, and is the all-purpose procedure for achieving pretreatment equality of groups, within known statistical limits," CAMPBELL & STANLEY, *supra*, at 6, and "[t]he usual statistics are appropriate only where individual students have been assigned at random to treatments [and controls]," *id.* at 23. Moreover, the "obvious design flaw" that Sander purported identified in Ho's work is equally applicable to his own, that is: (1) the coarseness of the tier location variable, and (2) the problem of omitted variables. With respect to the latter critique, Ho's analysis fares much better than Sander's analysis because it controls for 180 variables, whereas Sander takes less than ten into account. Ho, Evaluating Affirmative Action, *supra* note 23, at 8–9.

Sander has written another response to Ho's criticism, *see* Sander, *Replication of Mismatch*, *supra* note 4, but he fails to offer a credible rebuttal to Ho's article that faithfully adheres to the core requirements of causal inference emphasized in Ho's assessment of "Systemic Analysis." Moreover, in his response to Ho, Sander reproduces the very mistakes that Ho and other critics have cautioned against since "Systemic Analysis" was first published (instability bias, nonresponse bias, omitted variable bias, extrapolation bias, and measurement error bias). *See infra* Parts II–IV.

26. Kidder & Lempert, *Mismatch Myth*, *supra* note 22, at 109, 112 (noting that Ayres and Brooks's analysis, which was later adopted by Sander and Williams to provide support for mismatch effects, cannot yield reliable conclusions because of implausible assumptions, and identifying Ho's analysis as avoiding these problems).

evaluate[s] and analyze[s] [Sander's] study in a statistical framework of causation."[27] Ho's approach is unique because he clarifies several of the core assumptions of causal inference. And Ho was well-suited to do so because at the time he published his review, he had already written a Ph.D. dissertation at Harvard University (Harvard) titled, "Causal Inference in Political Science and Law,"[28] and studied under two of the giants in the field of causal inference.[29] In fact, the accuracy and comprehensiveness of Ho's critique of Sander was underscored in an amicus brief written by a distinguished group of theoretical and applied statisticians (hereinafter the "Empirical Scholars Brief" or "ESB"),[30] filed in response to Sander's own amicus brief,[31] in *Fisher v. University of Texas (Fisher I)*.[32] Even assuming for the sake of argument that Sander's criticisms of

27. Ho, Evaluating Affirmative Action, *supra* note 23, at 2.
28. Daniel E. Ho, Causal Inference in Political Science and Law (2004) (unpublished Ph.D. dissertation, Harvard University) (on file with Harvard University).
29. Ho was mentored by Gary King and Donald J. Rubin. King is currently the Albert J. Weatherhead III University Professor at Harvard University. He was elected to the American Academy of Arts and Sciences and the National Academy of Sciences (NAS) in 1998 and 2010, respectively. He was elected to the American Association for the Advancement of Science (AAAS) and American Statistical Association (ASA) in 2004 and 2009, respectively. Rubin is the John L. Loeb Professor of Statistics at Harvard University. He was elected to the American Academy of Arts and Sciences, NAS, AAAS, and ASA in 1993, 2010, 1984, and 1977, respectively.
30. The brief was written by eleven of the foremost experts in statistical and causal inference in support of the University of Texas in *Fisher v. University of Texas (Fisher I)*, 570 U.S. 297 (2013), "in order to point out to the Court the substantial methodological flaws in the research discussed in the Brief Amici Curiae for Richard Sander and Stuart Taylor, Jr. in Support of Neither Party." Brief of Empirical Scholars as Amici Curiae in Support of Respondents at 1, *Fisher I*, 570 U.S. 297 (No. 11-345) [hereinafter Empirical Scholars Brief in *Fisher I*]. Sander himself acknowledged the gravitas of the signatories: "There were eleven signatories to the ESB . . . . It was an eminent group, and the brief emphasized the prestige they brought to the enterprise by giving a short biography of each author. . . . These names alone gave the ESB an air of credibility and serious purpose." Richard Sander, *Mismatch and the Empirical Scholars Brief*, 48 VAL. U. L. REV. 555, 566–67 (2014) [hereinafter Sander, *Mismatch & the ESB*].

    The brief identifies three key problems with Sander's work: (1) invalid comparisons because he compares black to white students and not black to black students (Sander's Ayres and Brooks replication attempts to avoid this problem); (2) invalid adjustment for posttreatment outcomes (for example, law school graduation and law school grades); and (3) insufficient controls for pretreatment characteristics (something Ho's critique strongly emphasized). *See* Empirical Scholars Brief in *Fisher I*, *supra*, at 19–25.
31. The ESB states that "[r]esearch that applies these principles [of causal inference] has not found any substantially and statistically significant effects on bar passage. . . . Taking into account these principles of research design, there is simply no evidence of the harms of mismatch suggested by [Richard Sander's] Brief." Empirical Scholars Brief in *Fisher I*, *supra* note 30, at 25 (footnote and citation omitted). The brief specifically cites Ho's initial critique of "Systemic Analysis." *Id.*
32. 570 U.S. 297. In *Fisher I*, a white student, Abigail Fisher, challenged the constitutionality of race-conscious undergraduate admissions at the University of Texas. *Id.* at 297.

the Ayres and Brooks article are spot-on, *Mismatch* still does not provide the support for the unequivocal conclusions that Sander and Taylor announce because their analyses suffer from numerous other flaws that may not have been central to Ayres and Brooks's analysis (or even identified by Ayres and Brooks), but were highlighted by both Ho and the eleven signatories on the ESB (which also included Ho)—nearly all of whom are considered giants in the field of causal inference across a host of academic disciplines.[33] In fact, Ho expressly stated that the initial critiques of "Systemic Analysis," including the Ayres and Brooks paper, failed to identify the full range of inferential mistakes contained in that article.[34] Ho also identified additional problems with "Systemic Analysis" that he, himself, did not specifically address in his responses to Sander.[35]

As I explain below, Ho identifies multiple fundamental flaws in "Systemic Analysis" that *Mismatch* chose to simply ignore. Moreover, the one alleged flaw in Ho's work that Sander discusses in *Mismatch* relates to the manner in which law schools are grouped together in the Bar Passage Study (BPS) dataset;[36] this critique, however, applies equally to all of Sander's analyses.[37] In a recently published article, Sander provides a more detailed response to the criticisms against "Systemic Analysis" levied by Ho thirteen years earlier, and remarks, "Overall, I think it is fair to say that [Ho's] test does not tell us very much. We would not want to build a theory of mismatch around [Ho's] evidence. . . . [O]n the whole, [the evidence is] quite consistent with it."[38] Sander's analysis remains irreconcilably flawed, however, because it both

33. *See* Empirical Scholars Brief in *Fisher I*, *supra* note 30, at 16 ("The chief empirical research offered by [Sander's brief] . . . violates basic principles of causal inference that are widely accepted in the scientific community."); *see also infra* note 430 and accompanying text.
34. Ho, *supra* note 7, at 1997 n.2 ("[Sander's] article has already engendered a host of critical responses. . . . This Comment is the first, however, to point out the study's inferential flaws of post-treatment bias and extrapolation." (citations omitted)).
35. Ho, Evaluating Affirmative Action, *supra* note 23, at 10 n.7 ("While I have focused on one particular aspect of Sander's analysis, there are a host of other statistical issues that could threaten the validity of inferences. . . . Nonetheless, even staying within the framework provided by Sander, the conclusions fail."); Ho, *supra* note 7, at 1997–98 ("[My] reanalysis suggests that Sander's conclusions are untenable on their own terms.").
36. *See* LINDA F. WIGHTMAN, LAW SCH. ADMISSION COUNCIL, LSAC NATIONAL LONGITUDINAL BAR PASSAGE STUDY (1998) [hereinafter WIGHTMAN, BAR PASSAGE STUDY] (data on file with author and available with permission of LSAC).
37. *See infra* notes 243, 327 and accompanying text (discussing the problems associated with using tier location as a proxy for school selectivity for the purposes of measuring a mismatch effect).
38. Sander, *Replication of Mismatch*, *supra* note 4, at 87.

mischaracterizes some of Ho's critiques and fails to specifically address others that remain central to his own analyses.[39]

In the aftermath of the Empirical Scholars Brief, rather than retreat from his initial conclusions or significantly qualify them, Sander has doubled-down, frequently attacking the character and credibility of the ESB and its signatories,[40] but largely leaving unaddressed the mistakes identified in Ho's initial critiques, as well as the ESB.[41] As challenges to race-conscious admissions policies are, once again, making their way through the federal courts,[42] it is reasonable to assume that research proclaiming robust support for the mismatch hypothesis, and other alleged harms of race-conscious admissions policies, will find its way to the courts through amicus briefs and expert testimony. It is crucial, then, that empirical research used to support claims by parties on either side of the affirmative action debate adhere to the fundamental precepts of causal inference.[43] But the literature on causal inference is both vast and dense, and consequently, many jurists, legislators, and laypersons interested in understanding both the intended and unintended consequences of affirmative action are ill-equipped to understand the debate—especially when quantitative social scientists on both sides of the debate

39. *See infra* Part II (identifying persistent shortcomings in Sander's analysis of the mismatch hypothesis); *see also* Bjerk, *supra* note 23, at 3–5 (noting that Sander claims that his reanalysis of the Ayres and Brooks and Ho critiques both corrects their methods and provides evidence of large mismatch effects in law school associated with preferential admissions given to black students are unsubstantiated).

40. *See, e.g.*, Brief *Amicus Curiae* for Richard Sander in Support of Neither Party at 28, Fisher v. Univ. of Tex. at Austin (*Fisher II*), 136 S. Ct. 2198 (2016) (No. 14-981) [hereinafter Sander, Brief in *Fisher II*] (calling the Empirical Scholars Brief "laughably inept throughout"); Sander, *Mismatch & the ESB*, *supra* note 30, at 584 ("And the sort of zealotry on display in the [Empirical Scholars Brief] will, increasingly, be recognized for what it is—ideology wearing empirical makeup."); Sander, *Stylized Critique*, *supra* note 8, at 1641, 1657 n.92 (claiming that "[a]nother example of the Zealots avoiding any of the first-order mismatch issues, or the literature finding compelling evidence of them, is the Empirical Scholars Brief," and that these "ideological Zealots . . . cling to conventional affirmative action policies with almost religious fervor and see their attacks on mismatch as a sort of holy war").

41. *See, e.g.*, Empirical Scholars Brief in *Fisher I*, *supra* note 30, at 8 ("Sander's research has major methodological flaws—misapplying basic principles of causal inference—that call into doubt his controversial conclusions about affirmative action.").

42. *See, e.g.*, Students for Fair Admissions, Inc. v. President & Fellows of Harvard, 397 F. Supp. 3d 126 (D. Mass. 2019) (challenging Harvard's race-conscious admissions policies as discriminatory toward Asian Americans).

43. *See, e.g.*, Defendant's Memorandum in Opposition to Plaintiff's Motion for Summary Judgment at 6–15, *Students for Fair Admissions*, 397 F. Supp. 3d 126 (No. 1:14-cv-14176-ADB), ECF No. 435 (defendant claiming that statistical evidence provided by the plaintiff's statistical expert is "fundamentally unreliable" and identifying several errors in the regression analyses).

appear to draw conflicting (though not necessarily equally credible) inferences from the same data.

The purpose of this Article is to lay bare the core requirements of credible causal inference to the uninitiated, highlighting how inattention to (and sometimes outright disregard for) these rules has muddied the debate over mismatch effects in law school and in college admissions, more generally. I attempt to "introduce concepts incrementally so that there are no steep patches or discontinuities in the learning curve,"[44] but I acknowledge, at the outset, this is a tall order. The causal inference literature is quite challenging, in no small part because of both the rapid growth of the field and a lack of consensus on certain issues.[45]

Part I provides the conceptual and theoretical foundation for establishing credible causal claims. Part II offers a thorough critique of the existing empirical literature on the causal effect of affirmative action in legal education. Specifically, six core methodological deficiencies present in much of this research are identified and discussed: (1) posttreatment bias, (2) nonresponse bias, (3) omitted variable bias, (4) interpolation bias, (5) extrapolation bias, and (6) measurement error bias. Part III discusses the most notable critique of the mismatch thesis, the Empirical Scholars Brief, and the responses to that brief. Part IV both explains why attempts to replicate the early findings that were supportive of mismatch have been unsuccessful and highlights research examining alternative theories of racial/ethnic differences in law school performance and employment outcomes. Part V offers a template for the proper investigation of the effect of affirmative action on law school-related outcomes that attempts to avoid or significantly ameliorate the common shortcomings of prior research in this area. The Article concludes by underscoring the importance of the proper vetting of empirical work on legal issues by the scientific community.

44. JOHN THOMPSON, BAYESIAN ANALYSIS WITH STATA (2014) (back cover).

45. *Compare* Michael E. Sobel, *Discussion: 'The Scientific Model of Causality'*, 35 SOCIO. METHODOLOGY 99 (2005), *with* James J. Heckman, *Rejoinder: Response to Sobel*, 35 SOCIO. METHODOLOGY 135 (2005).

For those interested in a general introduction to statistical inference that addresses several of the issues discussed in this Article and is written for a legal audience, I recommend the popular text, *Statistics for Lawyers*. MICHAEL O. FINKELSTEIN & BRUCE LEVIN, STATISTICS FOR LAWYERS (3d ed. 2015). A less formal and less technical introduction to statistical inference is also available. MICHAEL O. FINKELSTEIN, BASIC CONCEPTS OF PROBABILITY AND STATISTICS IN THE LAW (2009).

I. A PRIMER ON ESTABLISHING CAUSAL CLAIMS

Before discussing the specific problems with much of the current research on mismatch theory, it is first important to provide some theoretical background for establishing causal claims.[46] After reviewing the core components of causal inference, I explain how Sander's initial examination of mismatch effects presented in his article, "Systemic Analysis," fails to satisfy these fundamental requirements and how his responses to critics, as well as the responses of other proponents of the mismatch hypothesis, repeat those same mistakes.[47] I also reexamine the approaches Sander and others have adopted in their replies to critics, and avoid several of their key errors. My reanalysis underscores how Sander's statistical models rest heavily on implausible assumptions that violate the basic rules of statistical inference.

"[C]ausality is not something that we can observe, but must be inferred from an association or set of associations. Inference, however, is only possible in the context of a theory that provides a set of assumptions about how various variables are possibly causally related to each other."[48] In other words, causal inference requires that two events are associated (correlated), but a causal connection between the two events does not necessarily follow from their observed association. A necessary condition to establish a causal relationship is that one can reasonably rule out all other reasons that the treatment and control groups systematically differ with respect to an outcome besides the influence of the treatment. Importantly, "[f]airly strong assumptions are needed for the estimates of direct and indirect [or mediated] effects to be interpreted causally."[49] These assumptions can be difficult to understand in

46. For an excellent nontechnical review of the requirements for causal inference, see Antonakis et al., *supra* note 1.

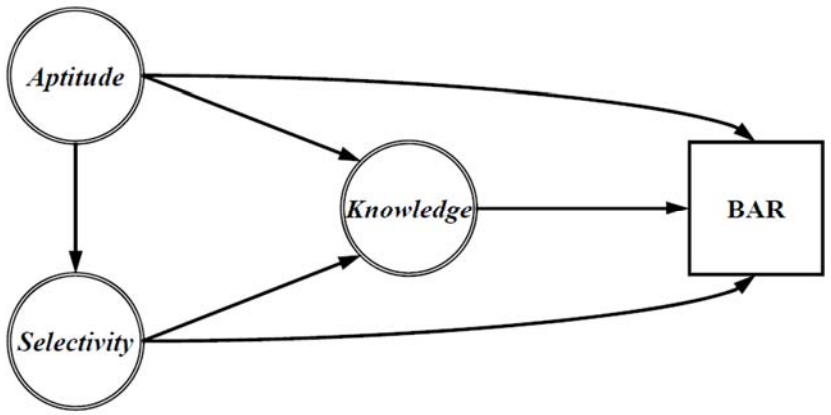
47. *See infra* Parts II–III.

48. Carly R. Knight & Christopher Winship, *The Causal Implications of Mechanistic Thinking: Identification Using Directed Acyclic Graphs (DAGs)*, *in* HANDBOOK OF CAUSAL ANALYSIS FOR SOCIAL RESEARCH 275, 284 (Stephen L. Morgan ed., 2013).

49. Tyler J. VanderWeele, *Mediation Analysis: A Practitioner's Guide*, 37 ANN. REV. PUB. HEALTH 17, 19 (2016). This assumption states that assignment to the treatment or control group is independent of the outcome. In observational studies in which inclusion in the treatment and control are nonrandom, the assumption is one of conditional independence—the independence assumption is satisfied if the assignment is random after adjusting for relevant covariates (also called "conditional exogeneity" or "conditional ignorability"). To be clear, the conditional independence assumption is a necessary, but insufficient condition for causal inference. The two other criteria are (1) regularity of succession (the cause must precede the effect in time), and (2) the two variables are empirically correlated with one other. It is the unconfoundedness assumption, however, that is most frequently violated and the bulk of the literature on causal inference has been devoted to this issue. *See, e.g.*, Antonakis et al., *supra* note 1.

the abstract, so the ensuing discussion attempts to carefully unpack them. The plausibility of the assumptions to establish causal relationships is made clear through causal diagrams which require the analyst to specify all common causes of all variables in the model, therefore providing a visual representation of the theoretical claim about the phenomena of interest.[50] In other words, these diagrams facilitate an understanding of why variables may be associated in the first place.[51]

Figure 1: Sander's Causal Model (Latent Constructs)



50. Perhaps the most popular type of causal diagram is a directed acyclic graph (DAG). *See* JUDEA PEARL, CAUSALITY: MODELS, REASONING, AND INFERENCE 12–13 & fig.1.1(b) (2000). I adopt a slight modification of DAGs in order to make the illustration and ensuing discussion more intuitive.

51. Felix Elwert & Christopher Winship, *Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable*, 40 ANN. REV. SOCIO. 31, 35 (2014) ("The power of [causal diagrams] lies in their ability to reveal all marginal and conditional associations and independences implied by a qualitative causal model. This enables the analyst to discern which observable associations are solely causal and which ones are not—in other words, to conduct a formal nonparametric identification analysis." (citation omitted)).

Figure 2: Sander's Causal Model (Manifest Indicators)

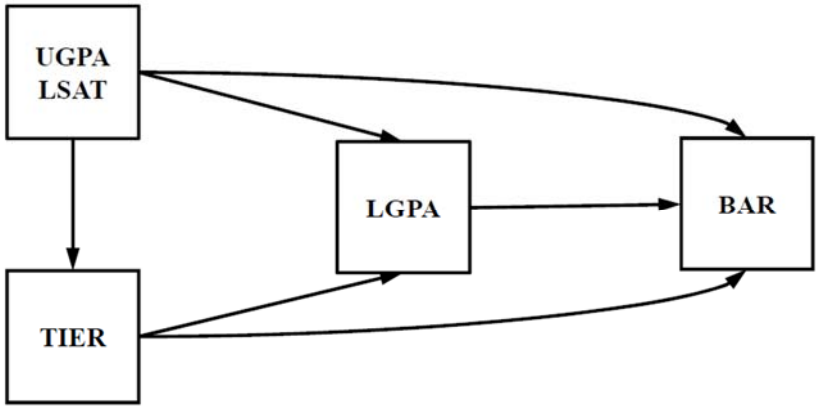
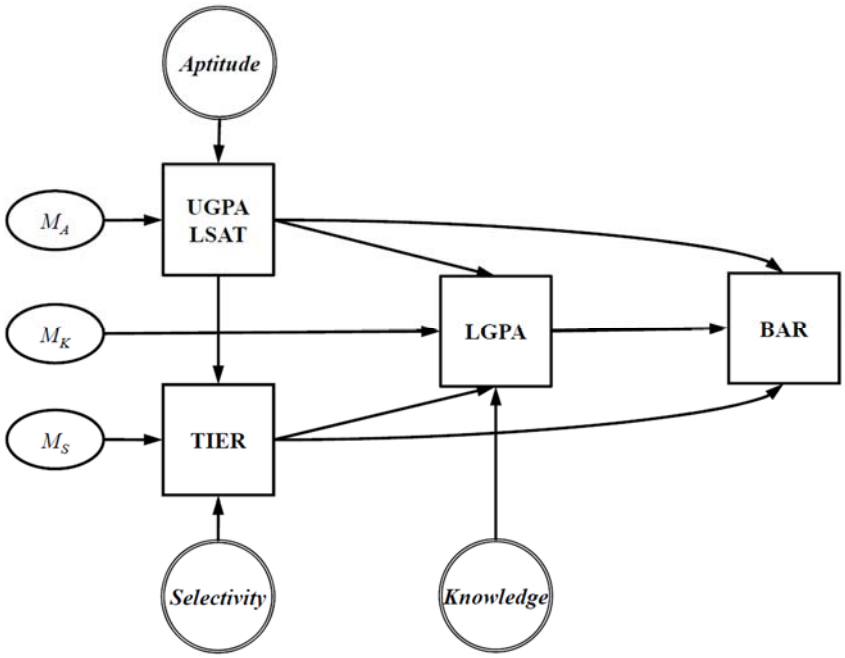


Figure 3: Sander's Causal Model (Latent Constructs & Manifest Indicators)



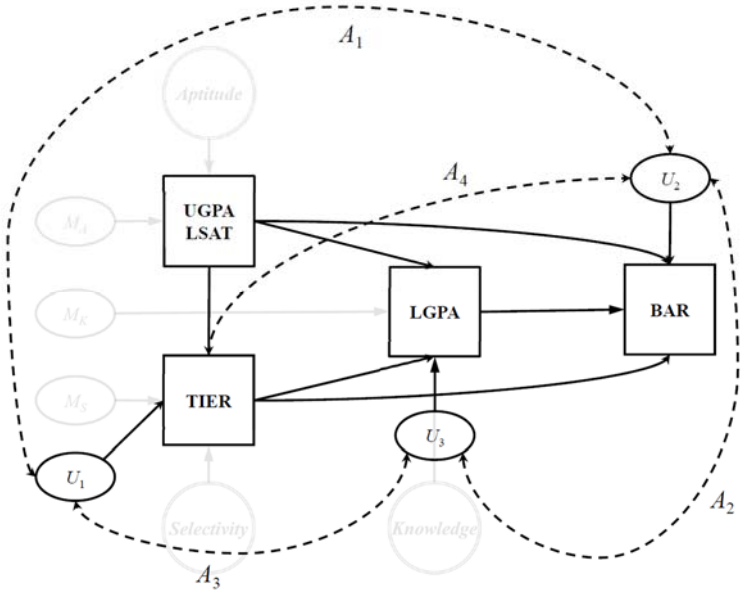
Note: Boxes represent observed variables; circles represent unobserved variables. Straight lines are possible causal effects.

Recall that mismatch theorists posit that beneficiaries of affirmative action in law schools learn less than they would have in the absence of such policies, and as a result, suffer the negative consequences of being outmatched by their peers in law school.[52] Three core relationships purportedly measured in Sander's mismatch analysis are: (1) aptitude for legal training, (2) school selectivity/competitiveness, and (3) acquired legal knowledge/reasoning skills. Figure 1 is a diagram of Sander's proposed mismatch model in "Systemic Analysis." *Aptitude*, *Selectivity*, and *Knowledge* are theoretical (latent) constructs than cannot be measured directly, and are therefore depicted with double-ringed circles. Bar passage, on the other hand, is called a manifest variable (also called a manifest indicator) because it can be directly measured, and is depicted with a box. The diagram in Figure 2 represents the manifest variables (shown in boxes) of undergraduate grade point average (UGPA), Law School Admission Test score (LSAT), school tier location (which is a proxy for law school competitiveness), and law school grade point average (LGPA), which are imperfect indicators of, respectively, aptitude, selectivity, and knowledge. The unidirectional arrow represents a causal relationship: The selectivity/competitiveness of the law school in which a student is enrolled is caused, at least in part, by the student's aptitude for legal training (Figure 1); and, the tier location of the law school where the student matriculates is caused, at least in part, by the student's UGPA and LSAT (Figure 2).[53] Figure 3 combines Figures 1 and 2 and illustrates how the theoretical constructs influence the manifest indicators (via single-direction arrows), as well as how the manifest indicators are related to one another. The small single-ring ovals (M_A, M_S, and M_K) represent measurement errors that impact the manifest constructs. That is, the ovals represent discrepancies between the proposed concept and its empirical referent.

52. *See supra* note 8 and accompanying text (positing three negative consequences of affirmative action: learning mismatch, competitive mismatch, and social mismatch).

53. Although Sander's visual depiction of his causal model incorrectly specifies a noncausal association between UGPA/LSAT and tier location by using a bidirectional arrow, mismatch theory expressly posits a causal association, which is traditionally depicted using a unidirectional arrow. I use the unidirectional arrow to accurately convey the relationship mismatch theory posits. See *infra* note 54 and accompanying text for a more detailed explanation of the choice to use a unidirectional arrow.

Figure 4: Sander's Causal Model



Note: Boxes represent observed variables; circles represent unobserved variables. Straight lines are possible causal effects; curved lines are correlations. Dashed lines indicate spurious relationships that Sander assumes do not exist. A1–A4 identify the specific assumption.

Figure 4 should both assist the reader in developing intuitions about causal inference and clarify the shortcomings of Sander's analytical framework.[54] The latent constructs and their measurement errors (M_A, M_S,

54. Sander formally presents his causal diagram in his response to Ho, Sander, *Mismeasuring*, *supra* note 24, at 2008, but he posits identical causal associations, either explicitly or implicitly, in his initial article, *see* Sander, *Systemic Analysis*, *supra* note 7, at 444. Figure 4 depicts a unidirectional arrow from UGPA/LSAT to tier, whereas Sander's model displays a bidirectional arrow from UGPA/LSAT to tier. Conceptually, Figure 4 is correct because Sander posits a causal relationship: A student's UGPA/LSAT effects the student's tier location. According to the mismatch hypothesis, tier location is endogenously determined, meaning that it is caused, at least in part, by other variables in the model, namely UGPA and LSAT. Sander, *Mismeasuring*, *supra* note 24, at 2008 n.16 (describing his mediation model as "specifying . . . a 'process' in which the variables interact with one another as they shape the eventual outcome").

A bidirectional arrow is conceptually incorrect because it implies either (a) tier location and UGPA/LSAT simultaneously cause one another (known as reciprocal causation) or (b) there is no causal relationship between UGPA/LSAT and tier location (spurious association). The first implication is implausible because tier location cannot affect UGPA/LSAT because these variables occur before matriculation to law school (it violates the regularity of succession requirement). *See supra* note 49. The second implication is that students do not necessarily attend law schools where they might be

and M_K) are "greyed out" from Figure 4 to minimize clutter, but I return to their impact on the analysis of mismatch theory in Subpart II.F.[55] Figure 4 assumes that the manifest indicators are error-free measures of the latent constructs. Sander makes four implausible assumptions (A_1, A_2, A_3, and A_4) which are depicted in Figure 4. Some of these assumptions have been expressly identified and challenged by critics of "Systemic Analysis," but others have not.[56] To date, none of Sander's critics have used causal diagrams to formally evaluate the structure of Sander's causal inferential framework. The circles in Figure 4 (U_1, U_2, and U_3) represent unobserved factors that influence the observed variables (sometimes called "disturbances"). These unobserved variables (U_1, U_2, and U_3) are distinct from the aforementioned measurement errors (M_A, M_S, and M_K), which represent discrepancies between the manifest indicators and the latent constructs those manifest indicators are supposed to capture.[57] The possible causal effects Sander hypothesizes are shown by straight-solid unidirectional arrows (noncausal associations, or correlations, would be depicted by curved-solid dual-ended arrows). The dashed lines represent causal and noncausal relationships that Sander assumes do not affect his analyses and bias his conclusions.

The first assumption is that there are no unobserved common causes of tier location (or more accurately, matriculating to a law school in a specific tier) and the likelihood of first-time bar passage (A_1). Second, there are no unobserved common causes of LGPA and the likelihood of first-time bar

outmatched based on their entering credentials, but rather some other factors outside of UGPA/LSAT explain the relationship between tier location and UGPA/LSAT. If this is the case, then Sander's analytical framework is not even testing the mismatch hypothesis because his model no longer posits that entering credentials independently affect where one attends law school. *See supra* note 49. In addition to this conceptual distinction, properly identifying the interrelationships between the key variables identified by Sander has important implications when examining the plausibility of Sander's assumptions. *See infra* Subpart II.A.2.

55. Error-prone measurements undermine causal inference via the misidentification of the treatment and control groups or the failure to balance the treatment and control group with respect to the confounding variables. *See infra* Subpart II.F.

56. *See, e.g.*, Ho, *Reply to Sander*, *supra* note 23, at 2013 ("Sander now proposes a structural equations model. This represents a different approach based on different assumptions from the original article, whereas the whole point of my reanalysis was to *reduce* the role of unfounded and unnecessary assumptions." (footnote omitted)); *see also* Avidit Acharya, Matthew Blackwell & Maya Sen, *Explaining Causal Findings Without Bias: Detecting and Assessing Direct Effects*, 110 AM. POL. SCI. REV. 512, 518 (2016) (noting that the assumption that no intermediate confounding variables exist is unrealistic for the majority of social science research).

57. *See generally* KENNETH A. BOLLEN, STRUCTURAL EQUATIONS WITH LATENT VARIABLES 319 (1989) (noting that measurement errors pertain to the relationship of observed variables to latent variables, whereas structural errors relate to the influence of variables on each other).

passage (A_2). Third, there are no unobserved common causes of tier location and LGPA (A_3). Lastly, no unobserved common causes of LGPA and first-time bar passage are affected by tier location (A_4).[58]

This final assumption, sometimes called the "no posttreatment confounders" assumption, is especially strong and essentially requires that there are no other unmeasured mediators (intervening causes) linking tier location to bar passage that also influence LGPA. This assumption, A_4, is more plausible if the mediator, LGPA in this case, occurs shortly after the exposure because:

> [i]f a substantial gap exists, then [A_4] requires that nothing on the pathway from the [treatment] to the mediator itself also independently affects the outcome. This assumption likely becomes increasingly less plausible the more time that elapses between the [treatment] and the mediator.[59]

Often "event[s] exist[] in a long chain of causation, and most events have multiple causes. The further back in time we go, the more causes that must be held constant."[60] Based on Sander's own explanatory framework, A_4 is unlikely to hold because of the temporal distance between what is learned in school (proxied by LGPA) and when a student takes the bar exam.[61] A_1 and A_3 are

58. While it may not be immediately obvious from Figure 4, A_4 incorporates A_2 & A_3 because of the TIER→LGPA→BAR causal chain. For simplicity, I omit the direct effect of TIER on BAR in this discussion, but that direct effect is properly depicted in Figure 4. The first part of the chain (TIER→LGPA) implies that U_1 and U_3 are uncorrelated (A_3), and the second part of the chain (LGPA→BAR) implies that U_2 and U_3 are uncorrelated (A_2). A_4 posits that the transmission of the causal effect of TIER on BAR is through law school grade point average (LGPA) and not some other unobserved mediator that also impacts the second part of the chain. One can easily depict the causal relationships that violate A_4 as a series of causal chain segments: TIER→U_2, U_2→LGPA, and U_2→BAR. Also note that because violation of A_4 also implies a violation of A_2, then U_2=U_3. It should be clear that A_4 implies there are no back door paths through additional mediating variables. Alternatively, one could depict A_4 in Figure 4 with a dotted straight line from TIER to U_3 and a solid straight line from U_3 to BAR.

To justify a causal interpretation of a mediated effect, the mediator must be isolated, but it need not be exhaustive. An isolated mediating variable is a variable that is not influenced by any other mediators; on the other hand, an exhaustive mediating variable is the only variable through which the treatment variable affects the outcome. The key advantage of an exhaustive mediator is that it permits the analyst to assess the full effect of the treatment variable, whereas an isolated mediator can only assist in identifying the portion of the effect of the treatment that is transmitted through that particular mediator (partial causal effect). STEPHEN L. MORGAN & CHRISTOPHER WINSHIP, COUNTERFACTUALS AND CAUSAL INFERENCE: METHODS AND PRINCIPLES FOR SOCIAL RESEARCH 229–230 (2007).

59. VanderWeele, *supra* note 49, at 21.

60. JOSEPH S. NYE, JR., UNDERSTANDING INTERNATIONAL CONFLICTS: AN INTRODUCTION TO THEORY AND HISTORY 51 (4th ed. 2003).

61. *See infra* note 457 and accompanying text (describing conceptual flaws in Sander's explanation that mismatch only influences first-time bar passage, but not eventual bar passage).

necessary assumptions in observational studies for the proper estimation of total causal effects; that is, if students were randomly assigned to law schools, then A_1 and A_3 would be satisfied. This obviously is not the case with the BPS data, so A_1 and A_3 will only hold if Sander's models adjust for all potential pretreatment confounders (overlapping variables included in U_1 & U_2 and U_1 & U_3). Adjusting for all pretreatment confounders is typically referred to as conditional exogeneity/ignorability.[62] Assuming for the sake of argument that students were randomly assigned to each tier (or, alternatively, Sander's analysis satisfied the conditional exogeneity/ignorability requirement), Sander's mediation analysis framework does not necessarily satisfy assumptions A_2 and A_4.[63] This is because even if the treatment is randomized (or conditionally independent of confounders after statistical adjustment), the mediator generally will not be.[64] As Jacob M. Montgomery and colleagues explain:

> Conditioning on posttreatment variables eliminates the advantages of randomization because we are now comparing *dissimilar* groups. . . . [C]oncerns about posttreatment bias are not really (or only) about the posttreatment variable itself. The problem is that by conditioning on a posttreatment variable, we have unbalanced the treatment and control groups with respect to *every other possible confounder*.[65]

Violations of any of these assumptions, A_1–A_4, undermine an analyst's ability to make credible causal claims by injecting bias into the identification

62. Properly controlling for these pretreatment confounders is also called "back-door criterion" for causal inference because confounding pretreatment variables provide an indirect, noncausal channel along which associations can occur—that is, these confounding pretreatment variables allow spurious associations through the back door. PEARL, *supra* note 50, at 79.

63. The isolation of causal effects through adjusting for mediating variables—either isolated or exhaustive—is also referred to as "front door criterion." *Id.* at 81.

64. *See* TYLER J. VANDERWEELE, EXPLANATION IN CAUSAL INFERENCE: METHODS FOR MEDIATION AND INTERACTION 64 (2015) ("Mediator-outcome confounding can be present even if the exposure is randomized (since the mediator is not randomized."); D. James Greiner, *Causal Inference in Civil Rights Litigation*, 122 HARV. L. REV. 533, 564 (2008) ("[R]andomization balances variables that might have a role in determining the potential outcomes, subject to an important set of exceptions: those variables that are themselves affected by the treatment. This is as it should be, as analysts do not ordinarily want balance in variables affected by treatment."); Ho, *supra* note 7, at 2000 n.16 (explaining that even if the treatment is randomly assigned in a classic experiment, adjusting for a concomitant variable would bias the measurement of a causal effect).

65. Jacob M. Montgomery, Brendan Nyhan & Michelle Torres, *How Conditioning on Posttreatment Variables Can Ruin Your Experiment and What to Do About It*, 62 AM. J. POL. SCI. 760, 762–63 (2018). I provide a detailed discussion of this aspect of Sander's mediation analysis in Subpart II.A.1.a, *infra*, describing how these assumptions can be tested and demonstrating that Sander fails to satisfy them.

of causal effects. This should make clear, then, that careful causal inference requires just as much care about potential unobserved relationships as it does about posited observed relationships. In the next Part, I discuss how each of these assumptions of causal inference has been routinely ignored (or inadequately assessed) in research on mismatch effects.

II. UNPACKING THE METHODOLOGICAL CRITIQUES

The academic mismatch hypothesis is not new. Earlier versions have been advanced by opponents of race-based affirmative action almost from the time such programs began and it has been a reoccurring theme of arguments against affirmative action ever since.[66] Proponents of affirmative action have routinely acknowledged that mismatch theory should be taken seriously because it posits facially plausible mechanisms through which affirmative action might harm its intended beneficiaries.[67] In fact, four of Sander's earliest critics stated:

> Indeed, we take the problem so seriously that despite the high value we place on racial diversity within law schools, the four of us would not support affirmative action as currently practiced in law school admissions if we believed that employing race-neutral admissions criteria would in fact lead to a net increase in the number of African Americans passing the bar.[68]

But these same scholars also cautioned that "[t]aking the [mismatch] hypothesis seriously does not, however, mean [uncritically] accepting it. Rather, it requires putting it to empirical tests."[69] But "while Sander has appropriately forced us and others to take a hard look at the actual workings of affirmative action . . . [his] conclusions in *Systemic Analysis* rest on a series of statistical

66. Cheryl I. Harris & William C. Kidder, *The Black Student Mismatch Myth in Legal Education: The Systemic Flaws in Richard Sander's Affirmative Action Study*, J. BLACKS HIGHER EDUC., Winter 2004/2005, at 102, 102 ("These arguments are not new. Many will recall earlier versions of the 'mismatch' thesis advanced by conservatives . . . with respect to affirmative action and college graduation rates. Time and time again the basic premises of these claims [about affirmative action and mismatch] have been refuted.").

67. *See, e.g.*, Richard Lempert, *Growing Up in Law and Society: The Pulls of Policy and Methods*, 9 ANN. REV. L. & SOC. SCI. 1, 19 (2013) ("[I]t is no accident that Professor Sander's mismatch thesis has received considerable public attention. His claims are intuitively plausible and have great appeal for those opposed to affirmative action.").

68. David L. Chambers, Timothy T. Clydesdale, William C. Kidder & Richard O. Lempert, *The Real Impact of Eliminating Affirmative Action in American Law Schools: An Empirical Critique of Richard Sander's Study*, 57 STAN. L. REV. 1855, 1857 (2005).

69. Kidder & Lempert, *Mismatch Myth*, *supra* note 22, at 105–06; *see also* Ho, *Reply to Sander*, *supra* note 23, at 2016 ("The empirical investigation of affirmative action is important and should be subjected to scientific scrutiny.").

errors, oversights, and implausible assumptions."[70] Theories will inevitably encounter negative or disconfirming evidence; therefore, it is only after continued rigorous empirical testing that a theory's validity can truly be known: "[T]he sole purpose of [scientific explanation] was to establish as firmly as possible the surviving theory . . . [as] the true one . . . [that is,] the best tested theory . . . of which we know."[71]

The empirical analyses in "Systemic Analysis," as well as those in *Mismatch* are plagued by several significant shortcomings that Sander must fully address before claiming any empirical support for mismatch theory. Several of these shortcomings have been identified by critics of Sander's early work on mismatch, though often incompletely explained, so my goal is to sufficiently clarify them, as well as identify additional deficiencies that have gone unaddressed, in order to underscore their relevance not only to Sander's work, but to scholarship on mismatch, more generally.[72] The deficiencies impacting Sander's analyses, many of which are discussed in introductory econometrics textbooks,[73] fall into six categories: (1) posttreatment bias, (2) nonresponse bias, (3) omitted variable bias, (4) interpolation bias, (5) extrapolation bias, and (6) measurement error bias.[74]

70. Chambers et al., *supra* note 68, at 1857.
71. POPPER, *supra* note 5, at 419 (emphasis omitted).
72. *Cf.* Empirical Scholars Brief in *Fisher I*, *supra* note 30, at 17 ("Much of the cited research has been in the law school context. Amici therefore focus the rest of their arguments on the methodological flaws contained in Sander's and economics professor Doug Williams's law school mismatch research . . . although the same methodological challenges also affect mismatch research in other settings.").
73. *See, e.g.*, DAMODAR N. GUJARATI, BASIC ECONOMETRICS 508–10 (4th ed. 2003) (discussing specification errors, five of which are present in Sander's work, and their consequences for causal inference).
74. See *infra* Table 1 for brief descriptions of these biases. Posttreatment bias and nonresponse bias can be viewed as special cases of "collider bias" (also called endogenous selection bias), which results from improperly adjusting for a common effect of two variables, and the offending variable (referred to as a "collider") is either on or off the causal pathway from the treatment to the outcome. Elwert & Winship, *supra* note 51, at 35–36. Collider bias and confounding bias (failing to adjust for a common cause of a variable) are closely related because adjusting for a collider variable can induce confounding bias by creating a spurious relationship between variables that did not exist before adjusting for the collider. On-causal pathway collider bias, in the form of concomitant variable adjustment, and off-causal pathway collider bias, in the form of nonrandom sample selection, impact research on mismatch effects in law school. Bias from nonresponse—missing data—also impacts analyses of mismatch effects in law schools and may be on-pathway and off-pathway. It must be noted, however, that the terminology in the causal inference literature with respect to collider variables has been inconsistent, and some scholars label mediator variables as collider variables, while others do not. Colliders can be both pretreatment and posttreatment variables, and there is an ongoing debate in the causal inference literature concerning the tradeoffs between controlling for pretreatment variables that may be both confounders and colliders

Each of these biases may significantly impair the identification of causal processes by undermining the analyst's ability to compare cases that are equivalent along all relevant dimensions save the variable of the interest. Which is to say that these biases may produce a spurious relationship between the key explanatory variable and the outcome. In the discussion that follows, I discuss each, in turn.[75]

Table 1: Biases Negatively Impacting Causal Inference

Type	Definition
Posttreatment Bias	Relationships between variables are distorted when (a) adjusting for an intermediate effect of the treatment or (b) subsetting on an intermediate effect of the treatment
Nonresponse Bias	Relationships between variables are distorted when values on variables are systematically missing
Omitted Variable Bias	Relationships between variables are distorted when they have unobserved common causes
Interpolation Bias	Relationships between variables are distorted when incorrectly specifying how they covary within the observed range of the data
Extrapolation Bias	Relationships between variables are distorted when incorrectly specifying how they covary outside the observed range of the data
Measurement Error Bias	Relationships between variables are distorted when (a) the treatment and control groups are misidentified; (b) the treatment dosage is misjudged; or (c) variables are systematically mismeasured

(because colliders are path-specific). *See id.* at 45–48; Guido W. Imbens, *Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics* 46–49 (Nat'l Bureau of Econ. Rsch., Working Paper No. w26104, 2009), https://ssrn.com/abstract=3428157 [https://perma.cc/2YW4-ZK8B]. In the interest of space, this Article focuses on collider bias resulting from concomitant variable adjustment, nonrandom sample selection, and nonresponse, but analysts of mismatch effects should also examine other potential instances of collider bias.

75. As Coryn Bailer-Jones explains succinctly:

> Science is fundamentally about learning from data, and doing so in the presence of uncertainty. Uncertainty arises inevitably and avoidably in many guises. It comes from noise in our measurements: we cannot measure exactly. It comes from sampling effects: we cannot measure everything. It comes from complexity: data may be numerous, high dimensional, and correlated, making it difficult to see structures.

CORYN A. L. BAILER-JONES, PRACTICAL BAYESIAN INFERENCE: A PRIMER FOR PHYSICAL SCIENTISTS 1 (2017).

A. Posttreatment Bias

Violations of assumptions A_2 and A_4 result in posttreatment bias, which comes in two forms. One form occurs when the analyst "controls" for the consequence of a cause. Adjusting for the effect of a causal variable can bias the estimate of the treatment effect in either direction ("concomitant variable adjustment bias"). The other source of posttreatment bias is conceptually equivalent and occurs when the analyst drops observations or subsets observations based on posttreatment criteria ("nonrandom sample selection bias"). Below, I describe each type of bias and how it undermines Sander's analyses of mismatch effects.

1. Concomitant Variable Adjustment Bias

"[F]rom the perspective of applied researchers, the problems associated with conditioning on posttreatment variables can be extremely serious."[76] Sander's prior analyses of mismatch suffer from this type of bias because he posits that LGPA is influenced by tier location, and then adjusts for LGPA when analyzing the impact of tier location on bar passage rates. As the Empirical Scholars Brief explains:

> Adjusting for such outcomes (rather than pre-existing characteristics), as the Sander studies do, contaminates inferences about the causal effect of law-school tier. Indeed, controlling for graduation and grades leads Sander [in prior work] to claim that there is no economic return to attending an elite law school at all, regardless of ethnicity.[77]

---

76. Acharya et al., *supra* note 56, at 514–15; *accord* 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).

According to causal inference expert Donald Rubin, "the concepts of direct and indirect causal effects are generally ill-defined and often more deceptive than helpful." Donald B. Rubin, *Direct and Indirect Causal Effects via Potential Outcomes*, 31 SCANDINAVIAN J. STAT. 161, 162 (2004).

77. Empirical Scholars Brief in *Fisher I*, *supra* note 30, at 23–24 (emphasis omitted) (citing Richard H. Sander & Jane Yakowitz Bambauer, *The Secret of My Success: How Status, Eliteness, and School Performance Shape Legal Careers*, 9 J. EMPIRICAL LEGAL STUD. 893, 915–17 (2012). Sander and Jane Bambauer did find some support for the impact of law school eliteness on LGPA after controlling for law school grades, but only for a small subset of schools.

Other work conducted by Sander purporting to explore various aspects of race-conscious admissions practices has also come under significant criticism from social scientists for reporting anomalous, inaccurate, and misleading results. *Compare* Richard Sander, The Consideration of Race in UCLA Undergraduate Admissions 1 (Oct. 20,

Sander has denied the presence of posttreatment bias arising from adjusting for LGPA in his models predicting bar passage in "Systemic Analysis" and has responded to these criticisms with two different approaches. First, he attempts to assess the direct and indirect (through LGPA) effects of tier location on first-time bar passage focusing solely on black students, and claims that he finds no evidence of posttreatment bias.[78] I refer to this as his mediation analysis because he purports to assess the degree to which LGPA mediates the relationship between tier location and first-time bar passage.[79] Second, he claims to conduct "more than a dozen different tests to examine the degree [of posttreatment] bias . . . . [and concludes that] [e]ach of these tests found zero evidence of bias."[80] I label this approach his "sequential analysis" because he begins with a trimmed model with only a few variables and gradually includes additional variables (and their interactions) to assess how much of the effect of those variables changes as more variables are included in the model.[81] Both of Sander's approaches fail to address the central problem of adjusting for a concomitant variable. Sander's reanalysis simply reproduces the same biases that undermined his initial conclusions.[82]

2012) (unpublished manuscript), http://www.seaphe.org/pdf/uclaadmissions.pdf [https://perma.cc/8NQW-R2YE] (claiming that UCLA illegally considers race in admissions decisions at the undergraduate level), *with* David Stern, Are There Racial Disparities in UCLA Freshman Admissions? 1 (Nov. 23, 2012) (unpublished manuscript) (explaining that "Professor Sander's results contain apparent anomalies" and that Sander improperly "conflat[es] results for all colleges [and] overlooks some important differences among them"), *and* Richard Lempert, Observations on Professor Sander's Analysis of the UCLA Holistic Admissions System 12 (Nov. 19, 2012) (unpublished manuscript) (noting numerous problems with the manner in which Sander presents and interprets his data and concluding that he "find[s] little in [Sander's] paper that stands up to close scrutiny").

78. Sander, *Mismeasuring*, *supra* note 24, at 2008 (claiming that his analysis "bear[s] out [his] original regressions and undercut[s] any claim of bias").

79. *Id.* at 2007 ("One of the most elegant ways to show this point [that going to a more elite law school lowers one's expected LGPA] is with a structural equation model—a type of analysis specifically developed to deal with situations in which one is concerned about independent variables indirectly affecting one another. Structural equation models allow us to directly measure those indirect effects." (footnote omitted)).

80. *Id.* at 2007 (citing Richard H. Sander & Joseph W. Doherty, Supplemental Notes on the Relative Effect of Law School Grades and Law School Prestige Upon Bar Passage (Apr. 27, 2005) (unpublished manuscript) (on file with the *UCLA Law Review*)).

81. Technically speaking, both Sander's mediation and sequential analysis attempt to assess the role of LGPA in mediating the relationship between law school selectivity and first-time bar passage—that is, LGPA transmits the causal effect from the treatment to the outcome—so both frameworks can be labeled mediation analyses. I use different labels for Sander's two approaches in an effort to minimize potential confusion.

82. The deficiencies in Sander's response to Ho underscore another concerned raised by King: Many reoccurring "problems [in quantitative social science] are more than

Moreover, even when accepting Sander's methodological frameworks on their own terms, a closer inspection of his analyses reveals that they do not support the very conclusions he describes.

Measuring the portion of the effect of a treatment variable that is transmitted to the outcome variable through mediator variables is likely to result in biased estimates of mediation effects when mediators are not directly manipulated through an experimental design.[83] "Warnings about this problem have been issued for decades by statisticians, psychologists, and political scientists."[84] Formally, if the mediator and the outcome variable share common causes (confounders), randomization/confounder adjustment based on the treatment does not balance those mediator-outcome confounders, and adjusting for the mediator will bias the estimated causal effect. This stems from the fact that adjusting for the mediator in the analysis does not physically disable (block) the direct path from the treatment going through the mediator; it merely matches samples with equal values of the mediators (meaning that it breaks randomization), and thus induces spurious correlations among other variables in the analysis. "Once we start reasoning about direct and indirect effects, we are considering the effects of not only the exposure but also the mediator as well. . . . [F]ailure to control for mediator-outcome confounding in a randomized trial can substantially bias estimates of direct and indirect effects."[85] Substantive knowledge of potential mediator-outcome confounders is necessary to identify these spurious correlations, and the analyst must attempt to adjust for those mediator-outcome confounders.[86]

For readers unfamiliar with the rules of causal inference, the preceding discussion might not seem especially intuitive at first, so it is important to clarify how controlling for a concomitant variable would break randomization and could induce bias. For the sake of simplicity, assume LGPA is a binary variable (either high or low) and the statistical model adjusts

technical flaws; they often represent important theoretical and conceptual misunderstandings." King, *supra* note 3, at 666.

83. *See* John G. Bullock & Shang E. Ha, *Mediation Analysis Is Harder Than It Looks*, *in* CAMBRIDGE HANDBOOK OF EXPERIMENTAL POLITICAL SCIENCE 508 (James N. Druckman et al. eds., 2011); Kosuke Imai, Luke Keele & Teppei Yamamoto, *Identification, Inference and Sensitivity Analysis for Causal Mediation Effects*, 25 STAT. SCI. 51, 61 (2010) (noting that the strong assumptions of causal mediation analysis may not hold in many social scientific studies because even if the treatment is randomized, the mediator is not).

84. Bullock & Ha, *supra* note 83, at 508.

85. VanderWeele, *supra* note 49, at 21.

86. *Id.*

for LGPA when examining the effect of tier location on bar passage rates.[87] This entails comparing the likelihood of bar passage of individuals with low LGPAs who are mismatched with individuals with low LGPAs who are not mismatched. It also entails comparing individuals with high LGPAs who are mismatched to individuals with high LGPAs who are not mismatched. Recall that mismatch theory posits that mismatched individuals and nonmismatched individuals should not have the same GPA, so if affirmative action does lead to mismatch, these groups are likely to be dissimilar along important dimensions that are not accounted for in the model. The nonmismatched/low LGPA group will consist of individuals with unobserved characteristics that make them most likely to "learn less" in law school and fail the bar, compared to the mismatched/low LGPA group, who are less likely to possess the same unobserved characteristics; the mismatched/high LGPA group will consist of individuals with unobserved characteristics (such as study habits) that make them most likely to "learn more" in law school and pass the bar (compared to the nonmismatched/high LGPA group). As a result, Sander's models that adjust for LGPA almost certainly induce confounding that violates A_2.[88]

Also recall that A_4 requires that there is no effect of the treatment that confounds the mediator-outcome relationship (in other words, there are no unobserved mediators of tier location-bar passage relationship that influence LGPA or share common causes), and this assumption must hold for all of the mediators.[89] Unmeasured common causes of the LGPA–bar passage relationship will bias estimates of both the direct and indirect effects of tier location.[90] And these potential biases are in addition to the biases resulting from nonrandom assignment of tier location (violations of A_1 and A_3). The importance of properly accounting for bias resulting from unobserved factors in causal mediation analysis was underscored by Donald Rubin: "A more successful general approach [to mediation analysis] is to collect and use

87. Obviously LGPA is a continuous, not binary, variable. This distinction does not change the analysis, although the ensuing example becomes more cumbersome if one decides to separate LGPA into additional categories. A covariate adjustment entails partitioning the population into groups that are homogeneous relative to the covariate (such as LGPA), assessing the effect of tier location on bar passage in each homogeneous group, and then averaging the results. *See* PEARL, *supra* note 50, at 78.

88. *See* Acharya et al., *supra* note 56, at 514–15.

89. *See* VANDERWEELE, *supra* note 64, at 114.

90. *See id.* at 26. This might not seem obvious at first, but remember that mediation analysis is essentially about the treatment (law school tier) changing the mediator (LGPA), and the change in the mediator, in turn, changing the likelihood of bar passage (and other relevant outcomes, such as employment opportunities, etc.).

covariates that are predictive of both the intermediate potential outcomes [here, LGPA] and the final potential outcomes [here, bar passage]."[91]

Mediation analysis has gained considerable promise in the social sciences,[92] and when properly conducted, assists in making credible causal claims. Sander, however, fails to acknowledge that his approach rests on the highly implausible assumption that the association between tier location and bar passage could only be due to the hypothesized mechanisms of mismatch (which, according to Sander, is proxied by LGPA).[93] Sander's mediation analysis only takes into account five explanatory variables[94]—tier location, UGPA, LSAT, sex/gender, and LGPA—although nearly two hundred pretreatment variables are available in the BPS data.[95] In other analyses, Sander includes family income and part-time enrollment status, but his models fall short of accounting for the wide range of information present in the BPS data that would be predictive of intermediate and final outcomes.[96] The poor predictive ability of his statistical models examining the likelihood of passing the bar underscores this point.[97]

Sander's claims that "Ho does not run any of the standard tests to detect bias; he assumes that it exists and that it is fatal to the model"[98] and that "Ho's critique has inspired [Sander] to run more than a dozen different tests to examine the degree to which bias . . . might exist"[99] are particularly telling because they reveal that Sander understands neither the direct relevance of Ho's reanalysis to identifying the extent of bias in "Systemic Analysis" nor the inadequacy of his own standard tests to address the problems identified by Ho. Simply stated, Ho both avoids unwarranted modeling assumptions[100] and adheres to Rubin's aforementioned advice with respect to mediation analysis by "collect[ing] and us[ing] covariates that are predictive of both the intermediate potential outcomes [here, LGPA] and the final potential outcomes [here, bar

91. Donald B. Rubin, *Causal Inference Through Potential Outcomes and Principal Stratification: Application to Studies With "Censoring" Due to Death*, 21 STAT. SCI. 299, 305 (2006).
92. *See* Knight & Winship, *supra* note 48, at 275.
93. *See, e.g.*, *id.* at 284 ("The possible correspondence between a mechanism and an observed association does not imply causality unless it can be demonstrated that the association could *only* be due to the hypothesized mechanism.").
94. Sander, *Mismeasuring*, *supra* note 24, at 2008.
95. *See infra* Subpart II.C (describing variables included in the BPS).
96. *See* Sander, *Systemic Analysis*, *supra* note 7, at 439.
97. *See infra* Subpart II.C (discussing the poor predictive ability of Sander's models examining bar passage); *infra* Appendix A.4 (same).
98. Sander, *Mismeasuring*, *supra* note 24, at 2006–07.
99. *Id.* at 2007.
100. *See infra* Subpart II.D.

passage]."[101] After engaging in empirically-driven (rather than purely model-driven) counterfactual analysis[102] and taking into account as few as the four pretreatment variables that Sander examines, and as many as 170 additional pretreatment variables included in the BPS data, Ho discovers that tier location has no effect on passing the bar exam.[103] The key feature of Ho's analytical framework Sander misunderstands is that Ho measures the total effect of the treatment variable (tier location) without attempting to disaggregate this effect into its component parts.[104] Sander's analysis purports to measure the effects of these components of the total tier location effect—the direct and indirect effects—on first-time bar passage. A direct effect represents the part of the total effect that is not transmitted through the intervening variables, whereas indirect effect is the component of the total effect that is transmitted through intervening factors.[105] Adequately measuring the total (or composite) causal effect requires fewer assumptions than the direct and indirect effects because randomization does not guarantee the validity of direct and indirect effects. "Even when causal relationships are firmly established, demonstrating the mediating pathways is far more difficult—practically and conceptually—than is usually supposed."[106] Specifically, a total effect has a causal interpretation when A_1 is satisfied (there is no treatment-outcome confounding) and there exists regularity of succession (the cause precedes the effect in time). In contrast, the separate calculation of direct and indirect effects "requires an explicit specification of

101. *See* Rubin, *supra* note 91, at 305; *see also infra* Subpart II.C.
102. *See infra* Subpart II.E.
103. Ho, *supra* note 7, at 2002–03 ("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. . . . 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." (footnote omitted)); Ho, Evaluating Affirmative Action, *supra* note 23, at 7–9 ("Sander controls for only a few covariates . . . . I thereby collect more data to reassess the effect [sic] of law school tier. . . . We match on 180 of these pretreatment covariates to assess the effect of a top-tier school. . . . Clearly, top-tier students differ in a host of characteristics not accounted for by Sander . . . .").
104. *See* VANDERWEELE, *supra* note 64, at 21–22 ("[T]he total effect . . . represents simply the overall effect of exposure or treatment on the outcome . . . [but] to have a causal interpretation as direct and indirect effects, fairly strong assumptions about confounding need to be made.").
105. VanderWeele, *supra* note 49, at 18–19, 22.
106. Donald P. Green, Shang E. Ha & John G. Bullock, *Enough Already About "Black Box" Experiments: Studying Mediation Is More Difficult Than Most Scholars Suppose*, 628 ANNALS AM. ACAD. POL. & SOC. SCI. 200, 202 (2010).

additional intervening and mediating variables."[107] So, in addition to A_1, the analysis must also satisfy A_2 (no mediator-outcome confounding), A_3 (no treatment-mediator confounding), and A_4 (no mediator-outcome confounder that is impacted by the treatment).[108] In responding to Sander's mediation analysis, Ho specifically notes, "[T]he whole point of my reanalysis was to *reduce* the role of unfounded and unnecessary assumptions. Sander's discussion fails to do precisely what my Comment aimed to achieve, namely, to clarify the assumptions in substantively meaningful ways in order to validly infer a causal effect."[109]

Next, I provide a closer examination of Sander's mediation analyses and demonstrate that his conclusions are unreliable and invalid due to both strong reliance on implausible assumptions and obvious errors of interpretation. When these errors are corrected, Sander's results fail to survive even modest (and obvious) forms of statistical scrutiny. This criticism also holds for Sander's sequential analysis, although Sander purports to assess the potential impact of posttreatment bias and claims that there is "zero evidence of bias."[110] In the interest of space, I examine Sander's mediation analysis in the main body of the text and provide a thorough discussion of his sequential analysis in Appendix B.[111]

---

107. MORGAN & WINSHIP, *supra* note 58, at 223.

108. VanderWeele, *supra* note 49, at 19.

109. Ho, *Reply to Sander*, *supra* note 23, at 2013. It is well known that a treatment variable may have no total causal effect on an outcome because of opposing direct and indirect effects that counterbalance each other's influence. For some research questions, properly measured direct and indirect effects may still be of substantive interest—for example, when there is little to no overall association between the treatment variable and the outcome, but there are strong offsetting direct and indirect effects. None of these conditions would appear to apply for mismatch theory because: (1) the total effect of the policy intervention of race-based affirmative is the question of substantive interest; (2) the polychoric correlation between tier location and first-time bar passage in the BPS is substantial and statistically significant (*rho* = 0.20, std. err. = 0.02); and (3) according to Sander's own model specification, both the direct and indirect effects of tier location on the probability of first-time bar passage are extremely weak (respectively, 4 and -6.4 percent).

A polychoric correlation properly takes into account the differing measurement scales of tier location (ordinal) and first-time bar passage (binary). *See* Karl G. Jöreskog, *On the Estimation of Polychoric Correlations and Their Asymptotic Covariance Matrix*, 59 PSYCHOMETRIKA 381 (1994).

110. *See supra* note 80.

111. Sander's sequential analysis is wholly inadequate to properly investigate the influence of posttreatment bias. Furthermore, even evaluated on its own terms, his analysis does not support his conclusions because it suffers from at least three core flaws: (1) incorrect interpretation of the marginal effect of tier location; (2) miscalculation of the standard errors and associated tests of statistical significance; and (3) miscalculation of the marginal effects of race/ethnicity. *See infra* Appendix B.

It is worth noting that, recently, Sander has retreated from his initial claims that his analyses demonstrated that affirmative action causes blacks to fail the bar exam. Specifically, he now argues that "Ho's observation [of posttreatment bias] would be well taken if my [analyses] were a causal test of mismatch, but it is not. . . . I was not putting forth a causal model, but rather demonstrating a paradox that can be explained by the mismatch mechanism. The post-treatment bias critique is simply not relevant."[112] But Sander's statement is directly at odds with his earlier attempts to both advance a causal story linking race-conscious affirmative action to bar passage and employment outcomes and defend his work against criticisms of the inappropriateness of his analyses for causal inference.[113] For example, he expressly claims to:

112. Sander, *Replication of Mismatch*, *supra* note 4, at 87. *But cf.* POPPER, *supra* note 5, at 248–52 (writing that scientific explanation entails assessments of the degree of verification, which require the theory to be falsifiable, and not assessments of the degree of corroboration, which simply focus on supportive evidence); Robert K. Merton, *supra* note 5, at 468 (criticizing *post factum* explanations as unscientific because they remain at the level of plausibility (low evidentiary value) and do not lend themselves to falsification).

113. *E.g.*, SANDER & TAYLOR, *supra* note 4, at 58–60 ("Blacks were doing badly on the bar not . . . because law schools were somehow unwelcoming. . . . They were doing badly . . . because the law schools were killing them with kindness by extending admissions preferences (and often scholarships to boot) . . . ."); Sander, *Systemic Analysis*, *supra* note 7, at 447 ("[R]acial preferences in law school admissions significantly worsen blacks' individual chances of passing the bar by moving them up to schools at which they will frequently perform badly. I cannot think of an alternative, plausible explanation. If there were any other factor that somehow disadvantaged blacks—e.g., if blacks had more trouble affording bar-preparation classes and were therefore more likely to go it alone—then this would make being black an independently significant causal factor in bar passage rates. But it is not."); Sander, *Mismeasuring*, *supra* note 24, at 2010 ("[I]t is now *undeniable* that this system [of race-based affirmative action] . . . is producing grossly unequal results" (emphasis added)); Sander, *Mismatch & the ESB*, *supra* note 30, at 575 (explicitly defending Williams's causal analysis of mismatch effects in law school, using the first choice framework); Sander, Brief in *Fisher II*, *supra* note 40, at 19 (claiming that large racial preferences resulting from affirmative action produce, *inter alia*, learning mismatch and provide "a straightforward, direct causal link between the preferences and the effects"); Sander, *Stylized Critique*, *supra* note 8, at 1643 ("[M]ismatch is likely to produce poor grades and discouragement. In the case of science mismatch, it seems to cause the vast majority of science-interested students (if they receive large preferences) to abandon STEM fields."); Brief Amicus Curiae for Richard Sander in Support of Petitioner at 19, 22, Schuette v. Coal. to Defend Affirmative Action, 134 S. Ct. 1623 (2014) (No. 12-682) [hereinafter Sander, Brief in *Schuette*] (stating that racial preferences "cause students to receive lower grades," and that students who receive large preferences "appear to actually learn less than otherwise similar students attending less elite law schools, as evidenced by the latter students' much higher bar passage rates"); Richard H. Sander, *Listening to the Debate on Reforming Law School Admissions Preferences*, 88 DENV. U. L. REV. 889, 937 (2011) [hereinafter Sander, *Listening to the Debate*] ("According to mismatch theory, the explanation lies in the fact

> [D]irectly measure those indirect effects [of mismatch by] . . . specifying and estimating the interrelationships among independent variables by modeling them as a "process" in which the variables interact with one another as they *shape* the eventual outcome. This model is specified to estimate whether law school grades "suppress" the impact of law school tier on bar passage . . . controlling for the influence of other variables.[114]

By positing a process in which variables "shape the eventual outcome," Sander is clearly advancing a causal story. Moreover, hypothesizing (and analyzing) how law school grades "suppress the impact of law school tier" is simply another way of saying "mediates" the relationship between tier location and bar passage. Philosophers and social scientists have long recognized that one cannot speak of mechanisms without invoking causal logic. "First, a mechanism is identified by the kind of effect or phenomenon it produces. . . . Second, a mechanism is an irreducibly causal notion. It refers to the entities of a causal process that produces the effect of interest."[115] Also recall that "earlier versions [of the academic deficient/mismatch hypothesis] have been advanced by opponents of race-based affirmative action almost from the time such programs began and it has been a reoccurring theme of arguments against affirmative action ever since."[116] Thus, the only potential scholarly contribution that Sander could offer to the ongoing debate was the empirical demonstration of a causal link between race-conscious affirmative action and diminished outcomes for its beneficiaries.[117] Even assuming,

---

that law school racial preferences cause blacks to be clustered at the bottom of the credential distribution at the great majority of law schools. . . . [T]hese grades are so low that they signify (based on later bar performance) that little learning is going on. Most whites with comparable credentials go to much less elite schools, get better grades, learn more, and thus do far better on the bar."); Richard H. Sander, *The Racial Paradox of the Corporate Law Review*, 84 N.C. L. REV. 1755, 1811 n.145 (2006) [hereinafter Sander, *Racial Paradox*] (arguing that the BPS data strongly suggest that "poor grades *lead* many blacks to drop out of [law school]" (emphasis added)); Sander, *Reply to Critics*, *supra* note 10, at 2010 (claiming that the low grades of black law school graduates, relative to white graduates, "will substantially disadvantage blacks, and will cause the black-white earnings gap to increase"); *see also supra* notes 53–54 and accompanying text (discussing Sander's presentation and attempted analysis of a formal causal model of the mismatch hypothesis).

114. Sander, *Mismeasuring*, *supra* note 24, at 2007, 2008 n.16 (emphasis added).

115. Peter Hedström & Petri Ylikoski, *Causal Mechanisms in the Social Sciences*, 36 ANN. REV. SOCIO. 49, 50 (2010) (citations omitted).

116. *See supra* note 66 and accompanying text (noting that the black student mismatch thesis is not new and has been repeatedly refuted).

117. *See, e.g.*, Ho, *supra* note 7, at 1998 ("Sander is relegated to investigating . . . the causal effect of attending a higher-tier law school.").

arguendo, that Sander never intended to posit a causal model, his more modest (and decidedly less scientific) claim of the identification of a plausible mechanism through which race-conscious affirmative action impacts legal education and employment outcomes is also refuted by the data when the most basic and routine rules of statistical inference are followed.

a. Mediation Analysis

There are five core problems with Sander's mediation analysis of students: (1) the focus on fully standardized regression coefficients which gives the misimpression that the total tier effect on the probability of bar passage is nearly three times as large as it actually is;[118] (2) the miscalculation of the uncertainty (standard errors) of the total tier effect—when the correct approaches for the standard errors are used, the total tier effect is statistically indistinguishable from zero;[119] (3) the sensitivity of the results to removal of the tier containing all historically black law schools (HBLS), whereby removing the HBLS tier excludes less than one-fifth of the black students in the sample but renders the total effect of tier location practically zero and statistically insignificant;[120] (4) the extremely poor ability of the model to accurately predict either LGPA or bar passage—only one-quarter of the variability in LGPA and one-fifth of the variability in first-time bar passage is explained by the model;[121] and (5) the unreasonableness of his assumption of unconfoundedness (meaning that there is no bias due to omitted variables).[122]

In the interest of space, I focus on the final problem: omitted variable bias. Thorough discussions of the first four problems are provided in Appendix A. One might ask why it is necessary to specifically test the accuracy of Sander's

118. *See infra* Appendix A.1.
119. *See infra* Appendix A.2.
120. *See infra* Appendix A.3.
121. *See infra* Appendix A.4.
122. There are additional significant shortcomings with Sander's mediation analysis. Specifically, he improperly: (1) employs a linear probability model to estimate his mediation analysis, although bar passage is a binary variable (interpolation bias); (2) models tier location as a continuous explanatory variable, although it is an ordinal variable (interpolation bias); (3) separates the second and third tiers although they are not meaningfully distinguishable in terms of the mismatch-relevant factors of UGPA, LSAT, and selectivity (instability bias); (4) focuses on a subsample who took the bar (nonrandom selection bias); and (5) compares respondents across tiers who do not have overlapping values across his core variables (extrapolation bias). Rather than address all of these issues here, I describe each issue in subsequent Parts of this Article and demonstrate that correcting for these problems undermines Sander's conclusions about mismatch effects.

unconfoundedness assumption in light of all of the other aforementioned errors. I offer the ensuing discussion because the identification of the causal effect (validity) is conceptually and analytically distinct from statistical inference based on reliability (the standard errors, confidence intervals, and *p*-values)[123] and predictive accuracy.[124] Even if Sander reported a highly statistically significant (precise) mismatch effect and a highly predictive model, the effect would still be biased if the causal estimate were impacted by confounding variables omitted from the model because the underlying statistical model misrepresents the actual process that generated the outcome.

Testing Sander's Unconfoundedness Assumptions. I reanalyzed Sander's mediation model to test the sensitivity of his results to the A_1–A_3 assumptions in several respects.[125] It is important to reiterate that Sander's use of LGPA to test the direct and indirect effects of tier location on bar passage is only valid if one has confidence that there are no unobserved variables that influence the tier-bar passage relationship (A_1) for direct effects, or the LGPA–bar passage (A_2 & A_4) and tier-LGPA relationships (A_3) for indirect effects.[126] But law students are neither randomly assigned to law schools nor randomly receive law school grades, so these assumptions are extremely difficult (if not impossible) to defend. One might plausibly assume conditional exogeneity/ignorability after adjusting for confounding variables, but one must take into account a wide range of potential confounders.[127] Sander only

123. A pioneer of modern statistics, Ronald A. Fisher, famously remarked that the concerns of statistics are "a) [t]o discover what quantities are required for the adequate description of a population . . . [and] b) [t]o determine how much information, and of what kind, respecting these population values is afforded by a random sample." Ronald A. Fisher, *On the "Probable Error" of a Coefficient of Correlation Deduced From a Small Sample*, 1 METRON 3 (1921).

124. Paul Allison explains:

> There are two major uses of multiple regression: prediction and causal analysis. In a prediction study, the goal is to develop a formula for making predictions about the dependent variable, based on the observed values of the independent variables. . . . In a causal analysis, the independent variables are regarded as causes of the dependent variable. The aim of the study is to determine whether a particular independent variable really affects the dependent variable, and to estimate the magnitude of that effect, if any.

PAUL D. ALLISON, MULTIPLE REGRESSION: A PRIMER 1–2 (1999) (emphasis omitted).

125. Assumption A_4 could not be examined with the available data. *See infra* note 146 and accompanying text.

126. *See supra* Part I.

127. Matthew Blackwell, *A Selection Bias Approach to Sensitivity Analysis for Causal Effects*, 22 POL. ANALYSIS 169, 171 (2014) ("In an observational study, however, the analyst's goal is to collect as many variables as possible to . . . make [the unconfoundedness assumption] as plausible as possible.").

controls for a handful of variables in his mediation model, so one cannot have much confidence that A_1–A_4 are satisfied in his analyses.

In the absence of randomization or appropriate controls for confounding variables, a longstanding tradition in economics, dating back to the 1920s, is the use of instrumental variables, which are devices one employs to identify the causal effect by eliminating potential bias.[128] A detailed discussion of the instrumental variable framework is beyond the scope of this Article, but the core idea is that Sander's mediation analysis should have included variables that affect his key explanatory variables (tier location and LGPA) but do not affect the dependent variables (LGPA and bar passage) except through their effect on those key explanatory variables (treatment variables).[129] The goal of instrumental variable analysis is not to control for confounding variables, but to indirectly estimate the relationship between the treatment and the outcome variable. Stated differently, instrumental variables induce changes in the explanatory variable, but have no independent effect on the outcome variables. The analyst can identify a causal effect because the instrumental variable is uncorrelated with any unobserved factors that simultaneously influence the explanatory variable and the outcome variables. The causal effect of the treatment is estimated by the indirect effect of the instrumental variable on the outcome that operates through the treatment. Unfortunately, the validity of an instrumental variable cannot be confirmed with data, so it must be defended based on theory and the plausibility of the unconfoundedness assumption.[130]

Most empirical scholarship in economics offers an extended discussion of the use of (or search for) valid instrumental variables, especially when key

128. *See, e.g.*, Guido W. Imbens, *Instrumental Variables: An Econometrician's Perspective*, 29 STAT. SCI. 323 (2014) (describing early applications of instrumental variables, sometimes called exclusion restrictions, in economics).

129. The instrumental variable can be correlated with both the treatment and outcome variables, but it can only have a direct causal effect on the treatment variable. LGPA is a mediating variable, so it is a dependent variable in the first equation and an endogenous treatment variable in the second question predicting bar passage.

130. Technically, instrumental variables must satisfy four conditions: (1) uncorrelated with omitted variables impacting the outcome variable (so the instruments and the outcome do not share causes); (2) correlated with the endogenous (causal) variable of interest; (3) the instrument does not affect the outcome except through its potential effect on endogenous (causal) variable; and (4) strongly correlated, rather than weakly correlated, with the causal variable of interest. The first and third conditions pertain to the validity of the instrument, whereas the second and fourth conditions relate to the relevance of the instrument. *See* A. COLIN CAMERON & PRAVIN K. TRIVEDI, MICROECONOMETRICS: METHODS AND APPLICATIONS 99–100 (2005) (explaining the requirements of instrumental variable estimation for regression analysis).

explanatory variables are endogenous such as LGPA, but Sander's work is silent on this issue. As one econometrician explained, "The study of identification [nonconfounding] logically comes first. Negative identification findings imply that statistical inference [significance tests and *p*-values] is fruitless . . . ."[131] Jesse Rothstein and Albert Yoon, as well as Doug Williams, implement an instrumental variable framework when examining mismatch effects in the BPS data, but arrive at contradictory conclusions.[132] Whereas Jesse Rothstein and Albert Yoon fail to uncover support for the mismatch hypothesis, Williams reports extremely strong evidence of mismatch effects.[133] Williams's results are highly suspect, however, because of his choice of an invalid instrument, his extremely limited set of control variables, and his removal of 61 percent of the law schools (and 63 percent of the law students) from the analysis.[134] In reviewing Williams's analysis, Peter Arcidiacono and Michael Lovenheim note that "[Williams] instrumented for law-school tier with the first choice variable, resulting in massive mismatch effects. However, the estimated effects are so large as to not be plausible."[135]

131. CHARLES F. MANSKI, IDENTIFICATION PROBLEMS IN THE SOCIAL SCIENCES 6 (1995); *see also* Carlos Brito & Judea Pearl, *A New Identification Condition for Recursive Models With Correlated Errors*, 9 STRUCTURAL EQUATION MODELING 459, 459 (2002) ("Before [causal models] can be estimated and evaluated against data, a researcher must make sure that the parameters of the model are identified . . . ." (emphasis omitted)).

132. *See* Rothstein & Yoon, *supra* note 21, at 22 ("[W]e conclude that the use of affirmative action at these schools does not generate meaningful mismatch effects.").

133. Williams, *supra* note 22, at 194 ("This article demonstrates . . . evidence for mismatch effects in legal education.").

134. *See* Arcidiacono & Lovenheim, *supra* note 7, at 20 n.30 (doubting the accuracy of Williams's instrumental variable results); Kidder & Lempert, *Mismatch Myth*, *supra* note 22, at 109 (describing bias in Williams's analysis resulting from his choice of an invalid instrument and his removal of the third and fourth law school tiers).

135. Arcidiacono & Lovenheim, *supra* note 7, at 20 n.30. The instrumental variable that Williams uses is the first choice variable that was employed by Ayres and Brooks and, in several analyses, by Sander. The first choice variable purportedly indicates whether a student elected to enroll in their first, second, or third choice law school, provided the student was admitted to multiple schools and, therefore, had multiple options available to them. But all of these analyses using the choice variable instrument rest on the assumption that second choice students are equally strong, academically, as their first choice counterparts; this is unlikely to be true, however, because many second choice students are academically stronger than their first choice counterparts based on factors observable to admissions officers, but not captured by the BPS (or not included in Ayres and Brooks, Sander, or Williams's models). Williams acknowledges that the first choice variable will only be a valid instrument if "the decision of whether or not to attend one's first-choice school is uncorrelated with unobservable ability but correlated with selectivity," Williams, *supra* note 22, at 185, which would imply that "the exogenous factors that determine the value of CHOICE . . . are uncorrelated with the disturbance [for example, omitted variables] in the outcome [e]quation." *Id.* at 191. *See also infra* Subpart II.F.1.b (discussing problems with the use of the first/second choice framework).

Williams appears to recognize the questionable credibility of his estimates when stating, "[t]he magnitudes of the coefficients are large—in fact much larger than what is required to explain racial differences [for blacks]," Williams, *supra* note 22, at 191, and then switches his focus to the results for the "minority subgroup" composed of black, Native American, and Latinx students, *id.* at 172, because the effects are significantly smaller but still very large. Williams cautions, "[a]lthough the large magnitudes could reflect model misspecification, they could also simply reflect the small sample size. . . . Although these [instrumental variable] results should be interpreted carefully because of the magnitudes, they are consistent with mismatch effects." *Id.* at 192–93.

Williams reports results from a test (the *F*-statistic) purportedly showing that the first choice variable is a sufficiently strong instrument in most of his models. *Id.*; *see* CAMERON & TRIVEDI, *supra* note 130, at 105 (describing the utility of the *F*-statistic in instrumental variable estimation). A strong instrument is one that is highly correlated with the endogenous variable. *Id.* at 103–04. But Williams is much less transparent about what the *F*-statistic does not reveal: whether his instrument is a *valid* instrument. Williams assumes, without providing adequate justification, that his instrument is correct, and then examines whether the instrument is *weak*. But an instrument may be both strong and invalid, so a test focusing solely on the strength of an instrument says nothing about whether it is a proper instrument. Scholars have noted that potential instrumental variables must be subject to "intuitive, empirical, and theoretical scrutiny to reduce the risk of using invalid instruments." Michael P. Murray, *Avoiding Invalid Instruments and Coping With Weak Instruments*, J. ECON. PERSP., Fall 2006, at 111, 130. "[I]nstrumental variable estimation requires both a valid instrument on which to stand and an instrument that isn't too short (or 'too weak')". *Id.* at 111; *see also* CAMERON & TRIVEDI, *supra* note 130, at 99–100. Relatively recent developments in econometrics permit, in some situations, a partial estimation of causal effects in the presence of potentially invalid instruments, but Williams does not employ any of these approaches to examine the robustness of his results. *See, e.g.*, Timothy G. Conley, Christian B. Hansen & Peter E. Rossi, *Plausibly Exogenous*, 94 REV. ECON. & STAT. 260 (2012) (discussing methodologies, introduced in 2008, for causal inference with instruments that are potentially invalid); Aviv Nevo & Adam M. Rosen, *Identification With Imperfect Instruments*, 94 REV. ECON. & STAT. 659 (2012) (same).

Even a cursory examination of Williams's results should raise strong suspicion about his models and the conclusions drawn from them. According to Williams's instrumental variable models that focus on the minority subgroup, and therefore allegedly reveal smaller mismatch effects than the models focusing exclusively on black students, the impact of mismatch on bar passage rates is still astronomical:

> Focusing on the results for the minority subgroup . . . the results suggest that eliminating two-thirds of the actual mean academic index distance would increase first-time bar passage by an amount seven times larger than the unexplained gap (68 percent compared to 10 percent), would increase ever passing the bar by an amount nine times larger than the unexplained gap (44 percent compared to 5 percent), and would increase lawyer completion rates by an amount six times larger than the unexplained gap (41 percent compared to 7 percent).

Williams, *supra* note 22, at 191–92. When one recognizes that the estimates for first-time bar passage for the black students-only model is 66 percent larger than the estimates for the minority subgroup (56 percent larger for eventual bar passage and 52 percent larger for lawyer completion), and 55 percent of the minority subgroup consists of black students, one can readily understand why Arcidiacono and Lovenheim stated that Williams's "estimated effects are so large as to not be plausible." An obvious explanation for Williams's farfetched findings is that the parameter estimates from his model, which are

Many social scientists—especially noneconomists—are skeptical of the instrumental variable framework because it is practically impossible to definitively determine whether a variable is, in fact, a valid instrument.[136] And even in the event that a valid instrument is available, it may be a "weak" instrument (it is only mildly associated with outcome variable through the treatment), and therefore incapable of providing much information about the true causal effect of the treatment.[137] An alternative approach—popular in econometrics, psychometrics, sociological methodology, and educational statistics—permits the testing of the A_1–A_4 assumptions by allowing for a correlation between disturbances that impact tier location, LGPA, and bar passage.[138] The approach assumes that shared unobserved student-specific factors influencing tier location, LGPA, and bar passage manifest themselves in the correlation of the disturbance terms of the equations. Not only is this test possible for the mediation analysis framework adopted by Sander, but it is routinely employed in mediation analysis when the assumption of treatment-outcome and mediator-outcome unconfoundedness seems unrealistic.[139] "[T]he correlation

based primarily on white law students, are inapplicable to the black and minority student subsamples and lead to nonsensical interpretations of causal effects. *See infra* Subparts II.D–II.E.

136. *Compare* MORGAN & WINSHIP, *supra* note 58, at 196–97 (explaining that the basic underlying assumption of the instrumental variable estimator is both very strong and untestable), *with* James J. Heckman & Richard Robb, Jr., *Alternative Methods for Evaluating the Impact of Interventions*, *in* LONGITUDINAL ANALYSIS OF LABOR MARKET DATA 156, 185 (James J. Heckman & Burton Singer eds., 1985) (arguing the advantage of the instrumental variable estimator is that it "is the least demanding in the a priori conditions that must be satisfied for its use"). Sociological methodologists Thomas DiPrete and Markus Gangl explain:

> [U]nder a set of [standard] assumptions, the [instrumental variable] estimator estimates the . . . average effect of the treatment for the subsample of the population that is induced by a specific change in the value of the [instrumental variable] to select themselves into treatment. [But] [t]he [instrumental variable] solution to the problem of endogeneity comes at the cost of introducing new sources of uncertainty about the true causal effect.

Thomas A. DiPrete & Markus Gangl, *Assessing Bias in the Estimation of Causal Effects: Rosenbaum Bounds on Matching Estimators and Instrumental Variables Estimation With Imperfect Instruments*, 34 SOCIO. METHODOLOGY 271, 276–77 (2004) (footnote omitted).

137. MORGAN & WINSHIP, *supra* note 58, at 23.

138. *See* Robert C. Luskin, *Estimating and Interpreting Correlations Between Disturbances and Residual Path Coefficients in Nonrecursive (and Recursive) Causal Models*, 22 AM. J. POL. SCI. 444, 450 (1978).

139. *See, e.g.*, Imai et al., *supra* note 83, at 61 ("[R]esearchers can easily check the robustness of their conclusion obtained under the sequential ignorability assumption via correlation between [the disturbances]."); Luskin, *supra* note 138, at 450; Ho, *Reply to Sander*, *supra* note 23, at 2015 n.24 ("To be clear, some regression techniques account for some types of unobserved heterogeneity, but not the logistic regression employed by Sander.").

between the disturbances of the . . . structural equations expresses the extent to which those equations fail to recognize major causes of their dependent variables that are either the same or correlated."[140] In particular:

> [A] nonzero correlation parameter can be interpreted as the existence of omitted variables that are related to both the observed value of the mediator . . . and the potential outcome[] . . . even after conditioning on the treatment variable . . . (and the other observed [pretreatment] covariates).[141]

As with all parametric statistical models, the correlated disturbances approach often rests on its own set of assumptions—most importantly, the joint distribution of the disturbances is multivariate normal. This, however, is a common assumption of most statistical models that examine multiple processes (equations) and prior simulation studies suggest that "most aspects of statistical inference are highly robust to this assumption" of normality of disturbances, "including estimation of covariate effects."[142] Moreover, when the multivariate normal assumption is violated, other methods are available that do not assume multivariate normality.[143] These methods are most appropriate when the number of variables in the model is small and the sample size is large.[144]

It is important to note, at the outset, that it is impractical to simultaneously test all of the assumptions of the (non)correlation of disturbances relevant to the mismatch hypothesis because such a test requires imposing constraints on model parameters that may be questionable on theoretical grounds.[145] It is also nearly impossible to test assumption (A_4) given

140. Luskin, *supra* note 138, at 450; *see also* Ronald S. Landis, Bryan D. Edwards & Jose Cortina, *On the Practice of Allowing Correlated Residuals Among Indicators in Structural Equation Models*, in STATISTICAL AND METHODOLOGICAL MYTHS AND URBAN LEGENDS: DOCTRINE, VERITY AND FABLE IN THE ORGANIZATIONAL AND SOCIAL SCIENCES, 193, 204 (Charles E. Lance & Robert J. Vandenberg eds., 2009) ("To the degree that two residuals correlate, there is evidence that there exists a cause of both of the variables to which the residuals are attached but that is not specified in the model.").

141. Imai et al., *supra* note 83, at 61.

142. Charles E. McCulloch & John M. Neuhaus, *Misspecifying the Shape of a Random Effects Distribution: Why Getting it Wrong May Not Matter*, 26 STAT. SCI. 388, 400 (2011).

143. Michael W. Browne, *Asymptotically Distribution-Free Methods for the Analysis of Covariance Structures*, 37 BRIT. J. MATHEMATICAL & STAT. PSYCH. 62 (1984) (introducing an asymptotically distribution-free method to analyze structural equation models when the data violate the assumption of multivariate normality).

144. Jöreskog, *supra* note 109 (explaining that asymptotically distribution-free methods require a large number of observations for accurate estimation).

145. *See* REX B. KLINE, PRINCIPLES AND PRACTICE OF STRUCTURAL EQUATION MODELING 137 (3d ed. 2011) (noting that "the imposition of . . . constraints generally requires [*a priori*] knowledge about relative effect magnitudes"). These constraints typically entail

the available data.[146] Consequently, these limitations render my results suggestive rather than conclusive. Nonetheless, the primary concern for Sander's mediation framework is the validity of A_2 and A_4, so if *either* assumption is violated, then any alleged mediated causal effect cannot be demonstrated, even if A_1 and A_3 are satisfied.[147] In an effort to make Sander's assumptions more plausible, given the available data, I include a richer set of pretreatment covariates to potentially satisfy the conditionally exogenous assumptions.[148] Specifically, the models I examine include sex/gender, marital

restricting certain correlations between observed or unobserved variables to "zero" (exclusion constraints) or requiring two correlations across pairs of variables to be equivalent (equivalence constraints). These constraints are mathematically necessary to obtain sensible results about hypothesized relationships. *See* PAMELA M. PAXTON, JOHN R. HIPP & SANDRA MARQUART-PYATT, NONRECURSIVE MODELS: ENDOGENEITY, RECIPROCAL RELATIONSHIPS, AND FEEDBACK LOOPS 26 (2011). There are important conceptual differences between the instrumental variable approach in econometrics and the correlated unobserved variables (or disturbances) analysis, although the exclusion restriction(s) required to estimate the correlated disturbances analysis share similarities to the instrumental variable approach. James B. Kirby & Kenneth A. Bollen, *Using Instrumental Variable Tests to Evaluate Model Specification in Latent Variable Structural Equation Models*, 39 SOCIO. METHODOLOGY 327, 333 (2009); *accord* Brito & Pearl, *supra* note 127, at 472 (noting that "criteria based on instrumental variables ensure the identifiability of one parameter at a time, while the criterion established in this paper [using correlated errors] ensures the identifiability of the model as a whole").

146. *See, e.g.*, Imai et al., *supra* note 83, at 61 (explaining that the correlated errors sensitivity test "does not address the possible existence of post-treatment confounders"). Recall that Sander claims that affirmative action has three "first-order effects"—learning mismatch, competition mismatch, and social mismatch—which provide the direct theoretical link between disparities in academic credentials and various outcomes. *See supra* note 8. In other words, these first order effects are the mechanisms through which mismatch influences "second order effects" (for example, lower graduation rates, failing the bar, lower wages, and more). *See supra* note 8. A plausible test of A_4, then, might be a model that takes into account all three of the first order effects. But the analysis is further complicated by the fact that Sander's framework also implies that the hypothesized causal mechanisms are not necessarily causally independent (for example, affirmative action may directly lead to learning mismatch and competition mismatch, but may also lead to competition mismatch through learning mismatch). *See* Kosuke Imai & Teppei Yamamoto, *Identification and Sensitivity Analysis for Multiple Causal Mechanisms: Revisiting Evidence From Framing Experiments*, 21 POL. ANALYSIS 141, 142–43 (2013) (discussing mediation analysis when multiple mediators have a causal relationship with one another). It is also worth noting, however, that Sander's distinction between first and second order effects is fuzzy. For example, are lower graduation rates a first order effect (competition mismatch) or a second order effect? According to Sander, competition mismatch occurs when students change their original or education career plans (such as by dropping out of law school) because they are outmatched. *See supra* note 8.

147. *See supra* Part I. The validity of A_2 & A_4 are necessary, but insufficient conditions for the identification of causal effects in the mediation analysis. The spuriousness of the associations between tier–bar passage (A_1) and tier-LGPA (A_3) will also undermine the mediation analysis.

148. *See supra* Part I.

status, number of children, mother's education, father's education, family income, financial debt load, UGPA, LSAT, number of law school applications, number of law school acceptances, college major, tier location, and LGPA.[149] Consistent with Sander's analysis, I examine the subsample of black students. In contrast to Sander's analysis, which improperly employs a linear probability model, my mediation analysis models first-time bar passage via a binary regression model and tier location via an ordinal regression model.[150]

My reanalysis of Sander's mediation model reveals that several of the unconfoundedness assumptions (A_1, A_2, and A_3) are violated across various model specifications. For example, when simultaneously examining the potential endogeneity of both tier location and LGPA when examining first-time bar passage, the correlations between unobservables influencing tier–bar passage (A_1), LGPA–bar passage (A_2), and tier-LGPA (A_3) and are all statistically significant; ranging from -0.53 to 0.34 across various model specifications (see Figure 4). The presence of these correlations of disturbances biases inferences about both the direct and indirect (mediated) effects estimated in Sander's model. Specifically, the violation of A_1 undermines inferences about the direct effect of tier location on bar passage (by contaminating the TIER→BAR relationship),[151] and the violations of A_2 and A_3 biases inferences about the indirect effect of tier

149. First-time bar passage is a dependent variable in all of the estimated models, *see supra* Figure 4 (depicting bar passage as a consequence of the direct and/or indirect effects of tier location, UGPA, and LSAT). Tier location and LGPA are explanatory variables in some models and dependent variables in other models, *see supra* Figure 4 (showing tier location is caused, in part, by UGPA and LSAT, and LGPA is caused, in part, by UGPA, LSAT, and tier location). In order to identify the parameter estimates and correlations, the effects of some of these variables were constrained to zero (known as an exclusion restriction) across various model specifications. *See supra* Figure 4.

150. Kosuke Imai, Luke Keele & Dustin Tingley, *A General Approach to Causal Mediation Analysis*, 15 PSYCH. METHODS 309, 316 (2010) (describing a framework for causal mediation analysis with noncontinuous mediator and outcome variables); *see also infra* Appendix C (underscoring the importance of employing the ordinal regression model when estimating the impact of entering credentials on tier location).

The results of the sensitivity analysis for the correlation of the disturbances were similar when employing Sander's linear probability model. The assumption of multivariate normality of the disturbances, which is required for the proper estimation of the correlations of the disturbances in the mediation analysis, is violated by linear probability model (because the error distribution is, in fact, binomial). This assumption, however, is not violated by the binary and ordinal (probit) models because the disturbances of underlying latent variables are assumed to satisfy multivariate normality. *See* Christopher Winship & Robert D. Mare, *Structural Equations and Path Analysis for Discrete Data*, 89 AM. J. SOCIO. 54, 76 (1983).

151. The violation of A_1 would also bias an inference about the total effect of tier location on bar passage in an unmediated model (that is, a model that did not control for posttreatment variables) because of the confounding resulting from correlation between U_1 and U_2. *See supra* Figure 4.

location on bar passage (by contaminating the TIER→LGPA and LGPA→BAR relationships, respectively). Consequently, it is necessary to examine the upper and lower limits of the estimated causal effect by varying the magnitude of the correlation between the unobservables, as well as "the proportion of previously unexplained variance (either in the mediator or in the outcome) that is explained by the unobserved confounder."[152] This causal bound testing sensitivity analysis is also appropriate (and advisable) even when analyzing mismatch effects outside of the mediation framework.[153]

In summary, unmeasured common causes appear to account for the results reported from Sander's mediation analysis. Once those confounding factors are taken into account, Sander's model does not provide evidence of mismatch effects. Causal inference scholars have cautioned:

> Investigators should think more carefully about and collect data on and control for such mediator-outcome confounding variables when mediation analysis is of interest. *If the investigator is aware that unmeasured confounding may be an issue in his or her study, sensitivity analyses should be implemented.*[154]

There are numerous statistical tests that allow the analyst to examine the sensitivity of Sander's model to the assumptions concerning the impact on these unobserved factors on tier location, LGPA, and bar passage.[155] Furthermore, scholars have also suggested ways analysts may avoid posttreatment bias when investigating direct and indirect causal pathways.[156] Yet Sander does not report (nor does it appear that he has examined) the robustness of his findings utilizing these available, and statistically defensible,

152. Imai et al., *supra* note 83, at 62 (discussing various sensitivity tests using the correlation of the disturbances).
153. *See infra* note 461 and accompanying text.
154. VANDERWEELE, *supra* note 64, at 26 (emphasis added) (citations omitted).
155. MORGAN & WINSHIP, *supra* note 58, at 169 (discussing the examination of causal relationships in the presence of unobservables); VANDERWEELE, *supra* note 64, at 66 (discussing the importance of sensitivity analyses to assess both the potential spuriousness of treatment-outcome relationships and the plausible range of values for the causal effect according to assumptions about the influence of unobservables); Blackwell, *supra* note 123, at 169 (presenting "a broad methodology for evaluating the sensitivity of causal effect estimation to specific critiques of bias"); *see also* Luskin, *supra* note 138, at 447–50 (emphasizing the importance of examining the potential correlation between unobserved factors in causal models that test for mediation effects).
156. Acharya et al., *supra* note 56 (describing the controlled direct effect (CDE) approach to causal inference); Imai & Yamamoto, *supra* note 142, at 141 (presenting methods for "multiple causal mechanism" analysis to examine alternative causal pathways); Gary King & Langche Zeng, *The Dangers of Extreme Counterfactuals*, 14 POL. ANALYSIS 131, 147 (2006) (proposing the "multiple-variable causal effects" framework).

analytical frameworks. As noted earlier, concomitant variable adjustment bias is very likely to occur in nonexperimental settings investigating mediation effects because these observational studies most heavily rely on the assumption of unconfoundness to identify causal effects. Moreover, as I explain *infra*, the likely reason Sander's models violate A_1–A_3 is because he neglects to control for a wide range of relevant variables that are available in the BPS data, although his critics have demonstrated that his results are sensitive to the inclusion of a proper set of pretreatment confounding factors.[157] Some of Sander's later work on mismatch avoids bias resulting from adjusting for a concomitant variable adjustment, but his early work is susceptible to this form of bias and his responses to this particular criticism fail to adequately acknowledge the seriousness of the problem and how it undermines inferences drawn from those models.

2. Nonrandom Sample Selection Bias

The second form of posttreatment bias, which Sander identifies but improperly analyzes (and likely exacerbates), results from dropping or subsetting observations based on posttreatment criteria.[158] Subsetting observations based on posttreatment criteria biases causal inference in a manner similar to the aforementioned discussion of adjusting for a posttreatment variable. Whereas the concomitant variable adjustment bias focused on LGPA, the nonrandom sample selection bias concern arises from Sander's analysis of bar passage rates for those individuals who graduated from law school and decided to take the bar without proper adjustment for this nonrandom subset law students. Malgorzata Wojtyś and colleagues explain:

157. *See infra* Subpart II.C (discussing research contradicting mismatch theory using the BPS data when a richer set of control variables are included in the analysis).

158. Sander, *Replication of Mismatch*, *supra* note 4, at 80 ("[Ayres and Brooks] only included people who actually sat for the bar exam. But . . . there is a good argument for including in the analysis some or all of those who did not take the bar exam."). Sander notes that this type of posttreatment bias was highlighted in the Empirical Scholars Brief, but was also ignored by Ayres and Brooks and Ho, *id.* at 88, however, Sander's statement is misleading with respect to Ho's work. While it is true that Ho does not specifically analyze the extent of the possible bias resulting from only analyzing law school graduates, Ho openly acknowledges that he did not examine all of the various problems with Sander's analysis, including nonrandom sample selection. *See infra* note 178 and accompanying text. There are various causes of nonrandom sample selection, including nonrandom selection resulting from the treatment. The proper methodological approaches to address nonrandom selection vary depending on the specific mechanisms of selection, but the various forms of nonrandom selection impacts causal inference in a similar fashion.

> Non-random sample selection arises when an output variable of interest is available only for a restricted non-random sub-sample of the data. This often occurs in sociological, medical and economic studies where individuals systematically select themselves into (or out of) the sample . . . . If the aim is to model an outcome of interest in the entire population and the link between its availability in the sub-sample and its observed values is through factors which cannot be accounted for then any analysis based on the available sub-sample will most likely lead to biased conclusions.[159]

This phenomenon is sometimes called "'censoring' or 'truncation' of data due to death" because it was heavily emphasized by biostatisticians.[160] Sander's analysis attempts to examine the impact of mismatch on bar passage rates by examining only those students who graduated from law school, but the analytical framework he employs is inappropriate precisely because one of the hypothesized effects of mismatch is dropping out/dismissal from law school. By focusing only the subsample of law school graduates, Sander's analysis creates the same problems discussed earlier when improperly adjusting for LGPA which is also a consequence of mismatch.[161] Figure 5 depicts the hypothesized causal processes that result in nonrandom sample selection. Note that there is no causal arrow linking law school graduation status (labeled "in/out" in Figure 5) to bar passage—that is, in/out is not on the causal pathway between UGPA, tier, and LGPA and bar passage.[162] This diagram simply illustrates the processes through which these explanatory variables impact whether a student ultimately sits for the bar exam.[163] The likelihood an individual drops out of law school may not only result from mismatch (for instance, poor grades and inadequate learning) but for other reasons as well,

159. Malgorzata Wojtyś, Giampiero Marra & Rosalba Radice, *Copula Based Generalized Additive Models for Location, Scale and Shape With Non-Random Sample Selection*, 127 COMPUTATIONAL STAT. & DATA ANALYSIS 1, 1 (2018) (citations omitted).

160. Rubin, *supra* note 91, at 300.

161. But it is worth emphasizing that there are data in the BPS that capture the reasons why individuals left law school. Financial considerations were among the most common reasons why students dropped out in their first year of law school and grades were infrequently mentioned. Linda F. Wightman, *The Threat to Diversity in Legal Education: An Empirical Analysis of the Consequences of Abandoning Race as a Factor in Law School Admission Decisions*, 72 N.Y.U. L. REV. 1, 37 (1997).

162. See *supra* note 74 and accompanying text for a discussion of adjusting for a posttreatment variable on or off of the causal pathway.

163. As with Figure 4, some features of the model (including latent constructs and LGPA) have been "greyed out" to reduce the amount of clutter in the diagram.

such as finances, poor health, or changing career goals.[164] If the statistical model does not account for these other reasons which may be correlated with both attending a school in a higher tier and the likelihood of bar passage, then the model cannot properly measure the causal effect of mismatch.[165]

Similar to the first type of posttreatment bias I discussed, sample selection bias also presents a problem that is akin to omitted variable bias. The appropriate analytical task entails comparing the bar passage rates of "mismatched" and "nonmismatched" individuals while taking into account the effect of unobserved factors that simultaneously influence graduation probability and bar passage. The Empirical Scholars Brief (ESB) provides an intuitive numerical example:

> Suppose, for example, that we conducted the ideal experiment, randomizing 200 students to attend selective and less-selective institutions. Assume that the results are that 95 out of 100 graduate at the selective institution, and that 70 pass the bar exam. Meanwhile, at the less selective institution, 80 out of 100 graduate, and 60 pass the bar. That experiment suggests that students benefit from attending a selective institution, both in terms of graduation and bar passage. If we focus only on those who graduated, however, the bar passage rate is 0.74 at the selective institution (70/95), and 0.75 at the less selective institution (60/80). Referring to these findings would lead a reader to wrongly infer that a more selective law school harms students. *The selective law school, possibly via improved teaching and better resources, manages to graduate more students. But the subsets of students graduating from either school are not fully comparable*.[166]

Various statistical adjustments designed to account for sample selection bias have been available since the 1970s.[167] These models are useful because:

> Sample selection models allow one to use the entire sample whether or not observations on the output variable were generated. In its classical form, it consists of two equations which model the probability of inclusion in the sample and the outcome variable

164. A similar argument can be made for individuals who graduate from law school but decide not to sit for the bar exam.

165. *See* Ho, Evaluating Affirmative Action, *supra* note 23, at 10 n.7 (noting that statistical inferences based on a nonrandom subset of cases with bar passage data in the BPS are likely to be biased).

166. Empirical Scholars Brief in *Fisher I*, *supra* note 30, at 23 (emphasis omitted and added).

167. *See* James J. Heckman, *Sample Selection Bias as a Specification Error*, 47 ECONOMETRICA 153 (1979); Rubin, *supra* note 91, at 301 (referencing the "Rubin Causal Model" based on a series of papers from the 1970s).

> through a set of available covariates, and of a joint bivariate distribution linking the two equations.[168]

The joint bivariate distribution allows the analysts to take into account unobserved factors that impact both decisions (in/out and outcome).[169] The specific approach to addressing nonrandom sample selection will depend on, among other things, whether nonrandom selection criteria is a consequence of the treatment variable (on the causal pathway).[170]

Rather than utilizing these techniques or fully acknowledging the potential biases that are likely to result from ignoring them, Sander incorrectly elects to add those dropout/nontakers cases to the analysis and count them as bar passage failures.[171] Not only is his approach statistically indefensible, but it may bias his results in favor of his mismatch hypothesis because there will be students who are apparently "mismatched," based on UGPA and LSAT, who leave law school in good academic standing (or graduate from law school but elect to not take the bar) and would have likely passed the bar exam, but Sander treats those individuals as bar failures.[172] For example, in the BPS, 12.6 percent of nongraduates (185 students) who reported first year grade point average (FYGPA) were at or above the median FYGPA for students who graduated and passed the bar on the first attempt. With respect to academic index scores (a combination of UGPA and LSAT), 30.1 percent of nongraduates (716 students) were at or above the median academic index score for students who graduated and passed the bar on the first attempt. With respect to black students, 8.7 percent of nongraduates (23 students) were at or above the median FYGPA, and 18.8 percent (62 students) were at or above the median academic index score for first-time bar passers. These data, taken together with the fact that FYGPA and academic index score are poor predictors of first-time bar passage (explaining, respectively, 14 percent and 12 percent of the variance in the BPS study), strongly suggest that it is

168. Wojtyś et al., *supra* note 159, at 1.
169. Early formulations of the sample selection model required a bivariate normal (Gaussian) distribution, but that restriction has been relaxed and a much wider range of probability distributions can be accommodated. *Id.* at 2.
170. Rubin, *supra* note 91, at 305.
171. Sander, *Replication of Mismatch*, *supra* note 4, at 80 (assessing mismatch by examining those who graduate, but do not take the bar and classifying "[f]ailers" as "both people who failed the bar on their first attempt and people who did not graduate from law school"); Sander, Whitepaper, *supra* note 4, at 20 tbl.10.
172. Jesse Rothstein and Albert Yoon make a similar analytical error, but they avoid other mistakes made by Sander and they find no evidence of mismatch (and some evidence of reverse mismatch). *See* Rothstein & Yoon, *supra* note 21, at 19–21, 35–37.

incorrect to assume that nongraduates would have failed the bar exam on the first attempt.[173]

Figure 6 combines the diagrams in Figures 4 and 5, illustrating the interrelationships between the core variables influencing LGPA, graduation/sitting for the bar exam, and bar passage. Similar to my analysis

173. Sander posits, "[I]f mismatch operates at the graduation level, then not including non-graduates could bias the analysis against mismatch," Sander, *Replication of Mismatch*, *supra* note 4, at 80 n.13, but he does not provide data to support this claim. As noted above, 30 percent of nongraduates were at or above the median academic index score of those students who passed the bar on the first attempt, and nearly 13 percent of nongraduates reported a first year grade point average (FYGPA) at or above the median FYGPA for first-attempt bar passers. It is also noteworthy that the correlation between FYGPA and first-attempt bar passage considerably varies by race and race-tier location pairings. I examine the biserial correlation between FYGPA and first-attempt bar passage, which is the most appropriate measure of association in this context because it captures the relationship between a dichotomous variable (here, passing/failing the bar exam) and another continuous variable (here, FYGPA). PETER Y. CHEN & PAULA M. POPOVICH, CORRELATION: PARAMETRIC AND NONPARAMETRIC MEASURES 35 (2002) (explaining that the correlation between a dichotomous variable and a continuous variable will be biased downward when one uses a test that assumes both variables are continuous).

My analysis reveals that this correlation is more than twice as large for white students than for black students (0.49 versus 0.22). When the race effect is disaggregated by tier location, the correlation for white students is rather consistent across tiers, ranging from 0.47 (Tier 3) to 0.54 (Tier 4); for black students, however, the correlation varies substantially across tiers—ranging from 0.19 (Tier 1) to 0.57 (Tier 6). The correlations for black students for Tiers 1–4 are very similar (ranging from 0.19 to 0.25), but the correlations for Tiers 5 and 6 much larger: 0.39 and 0.57, respectively. This suggests that FYGPA is only as predictive of bar passage for black students as it is for white students when black students attend school in the HBLS tier. More than 80 percent of black students with a reported FYGPA who did not graduate attended schools located in Tiers 1–5, which are the tiers where the relationship between FYGPA and first-time bar passage is significantly weaker for black students than for white students. Even if we include black students from Tier 5, nearly 70 percent of black students who did not graduate and a reported FYGPA were from schools in the top four tiers.

My examination of the FYGPA–graduation association yielded similar results as the FYGPA–bar passage analysis. The biserial correlation for white students was 182 percent greater than the biserial correlation for black students (0.49 versus 0.27). When disaggregated by tier, the correlations range from 0.29 (Tier 1) to 0.55 (Tier 5) for white students. For black students, the correlations range from 0.08 (Tier 1) to 0.74 (Tier 6). Not only was the correlation between FYGPA and graduation very weak for black students in Tier 1, it was also the only correlation that was not statistically significant. In contrast, FYGPA is very predictive of the likelihood of graduation for black students in Tier 6—the HBLS tier—but these are also the schools that are least likely to adopt race-conscious admission practices for black students. *See* Sander, *Systemic Analysis*, *supra* note 7, at 416 & tbl.3.2 (noting that the median academic index score differential between black and white students is the smallest in the HBLS tier). These results strongly suggest that FYGPA is not very predictive of graduation from law school and first-time bar passage for black students attending a law school outside of the HBLS tier, yet 82 percent of black students attended law schools outside of the HBLS tier.

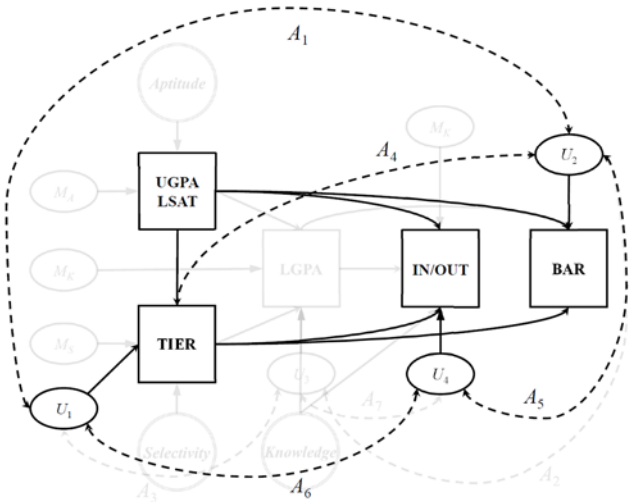
of concomitant variable adjustment bias,[174] it is possible to assess the relationships between unobservables impacting students' law school graduation rates ("in/out" decisions), law school tier location, and first-time bar passage rates. The results of this analysis reveal that the correlations between the disturbances impacting in/out-bar passage (A_5), tier-in/out (A_6), and LGPA-in/out (A_7) are statistically significant; ranging from -0.63 to 0.14 (see Figure 6).[175] A complete test of the influence of unobservables would entail the simultaneous consideration of mediation and nonrandom selection (A_1–A_7), but such a test is infeasible given the strong assumptions necessary to properly identify the additional correlations (A_5, A_6, and A_7) (see Figure 6); therefore, similar to my aforementioned discussion of the A_1–A_3 assumptions for the mediation analysis, these results should also be considered suggestive rather than conclusive.[176]

174. *See supra* Subpart II.A.2.

175. The models include gender, UGPA, LSAT, number of law school applications, number of law school acceptances, tier location, LGPA, mother's education, father's education, family income, and disability status.

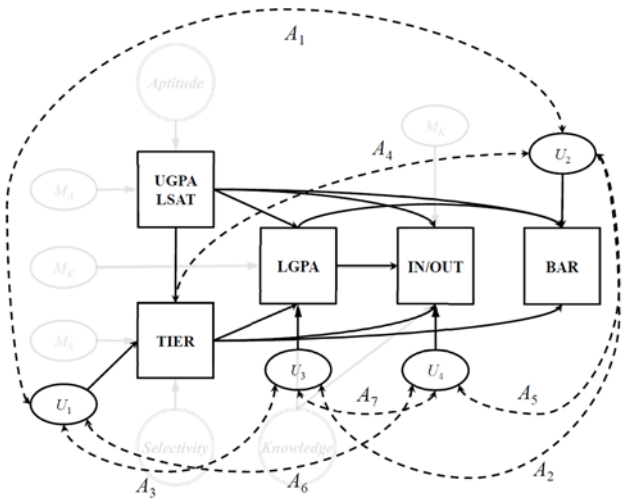
176. Amy Farley and colleagues' examination of several cohorts of students at the University of Cincinnati College of Law offers evidence that undermines Sander's claim that students who do not graduate (nongraduates) or who graduated but opted not take the bar (nontakers) should be coded as failing the bar when analyzing the impact of mismatch on bar passage. Amy N. Farley et al., *A Deeper Look at Bar Success: The Relationship Between Law Student Success, Academic Performance, and Student Characteristics*, 16 J. EMPIRICAL LEGAL STUD. 605, 617 fig.3 (2019). First, nongraduates had nearly identical UGPA/LSAT as graduates (nongraduates: 3.53/158; graduates: 3.51/159) but different law school grades (nongraduates: 2.69; graduates: 3.36). *Id.* Second, nontakers had very similar UGPA/LSAT/LGPA as takers (nontakers: 3.50/159/3.24; takers: 3.51/159/3.36). *Id.* Third, the nontakers had better UGPA/LSAT/LGPA than takers who failed the bar exam both instate and out of state (fail instate: 3.38/157/2.97; fail out of state: 3.42/156/3.02). *Id.* Finally, nontakers had very similar UGPA/LSAT/LGPA to takers who passed the bar exam both instate and out of state (pass instate: 3.51/159/3.44; pass out of state: 3.54/158/3.38). *Id.* It is obvious from these numbers that something other than mismatch is causing students to not graduate or not take the bar.

Figure 5: Sander's Causal Model



Note: Boxes represent observed variables; circles represent unobserved variables. Straight lines are possible causal effects; curved lines are correlations. Dashed lines indicate spurious relationships that Sander assumes do not exist. A1, A4–A6 identify the specific assumption.

Figure 6: Sander's Causal Model



Note: Boxes represent observed variables; circles represent unobserved variables. Straight lines are possible causal effects; curved lines are correlations. Dashed lines indicate spurious relationships that Sander assumes do not exist. A1–A7 identify the specific assumption.

B. Nonresponse Bias

Nonresponse bias occurs when values on variables are systematically (nonrandomly) missing. Sander's statistical models require that the data contain complete information on all variables used in the analysis for each respondent. If a value for any of those variables included in the model is missing, the analysis discards all of the information for that particular respondent (referred to "listwise deletion"). Three key problems may result from listwise deletion: (1) increased likelihood of failing to detect a statistically significant relationship (Type II error); (2) analysis of an estimation sample that is unrepresentative of the population from which the data were drawn; and (3) biased inferences from the estimation sample because the missing data may distort the relationships between the variables in the analysis.[177] Ho mentioned the possibility of nonresponse bias in Sander's work.[178] Specifically, Ho notes that Sander's analysis assumes that nonresponse to the survey questions is completely random.[179] If it is possible, however, to predict whether information for a variable is missing for an individual, then missingness is not completely random. King and colleagues emphasize that the "prediction required is not causal . . . [because] the purpose of an imputation model is to create predictions for the distribution of each of the missing values, not causal explanation or parameter interpretation."[180]

Sander's implicit assumption that the data missingness is completely random can be easily tested by modeling the likelihood that a respondent is missing a value for a particular variable with the other variables in his model. I performed this test on LGPA because nearly 7 percent of law school graduates had missing values for this variable.[181] Modeling missingness in LGPA with UGPA, LSAT, gender, race/ethnicity, family income, and tier location as explanatory variables, I discovered that all of these variables,

177. RODERICK J. A. LITTLE & DONALD B. RUBIN, STATISTICAL ANALYSIS WITH MISSING DATA 40 (1987) (explaining that listwise deletion can result in different magnitudes or signs of causal or descriptive inferences).

178. Ho, Evaluating Affirmative Action, *supra* note 23, at 10 n.7.

179. When data are missing completely at random (MCAR), it may be more difficult to get statistically significant results because the sample size is smaller, but results will not be biased. *See* LITTLE & RUBIN, *supra* note 177, at 40.

180. Gary King, James Honaker, Anne Joseph & Kenneth Scheve, *Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation*, 95 AM. POL. SCI. REV. 49, 51, 53 (2001).

181. I limit the analysis to respondents who graduated from law school because the focus of the analysis is on those missing values that result from participant nonresponse rather than structural nonresponse (such as a dropout). LGPA was missing for over sixteen hundred respondents who graduated from law school.

except LSAT and race/ethnicity, are statistically significant predictors of whether data are missing on LGPA. Recognizing that Sander's inferences are likely biased by nonrandom nonresponse, Ho emphasized that Sander should employ techniques developed specifically to ameliorate nonresponse bias.[182] Gregory Camilli and Darrell Jackson explicitly recognize this problem and properly account for nonresponse in the data.[183] They were able to partly quantify the extent of the nonresponse bias in the BPS data, reporting that "[a]n examination of the results revealed that the imputation process increased the error variance by about 30%."[184] The error variance reflects the variability that would be present in the data in the absence of missing data, so Camilli and Jackson's analysis has the advantage of minimizing bias in the estimation of the causal effects.[185]

C. Omitted Variable Bias

Posttreatment bias and nonresponse bias are special cases of omitted variable bias because adjusting for LGPA, subsetting the data based on law graduates/bar takers, and improperly accounting for missing data patterns can induce bias through unobservables even if the treatment was randomly assigned or conditionally exogenous (unrelated to unobservables after taking into account all relevant pretreatment variables) (Figure 4: A_2 & A_4). The most common (and widely discussed) type of omitted variable bias in nonexperimental studies occurs when the assumption of conditional exogeneity is implausible and there are unobserved common causes of both the treatment and the outcome(s). Ho criticized "Systemic Analysis" for failing to control for numerous variables that might simultaneously impact law school tier and the key outcomes of interest (LGPA and bar passage) (Figure 4: A_1 & A_3).

182. Ho, Evaluating Affirmative Action, *supra* note 23, at 10 n.7 (suggesting that Sander should employ a multiple imputation strategy, which makes educated guesses about those missing values based on the relationships between the data with the missing data and all other variables in the model and incorporates uncertainty in those predictions).

Nonresponse bias is akin to the aforementioned posttreatment bias arising from nonrandom sample selection. The key difference is that nonresponse bias need not result from the effect of the treatment. That is, in the posttreatment bias context, the missingness is caused (at least in part) by the treatment; however, this need not be the case for nonresponse bias.

183. Gregory Camilli & Darrell D. Jackson, *The Mismatch Hypothesis in Law School Admissions*, 2 WIDENER J.L. ECON. & RACE 165, 190 (2011).

184. *Id.*

185. *See* LITTLE & RUBIN, *supra* note 177, at 255.

Omitted variable bias is almost certainly operative in Sander's analyses because his statistical models rest on the implausible assumption that only a handful of variables capture the key differences between students that impact law school outcomes. For example, students who matriculate to elite law schools may differ along unmeasured credentials that are predictive of future academic and professional success.[186] Consequently, individuals are not necessarily outmatched at law schools simply because they tend to score lower than their classmates with respect to UGPA and LSAT, which are two narrow (and imperfect) proxies for academic preparedness and ability.[187] The assumption of the equivalency of treatment and control groups is much more plausible when researchers consider a host of pretreatment variables because the more similar individuals are across a wide range of observables, then the more alike they will be on unobservables.[188] The BPS contains information from a questionnaire administered to respondents during their orientation program, and "[s]ince the [q]uestionnaire was administered *before* law school began, most of these covariates are all plausibly pretreatment."[189] As Ho explains:

> The [BPS] Questionnaire contains a wealth of information (roughly 200 covariates) about the students in the dataset, including (a) personal and family background (e.g., student disabilities, whether the student's mother tongue is English, whether the student plans to attend law school full time, post-college full-time employment, prior legal employment, undergraduate work experience, female and male household head education and employment, family income, marital status, number of children, prior discrimination) (b) educational background (e.g., undergraduate major, prior degrees earned, year of graduation), and (c) financial status (e.g., financial dependents, prior loans, plans on working during law school).[190]

186. *See* Kidder & Lempert, *Mismatch Myth*, *supra* note 22, at 109 (noting that admissions officers have access to information beyond UGPA and LSAT scores, so students particularly strong along these non-UGPA/non-LSAT dimensions are likely to do better than students with similar UGPA/LSAT both during and after law school).
187. *See infra* Subpart II.F (describing the shortcomings of LSAT and LGPA as indictors of aptitude for legal education and acquisition of legal knowledge).
188. *See* Ho, *Reply to Sander*, *supra* note 23, at 2015 ("Unless the researcher gathers more data (as in my accompanying paper, which controls for 170 additional covariates), we have no idea what impact unobservables might have on the estimates." (footnote omitted)).
189. Ho, Evaluating Affirmative Action, *supra* note 23, at 8.
190. *Id.* Data were also collected on the selectivity of students' undergraduate institution (as measured by the Astin Index). This information was not provided in the once publicly available dataset; nonetheless, supplemental analysis provided by the BPS revealed that undergraduate selectivity was only weakly correlated with LSAT score, LGPA, and bar passage and, as a result, did not appreciably contribute to the predictive ability of the

Ho examines 180 pretreatment variables and finds that students in first tier schools significantly differed across a host of characteristics not accounted for in "Systemic Analysis" (nor are these variables accounted for in subsequent analyses performed by Sander).[191] Some of the key variables were wealth, age, desire to take the bar in different jurisdictions, English as their "best language," attending law school fulltime and during the day, financial dependents, marital status, and number of children. Ho also includes a variable indicating the region where the bar exam was taken, which was asked of study participants in a followup survey. This is an important variable because bar passage rates vary by state. Granted, region is an imperfect measure because it does not identify the specific state where bar exam was taken, but it should still appear in the models estimated by Sander.[192] After taking into account these and many other variables, Ho finds that the law school tier effect on bar passage rates vanishes. Camilli and Jackson reach a similar conclusion after considering a wider range of explanatory variables than Sander that are in the BPS, and their analysis also examined Asian American and Hispanic/Latinx students.[193] Jesse Rothstein and Albert Yoon, who also control for a richer set of variables than Sander in their analyses of the BPS data, find no differences between black and white students in terms of graduation rates, bar passage (if attempted), or employment outcomes (employed at least part-time, employed fulltime, and salary level).[194]

model. *See* WIGHTMAN, BAR PASSAGE STUDY, *supra* note 36, at 37, 40 n.70. The Astin Index defines strata for four-year colleges and universities based on the mean SAT or ACT score of entering freshman. Alexander W. Astin & James W. Henson, *New Measures of College Selectivity*, 6 RSCH. HIGHER EDUC. 1, 3 (1977).

191. Ho, Evaluating Affirmative Action, *supra* note 23, at 8 ("We match on 180 of these pretreatment covariates to assess the effect of a top-tier law school."); Sander, *Replication of Mismatch*, *supra* note 4, at 85 (failing to include the rich set of covariates identified by Ho that are available in the BPS data).

192. *See* Ho, Evaluating Affirmative Action, *supra* note 23, at 8 n.5 ("One might argue that students strategically select jurisdictions conditional on law school performance, but here we believe this bias to be small in comparison to omitted variable bias due to variation in jurisdictions.").

193. Camilli & Jackson, *supra* note 183, at 195 tbl.1, 195–97 (controlling for sex/gender, LSAT, UGPA, age, socioeconomic status, the number of lawyers in the family, law school paid for with loans, number of law schools to which the student applied, importance of school's academic reputation, importance of housing availability, and importance of number of minority students at a school).

194. Jesse Rothstein and Albert Yoon address selectivity by, essentially, estimating the probability of the candidates attending a selective school based on a host of factors: race, gender, age at matriculation, parents' education, disability status, ESL, gap year, fulltime work history, legal work history, undergraduate work history, father with white-collar occupation, and mother employed outside the home. Rothstein & Yoon, *supra* note 21, at 32 tbl.2. They find a strong "black" effect, indicating the presence of racial

The absence of a richer set of controls in Sander's analysis is especially puzzling for two reasons: (1) Sander has expressly acknowledged that a major advantage of the BPS data is that they contain "hundreds of variables on tens of thousands of law students"[195] and (2) other work by Sander on the impact of "mismatch" on attorneys' earnings includes a much richer set of controls.[196] Sander's seemingly cavalier approach to potential problems of omitted variable bias when assessing the impact of mismatch on bar passage is also directly at odds with his earlier statements about the influence of

considerations in admissions. Black applicants are, on average, 16 percentage points more likely to attend a selective school. *See id.* at 16. Jesse Rothstein and Albert Yoon note that their analysis captures, in a slightly different manner, what the relative tier analysis attempts to do. Jesse Rothstein and Albert Yoon find no mismatch effects and some reverse mismatch effects. *Id.* at 22–23 (concluding "that the use of affirmative action at these schools does not generate meaningful mismatch effects," and "[b]lack students are much more likely to obtain good jobs than are similarly-qualified white students, with a salary premium around 10–15 percent"); *accord* Ayres & Brooks, *supra* note 12 (reporting reserve-mismatch effects); Ho, Evaluating Affirmative Action, *supra* note 23, at 7 fig.3 (finding that black students in the BPS data in the elite tier have a better bar passage rate than black students in lower tiers with similar academic index scores, and this supports Ayres and Brooks reverse mismatch hypothesis; the difference is not statistically significant, however, and therefore may simply be the result of chance).

195. SANDER & TAYLOR, *supra* note 4, at 57.

196. Sander, *Systemic Analysis*, *supra* note 7, at 457 n.249, 463 tbl.7.3 (taking into account more than two dozen variables included in a study of nearly four thousand attorneys across the nation who passed the bar in 2000).

It is worth noting that Sander's findings concerning employment outcomes are contradicted by a vast literature demonstrating that affirmative action does not adversely impact employee performance and compensation. *See generally* BRUCE P. LAPENSON, AFFIRMATIVE ACTION AND THE MEANINGS OF MERIT 33–34 (2009) (summarizing studies finding that "employment establishments are not hurt by affirmative action"); *cf.* Major G. Coleman, *Merit, Cost, and the Affirmative Action Policy Debate*, REV. BLACK POL. ECON., Summer 1999, at 99 (reporting that the percentage of affirmative action hires does not negatively impact the profitability and efficiency of a business); Robert C. Davidson & Ernest L. Lewis, *Affirmative Action and Other Special Consideration Admissions at the University of California, Davis School of Medicine*, 278 J. AM. MED. ASS'N 1153 (1997) (noting that beneficiaries of affirmative action were just as likely as nonbeneficiaries to graduate from medical school, complete their residency, and receive positive performance evaluations from residency directors); Harry Holzer & David Neumark, *Are Affirmative Action Hires Less Qualified? Evidence From Employer-Employee Data on New Hires*, 17 J. LAB. ECON. 534 (1999) (finding no difference in performance ratings between affirmative action and non–affirmative action hires); Jonathan S. Leonard, *The Impact of Affirmative Action on Employment*, 2 J. LAB. ECON. 439 (1984) (finding that the use of affirmative action does not negatively impact business productivity); Nicholas P. Lovrich, Jr, Brent S. Steel & David Hood, *Equity Versus Productivity: Affirmative Action and Municipal Police Services*, PUB. PRODUCTIVITY REV., Autumn 1986, at 61 (discovering that police departments' aggressive use of affirmative action in hiring had no negative impact on operation costs, crime rates, and arrest rates).

unobservables. In his response to Ho's criticism of "Systemic Analysis," Sander states:

> We then confront a second problem [with Ho's analysis]: unobservable characteristics. Suppose we match a black student at the twentieth-ranked school against a black student at the thirtieth-ranked school by LSAT score, undergraduate GPA, and gender—the variables used by Ho. We still don't know key information about these students—in particular, their undergraduate college, their major, and their other skills and achievements. Because large majorities of law school applicants go to the most elite school that accepts them, it is very likely that the blacks at our twentieth-ranked school have stronger "unobservable" characteristics than do their counterparts at the thirtieth-ranked school. If so, Ho is comparing an academically stronger student with a weaker one. Thus, Ho is wrong when he suggests that we can view his matched students as experimental subjects "randomly assigned to a tier in an experiment." There are systematic, biasing reasons why one student is at the University of North Carolina and another with similar numbers is at Duke.[197]

Sander claims that the "first choice" analysis allows him to sidestep the omitted variable bias problem, but this is doubtful. Even assuming for the sake of argument that the aforementioned problem with the first choice analysis is not fatal,[198] scholars have explained that "many who choose not to attend their original first-choice law schools do so for financial reasons," and "[s]ince law schools use financial aid to lure top applicants who might otherwise go elsewhere, a portion of . . . second-choice students are likely to be particularly strong on both measured and unmeasured variables, and so they are likely to do better than their first-choice counterparts in law school and beyond."[199] And there is good reason to believe that admissions officers probe more deeply into characteristics other than UGPA and LSAT when deciding who

197. Sander, *Mismeasuring*, *supra* note 24, at 2010 (footnote omitted).

198. *See* Ayres & Brooks, *supra* note 12, at 1833 (explaining that the utility of the first choice framework to explore mismatch effects is seriously hampered because the data do not contain information on whether second (third) choice is at a more, less, or equally selective school as the first (second) choice school).

199. Kidder & Lempert, Mismatch Myth, supra note 22, at 109; see also Kevin Eagan, Jennifer B. Lozano, Sylvia Hurtado & Matthew H. Case, Higher Educ. Rsch. Inst., The American Freshman: National Norms Fall 2013, at 6 (2013), https://community engagement. uncg.edu/wp-content/uploads/2015/03/theamericanfreshman2013.pdf [https:// perma.cc/Z39G-6D34] ("Just over 40% of students said that being unable to afford their first-choice college was a 'very important' consideration in deciding to enroll in an institution other than their first-choice college.").

should receive financial aid (and to what degree) than they do in deciding whether to accept an applicant.[200] At minimum, Sander should test the sensitivity of his results to the possibility of omitted variable bias using several well-established causal bounds tests.[201] William Kidder and Richard Lempert provide the following thought experiment that nicely illustrates the deficiency of the first choice analysis:

> Consider two students with similar GPAs and LSAT scores, both of whom have been accepted at the Columbia and Northwestern law schools and both of whom would choose to attend Columbia, everything else being equal, because it is [six] or so places higher in the *U.S. News* rankings. Neither student stands out among Columbia's admittees, and Columbia offers neither a particularly generous financial aid package (in 1991 Columbia provided scholarships/grants to 45% of its enrolled minority students). Among Northwestern's applicants, however, the students are among the school's more attractive applicants and hence are candidates for a lucrative and prestigious tuition award (in 1991 Northwestern provided scholarships/grants to 76% of its minority students). Presumably Northwestern would offer its best financial packages to the student who based on all the information it has appears most able and most likely to be successful in law school and in her subsequent career. She receives the award and although she would have preferred to go to Columbia, having her tuition fully paid at Northwestern is too good a deal to turn down. In these circumstances, a comparison between the performance of the Northwestern and Columbia students, controlling for the admissions index, might be

200. Williams claims to address this concern by controlling for merit aid (with a dummy variable) and he reports that the effects for mismatch are even stronger "albeit slightly noisier." Williams, *supra* note 22, at 193–94 n.27. By slightly noisier, I assume he means the coefficients' level of statistical significance was lower. Williams is not entirely transparent in how he controls for merit aid, nor does he present the results from those analyses in his article. The BPS data include a question that asks respondents whether they attended a second (or even lower) choice school because the first choice school did not offer a large enough scholarship. This is a different question from whether the second (or lower) choice school offered sufficient merit aid to attract the student from their first choice school (and the amount of the award required to do so).

201. *See* PAUL R. ROSENBAUM, OBSERVATIONAL STUDIES 105 (2d ed. 2002) (describing various approaches to assess the sensitivity of the model to unobserved variables); Ho, *Reply to Sander*, *supra* note 23, at 2015 n.26 ("The standard advice in the literature is to first obtain balance on observables and then to conduct sensitivity analyses to examine to what degree inferences would change if some unobserved variable were correlated to the treatment and the outcome. In such an analysis, inferences could certainly change, but the actual truth cannot be ascertained from the data, because the data is by definition observed."). *See also infra* Subpart IV.B.

> misleading due to selection bias, *but the bias would favor the second rather than the first choice student*. Moreover, effects would be magnified since the cost of legal education is not irrelevant to the dependent variable of dropping out of law school, and graduating law school is a precondition for taking (and passing) the bar. Since Columbia is almost certainly a member of the BPS first tier of schools, but Northwestern could well be in the second tier, selection bias in this instance would favor rather than undercut the chance of finding apparent evidence of mismatch effects.[202]

A likely symptom of the underspecification (missing one or more important independent variables) of Sander's models is their low predictive ability.[203] David L. Chambers and colleagues highlight that Sander's models only correctly predict twenty-nine cases that passed the bar above what would have been predicted by chance alone.[204] This undermines Sander's claim that the statistically significant results are also substantively meaningful. It has been well established that "[s]tatistical significance may result from a small correlation and a larger number of [data] points. In short, the *p*-value does not measure the strength or importance of an association."[205] Moreover, given of the skewed nature of bar passage rates because nearly everyone eventually passes, greater attention needs to be given to the model's ability to correctly predict who fails the bar exam. Chambers and colleagues show that "Systemic Analysis" does very poorly at predicting "fails" (only 129 of 1074 fails (12 percent) were correctly predicted).[206] A more useful statistic to assess model fit in binary regression models is Tjur's *D*. The statistic simultaneously

202. KIDDER & LEMPERT, WHITE PAPER, *supra* note 22, at 8 n.11 (emphasis added) (footnote omitted). Williams claims that if students attend their second choice school because of cost considerations, "then it would be *harder* to find mismatch using second- or lower-choice indicator variable." Williams, *supra* note 22, at 186. As the quote by Kidder and Lempert underscores, first and second choice students are likely to differ in unobservable ways that bias findings in favor of mismatch, not against it. *Accord* Bjerk, *supra* note 23, at 4 ("[W]hatever caused second-choicers to not attend their first-choice must not be correlated with expected learning outcomes. This means, for example, it cannot be the case that second-choicers chose their second choice over the first choice because they saw that they would have better access to useful resources . . . or had strong connections with particular professors at their second-choice institution.").

203. *See infra* Appendix A.4 (discussing poor predictive ability of Sander's models).

204. Chambers et al., *supra* note 68, at 1870.

205. *Id.* at 1869 n.45 (quoting David H. Kaye & David A. Freedman, *Reference Guide on Statistics*, *in* REFERENCE GUIDE ON SCIENTIFIC EVIDENCE 333, 379 (Fed. Judicial Ctr. ed., 2d ed. 2000); *accord* PAUL CAIRNS, DOING BETTER STATISTICS IN HUMAN-COMPUTER INTERACTION 161 (2019) ("[I]t is better to look at effect sizes rather than *p*-values and to consider the meaningfulness and explanations behind effects as the outcomes of an exploration.").

206. Chambers et al., *supra* note 68, at 1871.

considers correctly predicted "passes" and "fails," and is preferable when the binary variable either has a lot of "1's" or "0's".[207]

Recent work by Sander underscores the extremely limited utility (poor model fit) of the first choice analysis in explaining both LGPA and first-time bar passage, although contrary to a reasonable interpretation of the data, he claims that "the law school mismatch hypothesis has been sufficiently well supported, from a wide variety of approaches."[208] Based on Sander's own analyses, the proportion of variance explained in first year law school grades ranges from 9 percent to 15 percent, and the proportion of variance explained in cumulative grades ranges from 14 percent to 18 percent.[209] Particularly telling is that Sander exclusively focuses on tests of statistical significance rather than the more important issue of the incremental improvement in model fit explained by including the first choice variable to the model. A similar criticism is applicable to his analysis of first-time bar passage and eventual bar passage rates. The proportion of variance explained for first-time bar passage ranges from 3 percent to 11 percent, and from 2 percent to 9 percent for eventual bar passage.[210] Chambers and colleagues note that the model fit Sander reports in "Systemic Analysis," which relied on a different analytical approach than first choice, was evidence of poor model fit (32.5 percent for the variance explained). Viewed individually and collectively, these model fit statistics (ranging from 2 to 18 percent) demonstrate that Sander's analytical models do a poor job of predicting outcomes and are highly suggestive of model misspecification.[211] The poor model fit statistics may also result from incorrect functional form assumptions, commonly known as interpolation and extrapolation bias.[212]

207. *See* Tue Tjur, *Coefficients of Determination in Logistic Regression Models—A New Proposal: The Coefficient of Discrimination*, 63 AM. STATISTICIAN 366, 370 (2009) (noting that the Tjur's *D* "is probably the measure of explanatory power that comes closest to the satisfaction of the eight ideal requirements [of a model fit measure]"); *see also* Tarald O. Kvålseth, *Cautionary Note About* R^2 39 AM. STATISTICIAN 279, 281 (1985) (listing desirable properties of a model fit statistic).
208. Sander, *Replication of Mismatch*, *supra* note 4, at 88.
209. *Id.* at 79 tbl.3 & tbl.4.
210. *Id.* at 80 tbl.5.
211. When model fit is particularly weak, it would be prudent for the analysts to incorporate prediction error in their estimated quantities of interest for individuals in the study. Gary King, Michael Tomz & Jason Wittenberg, *Making the Most of Statistical Analyses: Improving Interpretation and Presentation*, 44 AM. J. POL. SCI. 347, 349 (2000) (discussing how "researchers can . . . compute quantities of interest and account for uncertainty").
212. *See* WILLIAM D. BERRY & STANLEY FELDMAN, MULTIPLE REGRESSION IN PRACTICE 25 (1985) (explaining that rather than necessarily indicating the omission of theoretically relevant variables, poor model fit may result from misspecification of the functional form of the equation).

D. Interpolation Bias

The unconfoundedness/nonspuriousness assumption for credible causal inference—often called an "identifying assumption"—is more plausible in nonexperimental studies when more potential confounding variables are included in the analysis. A less appreciated, yet equally important, identifying assumption is that there is a nonzero probability of the nontreatment group receiving every level of treatment for every combination of values of covariates.[213] This is sometimes called the "positivity assumption" or the "overlap assumption." Treatment and control groups must be similar with respect to all of the relevant variables of interest so that the comparisons—the counterfactual—are close enough to the observed data to provide empirically supported answers. Even if positivity is only nearly violated because the probabilities of some treatment options are merely small for the treatment or control group (referred to as "weak overlap"), estimates of the treatment effect can be severely biased: "[V]iolations and near violations of the positivity assumption can increase both the variance [imprecision] and bias of causal effect estimates, and if undiagnosed can threaten the validity of causal inferences."[214] "It is not enough to imagine fruitful hypotheses. They must be carefully examined in relation to the known facts. . . . In examining the plausibility of a mental experiment, we must ask whether what is held constant is faithful to what actually happened."[215] A testable counterfactual statement is one in which its probability can be inferred from the data, so when counterfactuals questions invoke hypothetical questions that are incompatible with what is actually known or observed, "testing counterfactuals is fraught with conceptual and practical difficulties."[216]

It is often the case, however, that these two assumptions—unconfoundedness and positivity/overlap—are in tension with one another because covariate overlap is more difficult to achieve as more potential confounding variables

213. *See* King & Zeng, *supra* note 156, at 148–49.

214. Maya L. Petersen, Kristin E. Porter, Susan Gruber, Yue Wang & Mark J. van der Laan, *Diagnosing and Responding to Violations in the Positivity Assumption*, 21 STAT. METHODS MED. RSCH. 31, 32 (2012); *accord* GUIDO W. IMBENS & DONALD B. RUBIN, CAUSAL INFERENCE FOR STATISTICS, SOCIAL, AND BIOMEDICAL SCIENCES: AN INTRODUCTION 338 (2015) ("For covariate values with either few treated or few controls, it may be difficult to obtain precise estimates for treatment effects . . . .").

215. NYE, *supra* note 60, at 52.

216. Ilya Shpitser & Judea Pearl, *What Counterfactuals Can Be Tested*, *in* PROCEEDINGS OF THE TWENTY-THIRD CONFERENCE ON UNCERTAINTY IN ARTIFICIAL INTELLIGENCE 352, 352 (Ronald Parr & Linda C. van der Gaag eds., 2007).

are included in the analysis.[217] The root of the problem is that as more covariates are added to the model, stronger parametric assumptions about the relationships between those variables and the outcome are required to properly estimate their effects.[218] In statistics parlance, this is referred to as

217. *See, e.g.*, Gary King & Langche Zeng, *Empirical Versus Theoretical Claims About Extreme Counterfactuals: A Response*, 17 POL. ANALYSIS 107, 108–09 (2009) (explaining that credible causal inference requires treatment and control groups to be identical across all relevant covariates in the sample data being analyzed, and this becomes increasingly difficult as the number of covariates increases because random assignment only guarantees unconfoundedness, on average, across repeated experiments).

218. The following example, adapted with modification from Daniel E. Ho, Kosuke Imai, Gary King & Elizabeth A. Stuart, *Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference*, 15 POL. ANALYSIS 199, 209 (2007), should assist intuition about interpolation bias. Suppose we have a continuous dependent variable, *Y*, representing LGPA and a single six-category explanatory variable, *X*, representing the tier location of the law school, and the aim is to estimate the relationship between *X* and *Y* with linear regression, but without making any functional form (parametric) assumptions. To accomplish this, we need to estimate a total of six parameters: a parameter for the constant (intercept) and a parameter for each of the five dummy variables representing contrasts from the intercept. Equivalently, we could estimate a parameter for each for the six indicator variables describing the mean of *Y* for each value of *X*. Alternatively, the typical approach is to assume a linear relationship by directly including *X*, rather than the five indicator variables for each value of *X*. The simpler model requires the estimation of just two parameters: a parameter for the constant (intercept) and a parameter summarizing the change in conditional mean of *Y* for a unit-change in *X*. The appropriateness of a single parameter to capture the change in the conditional mean, however, is based purely on assumptions about a linear relationship between the two variables. If the underlying model only includes *X*, then estimating the four additional parameters may be feasible, but the inclusion of additional multivalued explanatory variables quickly makes the estimation of a separate parameter for each level of the multivalued variable unwieldy. For example, if another 38-category explanatory variable, *Z*, representing LSAT, *see infra* note 309 and accompanying text (explaining the LSAT scale used when the BPS was conducted), is included in the model, 228 parameters (6×38) must be estimated—a parameter for each unique pair of the values for the two variables that captures nonlinearity and interaction. If *W*, which represents the 26-category UGPA, is included in the model, then the number of estimated parameters becomes 5928 (228×26). Including gender in the model increases the number of parameters to 11,856 (5,928×2). *See also* Daniel E. Ho & Donald B. Rubin, *Credible Causal Inference for Empirical Legal Studies*, 7 ANN. REV. L. & SOC. SCI. 17, 26 (2011) (describing a model examining incarcerated person misconduct with eight variables that would require the estimation of over 69 million parameters in the absence of any parametric assumptions). Even if it is theoretically possible to estimate these parameters under a standard linear regression approach, it is impractical because social science datasets do not have enough observations. Slope coefficients can be biased to any degree if the functional relationship of *W* or *X* or *Z* is misspecified. In parametric causal inference, assumptions about the form of these relationships are necessary but seldom based on genuine knowledge and, as a consequence, there may be high levels of model dependence with insufficient reasons to adopt a particular set of assumptions. The parametric assumptions become more integral to the estimation of treatment effects if there are values of *W* or *X* or *Z* that must be interpolated for the control group.

"the curse of dimensionality"[219] because the model must simultaneously consider a large number of dimensions (variables) and "[t]he number of variables exceeds the number of observations that can be effectively exploited."[220] Stated differently, estimation of the model heavily relies on more dissimilar cases (more divergent counterfactuals) to measure the hypothesized treatment effect. "When groups differ sharply, regression may not credibly 'control' for confounding factors."[221] Empirical scholars have long been aware of problems associated with violations of the positivity assumption; they were initially identified in the causal inference literature thirty-five years ago by Paul Rosenbaum and Rubin,[222] and specifically identified in Sander's work on mismatch.[223]

Interpolation and extrapolation are, respectively, the processes of estimating a value of a function (for example, a mean or covariance) at a point inside or outside the range of observed values. Bias from interpolation and extrapolation (discussed in the next Subpart) occurs through correctly identifying the necessary control variables but failing to adjust for them properly.[224] Specifically, the region of interpolation is bounded by the minimum and maximum observed data points for all the variables in the model, whereas data points outside of the interpolation region are in the extrapolation region. To assist in building intuition about interpolation and extrapolation, as well as

---

When the dependent variable is categorical (noncontinuous), a nonlinear relationship will exist between the explanatory variable(s) on the natural metric of the dependent variable, and models that fail to a properly account for this nonlinearity also exhibit interpolation bias. *See infra* Appendix C (Functional Form Misspecification).

219. RICHARD BELLMAN, ADAPTIVE CONTROL PROCESSES: A GUIDED TOUR 94 (1961) (describing the curse of dimensionality as "a malediction that has plagued the scientist from earliest days"). The curse of dimensionality has two versions. The first is computational and relates to the memory burdens on computers to estimate the model irrespective of data suitability. The second is statistical and relates to the inadequacy of the data to properly estimate the model absent strong, and often unreasonable, assumptions. LARRY WASSERMAN, ALL OF NONPARAMETRIC STATISTICS 58 (2006).

220. RICHARD A. BERK, STATISTICAL LEARNING FROM A REGRESSION PERSPECTIVE 46 (2d ed. 2016).

221. Ho & Rubin, *supra* note 218, at 26; *see id.* at 26–27 (noting that "[f]or causal inference, the overwhelming recognition in applied statistics is that regression alone is fragile" and "regression [adjustment] does not amount to research design"); *see also* MORGAN & WINSHIP, *supra* note 58, at 88 ("[M]atching is usually introduced . . . as a nonparametric method of adjustment for treatment assignment patterns when it is feared that ostensibly simple parametric regression estimators cannot be trusted.").

222. Paul R. Rosenbaum & Donald B. Rubin, *The Central Role of the Propensity Score in Observational Studies for Causal Effects*, 70 BIOMETRIKA 41, 42 (1983).

223. *See infra* note 241 and accompanying text.

224. *See* King & Zeng, *supra* note 156, at 148. The widespread use of the terms interpolation bias and extrapolation bias is somewhat unfortunate because they are not especially intuitive without first understanding how the two concepts are related.

the problems they potentially pose for causal inference, I first provide written explanations of both phenomena and then conclude the discussion with a visual example. Guido Imbens and Rubin explain:

> An extreme case of imbalance occurs when the ranges of data values of the two covariate distributions by treatment differ, and as a result there are regions of covariate values that are observed in only one of the two treatment arms [extrapolation]. More typical, even if the ranges of data values of the covariate distributions in the two treatment arms are identical, there may be *substantial differences in the shapes of the covariate distributions by treatment status* [interpolation].[225]

When using a parametric model to adjust for potential confounding variables, interpolation bias arises from controlling for the confounders with the wrong functional form within the observed range of the data.[226] Even in the absence of omitted variable bias (and its various forms discussed earlier), interpolation bias will result if the model incorrectly postulates how the variables are related to the outcome. And this applies even if the form of the relationship between the treatment variable and outcome of interest is correct when "the multivariate density [distribution] of [the covariates] for the treatment group differs from that for the control group (within the region of interpolation)."[227] Political methodologist Jeremy Ferwerda and colleagues explain that in "parametric models, leaving out an important function of an observed covariate can result in the same type of omitted variable bias as failing to include an important unobserved confounding variable."[228] That is, if the parametric assumption is incorrect, then the statistical adjustment will fail to remove the bias from the included confounding variable(s).[229]

It is often the case that a parametric assumption provides an adequate approximation of the true relationship between variables, so minor departures from the actual form will not unduly impact parameter

225. IMBENS & RUBIN, *supra* note 214, at 337 (emphasis added).
226. A *functional form* is the conditional prediction of the dependent variable (Y) for a given value of an explanatory variable (X). In fully parametric regression models the functional forms are determined before fitting the model to the data. When the functional forms are determined as part of the fitting process itself, the models are nonparametric. A hybrid model in which some functional forms are determined a priori while others are determined as part of the fitting process are semiparametric. *See* BERK, *supra* note 220, at xvi.
227. King & Zeng, *supra* note 156, at 148.
228. Jeremy Ferwerda, Jens Hainmueller & Chad J. Hazlett, *Kernel-Based Regularized Least Squares in R (KRLS) and Stata (krls)*, J. STAT. SOFTWARE, July 2017, at 1, 2.
229. *See* King & Zeng, *supra* note 156, at 148 ("[I]nterpolation bias will exist if the density differences in [covariates] are not properly adjusted.").

estimates.[230] But it is extremely important to examine how much the counterfactual analysis (the "what if" question) depends on interpolation. When assessing a hypothesized treatment effect, the data may not contain any nontreatment group cases that have identical values on all relevant variables, save the treatment variable (or only contain very few cases). As a result, one cannot answer the "what if" question with observed data and must assume what the treatment effect would be if the control group were similar along all other relevant dimensions. Or in situations with very few cases, known as weak overlap, estimation of the treatment effect can be severely degraded because: (1) those cases exhibit insufficient variation across the range of included covariates in the model to properly isolate the effect of the treatment[231] and (2) the limited number of cases are unlikely to accurately represent the counterfactuals in the larger population. When the assumed data value (such as tier location) is bounded by the minimum and maximum observed values, the interpolation task is to "predict" what the effect would be for that particular value of the variable based on its conditional expectation derived from of the observed relationships in the "neighborhood" of the data. Without any parametric assumptions, the conditional expectation is totally unconstrained, so the imposition of a constraint through the specification of a functional form is necessary for the conditional expectation to be "smooth" (to have no sharp changes).[232] The consequence of the parametric assumption is to narrow the range into which the interpolated value can fall, thereby creating strong dependence on the correctness of the parametric assumption. To be clear, misspecification of the functional form of the relationships between variables need not require unobserved counterfactuals (or insufficiently observed counterfactuals) because an analyst can impose an incorrect parametric constraint, thereby biasing estimates, even when there are

230. Gary King & Langche Zeng, *When Can History Be Our Guide? The Pitfalls of Counterfactual Inference*, 51 INT'L STUD. Q. 183, 189 (2007) (explaining that interpolation is more constrained by the data than extrapolation, even when assuming the same conditional expectation relationship).

231. To reliably compute the conditional effect of the treatment on the outcome, (1) there should be a sufficient number of datapoints in the neighborhood of the other covariates in the model, and (2) those data points need to exhibit sufficient variation in the treatment. Otherwise, estimation of the treatment effect will essentially rely on interpolation or extrapolation of the effect of the treatment to an area of the data for which there is no observation or are only a few observations. Jens Hainmueller, Jonathan Mummolo & Yiqing Xu, *How Much Should We Trust Estimates From Multiplicative Interaction Models? Simple Tools to Improve Empirical Practice*, 27 POL. ANALYSIS 163, 167 (2019).

232. King & Zeng, *supra* note 230, at 188.

suitable counterfactuals in the data.[233] The key point is that causal relationships based on interpolated data are especially vulnerable to misspecification of the functional form because the data are not permitted to "speak for themselves." As scholars have cautioned, "[T]he greater the distance from the counterfactual to the closest reasonably sized portion of available data, the more model dependent inferences can be about the counterfactual."[234]

It has been "shown that methods such as linear regression adjustment can actually increase bias in the estimated treatment effect when the true relationship between the covariate and outcome is even moderately nonlinear, especially when there are large differences in the means and variances of the covariates in the treated and control groups."[235] Sander highlights substantial differences in entering credentials between white and black law students, but does not report any of the appropriate tests or conduct any of necessary adjustments to minimize or reduce the bias that results from employing regression analysis on data exhibiting these differences.[236] Reviewing Sander's work, Arcidiacono and Lovenheim note that "if there are cross-race differences in mismatch effects, generalizing these estimates to a sample of African American students could yield misleading conclusions about the extent of mismatch."[237] The estimates of black students will be biased

233. This occurs when the model itself is nonlinear because the dependent variable under investigation is noncontinuous—for example, binary (pass/fail) or ordinal (tier ranking). *See infra* Appendix C.

234. King & Zeng, *supra* note 230, at 188 (emphasis omitted); *see also* JOHN FOX, APPLIED REGRESSION ANALYSIS, LINEAR MODELS, AND RELATED METHODS 22, 425 (1997) (explaining that interpolation will negatively impact the precision of the parameter estimates (resulting in large standard errors) even if the parametric assumption is reasonable).

235. Elizabeth A. Stuart, *Matching Methods for Causal Inference: A Review and a Look Forward*, 25 STAT. SCI. 1, 3 (2010).

236. Sander, *Systemic Analysis*, *supra* note 7, at 415–16.

237. Arcidiacono & Lovenheim, *supra* note 7, at 17. Williams's analysis of mismatch effects repeats the same error by using "coefficients for whites . . . to predict the outcomes for each racial group using the actual credentials of the group." Williams, *supra* note 22, at 181 n.16.

The polychoric correlations between first-time bar passage and UGPA, LSAT, and LGPA are substantially stronger for white students than for black students. The correlations for UGPA (white students: 0.22, black students: 0.11) and LSAT (white students: 0.34, black students: 0.17) are approximately twice as large, and the effect of LGPA is 2.6 times as large (white students: 0.58, black students: 0.22). That is to say that UGPA, LSAT, and LGPA are much stronger predictors of first-time bar passage for white students than for black students. The correlation between first-time bar passage and tier location is similar for white students and black students (white students: -0.21; black students: -0.18). These correlations were virtually identical when students from the HBLS tier were excluded from the analysis, which is to be expected given that only 2.9 percent of students in the BPS attended law schools in the HBLS tier (18 percent of black students and 1.4 percent of white students). Recall that polychoric correlations take into

when there are crossracial differences because the mismatch effects measured by Sander are predominately based on white students (who compose 83.5 percent of the bar takers in the BPS data, whereas black students compose only 5.9 percent).[238] Political methodologist Donald Green and colleagues make this point most succinctly when they write, "The bottom line is that when subjects are governed by different causal laws, analyses that presuppose that the same parameters apply to all observations may yield biased results."[239] A nearly identical warning was given in 1990 by Linda F. Wightman, the principal investigator of the BPS, pertaining to predicting LGPA with the LSAT.[240] Ho

account the differing measurement scales of the variables, UGPA & LGPA (continuous), tier location (ordinal), and first-time bar passage (binary). *See supra* notes 109, 173 and accompanying text.

238. Michelle Landis Dauber's criticism of Sander's analysis of career outcomes for black lawyers underscores the problem of assuming the parametric relationships uncovered among white students equally apply to black students:

> [Sander] reports an analysis of roughly 1800 white lawyers and 200 black lawyers in order to estimate a model using law school grades and prestige to predict salaries. Therefore, white lawyers have much more weight in his results than do black lawyers, even though Sander wants to apply his conclusions solely to black lawyers.
>
> In using whites as stand-ins for blacks, Sander is implicitly claiming that whites and blacks face exactly the same labor market opportunities. But this (unstated) assumption is contradicted by reams of empirical evidence, both quantitative and qualitative, including research on labor markets for lawyers and other high-status professionals. . . .
>
> . . . While having 2000 cases for a regression model is a good thing because it increases the likelihood of finding statistical significance for real effects, it is of no use if the cases are the wrong ones. And even the magic of regression analysis can't turn Sander's 1800 white lawyers into black ones.
>
> . . . [I]ncluding "black" as a dummy variable means that Sander is allowing no way for the regression equation to testify about white/black differences in how the independent variables, notably grades and law school prestige, affect the dependent variable, second-year salaries. He is constraining racial difference to operate solely as an additive factor, through a pure discrimination effect. But race is more properly understood as a context variable, one that, if we are to heed the empirical literature . . . appears in the form of significant differences between blacks and whites in the operation of the factors affecting labor market outcomes.

Michele Landis Dauber, *The Big Muddy*, 57 STAN. L. REV. 1899, 1903, 1905 (2005) (footnote omitted). *See also generally* Sherod Thaxton, *Disentangling Disparity: Exploring Racially Disparate Effect and Treatment in Capital Charging*, 45 AM. J. CRIM. L. 95 (2018) (summarizing the statistical research literature on racially discriminatory practices across different life domains and underscoring the importance of carefully examining both the additive and interactive impact of race/ethnicity on discriminatory decisionmaking).

239. Green et al., *supra* note 106, at 206.

240. LINDA F. WIGHTMAN & DAVID G. MULLER, LAW SCH. ADMISSION COUNCIL, RESEARCH REPORT NO. 90-03, AN ANALYSIS OF DIFFERENTIAL VALIDITY AND DIFFERENTIAL PREDICTION FOR BLACK, MEXICAN AMERICAN, HISPANIC, AND WHITE LAW SCHOOL STUDENTS 25 (1990),

raised the issue of interpolation bias in "Systemic Analysis," and he and others suggest using nonparametric matching methods to avoid this problem (or test the sensitivity of parametric assumptions).[241] In fact, a closely related nonparametric matching procedure has been officially endorsed by the U.S. Food and Drug Administration as "Qualified for Scientific Use."[242]

Sander has expressly rejected Ho's suggestion to employ nonparametric matching for reasons that were insufficiently justified.[243] But as King and Zeng have emphasized, "Ultimately, whatever method of adjustment is used, the two [distributions of covariates] for the control and treatment groups need to be the same for interpolation bias be eliminated."[244] Prominent econometricians, including Nobel Laureate James Heckman, have noted the advantages of nonparametric matching to evaluate the impact of social programs: "Since nonparametric methods can be used to perform matching, the method does not, in principle, require that arbitrary functional forms be

https://files.eric.ed.gov/fulltext/ED468755.pdf [https://perma.cc/XNK7-8ZE2] ("When a regression equation is developed using combined data from white and minority students, the equation tends to overpredict law school performance for minority students."); *see infra* Subpart II.F (discussing both the random and systematic (nonrandom) measurement error associated with the LSAT).

241. *See* Ho, *supra* note 7, at 2001; Camilli & Jackson, *supra* note 183, at 170; Rothstein & Yoon, *supra* note 21, at 21; *see also* King & Zeng, *supra* note 156, at 149 ("Interpolation bias can also be adjusted for without a specified functional form via matching . . . . ").

242. Stefano M. Iacus, Gary King & Giuseppe Porro, *Causal Inference Without Balance Checking: Coarsened Exact Matching*, 20 POL. ANALYSIS 1 (2012) (noting that a nonparametric matching procedure called "Coarsened Exact Matching" was evaluated and endorsed by Office of Biostatistics & Epidemiology in the Center for Biologics and Research arranged for the FDA); *see also* W. G. Cochran, *The Effectiveness of Adjustment By Subclassification in Removing Bias in Observational Studies*, 24 BIOMETRICS 295 (1968) (introducing an earlier version of this nonparametric matching procedure to minimize bias from nonoverlapping groups).

243. *See* SANDER & TAYLOR, *supra* note 4, at 85 ("Ho used a 'matching' model to compare students who had similar academic indices but attended schools of differing eliteness . . . . However, in Ho's analysis he compared students attending schools very close to one another in their eliteness (and, thus, presumably very similar to one another in the level of mismatch facing black students). This is an obvious design flaw in Ho's approach; when this flaw is removed, the matching test shows levels of mismatch similar to those produced by other tests."); Sander, *Mismeasuring*, *supra* note 24, at 2010 ("Ho advertises his matching approach as a way to avoid bias. But in fact, because the BPS tiers overlap and because of the problem of unobservables, his method tends to *maximize*, rather than eliminate, bias. His technique and conclusions are thus invalid.").

Ho appropriately notes that Sander's criticism of his approach equally applies to Sander's own analysis because the mismeasurement of tier impacts all analyses that attempt to isolate the effect of tier. *See infra* note 326.

244. King & Zeng, *supra* note 156, at 149.

imposed to estimate program impacts."[245] Causal inference scholars have explained that "among social scientists who adopt a counterfactual perspective, matching methods are fast becoming an indispensable technique for prosecuting causal questions, even though they usually prove to be the beginning rather than the end of causal analysis on any particular topic."[246] This is because through matching:

> [R]esearchers can select a sample where the treatment and control samples are more balanced than in the original full sample [and] inferences are more robust and credible. . . . [This is] just like in the design phase of a randomized study [because] it precedes the phase of the study during which the outcome data are analyzed.[247]

The problem of interpolation bias—nonoverlap or very weak overlap—within the range of a single covariate across the law school tiers in the BPS data is shown in Figures 7 and 8. These figures plot the distributions of the academic index scores for all black and white students.[248] Figure 7 includes all students who enrolled in law school, whereas Figure 8 is limited to the students who eventually took the bar examination. Focusing solely on the region where the distributions intersect across the tiers, it becomes obvious that there are few students with overlapping index scores in nonadjacent tiers. This is true for both black and white students, but the limited overlap appears more pronounced for black students. To better understand how the distributions of the index scores across tiers impact causal inference about the effect of tier location on an outcome such as law school grades or bar passage, it is helpful to imagine a vertical line running across the distributions for each tier. The line's point of intersection across the tiers indicates the empirically based comparison group of similarly situated law students with respect to the academic index score across the various tiers. It should be clear that the data support a substantial number of comparisons between students from law schools in the top tier (Tier 1) and students in the next highest tier (Tier 2) in the overlapping region. This is less

245. James Heckman, Hidehiko Ichimura, Jeffrey Smith & Petra Todd, *Characterizing Selection Bias Using Experimental Data*, 66 ECONOMETRICA 1017, 1025 (1998); Guido W. Imbens, *Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review*, 86 REV. ECON. & STAT. 4, 11 (2004) ("[M]atching leads to consistent estimators for average treatment effects under weaker [assumptions than regression].").

246. MORGAN & WINSHIP, *supra* note 58, at 87.

247. IMBENS & RUBIN, *supra* note 214, at 337 (order of quotation edited for readability).

248. The academic index is simply a weighted combination of UGPA (40 percent) and LSAT (60 percent) that is normed to a 1000 point scale. *See also* Sander, *Systemic Analysis*, *supra* note 7, at 393 (explaining that an examination of admissions data of eight law schools from 2002 and 2003 confirmed the 60 percent LSAT and 40 percent weighing scheme).

so when comparing students from Tier 1 to students from Tiers 3–6. As a result, we can only be confident in inferences pertaining to the impact of tier location on outcome(s) of interest when adjusting for the academic index score if those comparisons are limited to situations when there are sizable treatment and control groups with similar index scores. Inferences based on balanced data are much more credible than those based on unbalanced data because they are less model-dependent, and therefore less likely to suffer from bias (or as much bias as analyses based on unbalanced data).

Most observational studies exploring potential causal effects adjust for more than a single variable. The inclusion of UGPA and LSAT separately, or the consideration of additional covariates (such as gender or family income) may further reduce the number of comparables across the tiers. Although the problems with employing an analysis that attempts to compare law students across nonadjacent tiers have been clearly articulated by scholars examining the impact of affirmative action on bar passage rates,[249] Sander continues to ignore this admonition in his recent work that, purportedly, addresses Ho's criticisms of his work.[250] The plausibility of the overlap assumption in the multivariate context can also be explored graphically by calculating the relative frequencies of the probability of a treatment assignment conditional on a set of covariates.[251] The probability of assignment statistic, often called a propensity score, is calculated for both treated and nontreated participants. Participants sharing a similar propensity score are considered comparable even though they may differ on values of specific covariates.[252] Propensity scores play a central role in causal inference because, when used appropriately, they serve as a balancing score representing the list of observed variables and avoid the curse of dimensionality problem by reducing a potentially long list of covariates to a one-dimensional score. By comparing cases with similar propensity scores, the analysis focuses on participants who are (nearly) equally likely to attend a law school in the same tier.

249. *See, e.g.*, Ho, *supra* note 7, at 2003; Kidder & Lempert, *Mismatch Myth*, *supra* note 22, at 110.
250. Sander, *Replication of Mismatch*, *supra* note 4, at 88 (finding no difference in first-time bar passage rates for black students when comparing adjacent tiers).
251. *See* IMBENS & RUBIN, *supra* note 214, at 319. In the context of Sander's work on mismatch, the probability of treatment assignment is the probability of enrollment in a specific law school tier.
252. Treated and control participants with the same propensity score may have differing values on an observed covariate, but the differences will be chance differences rather than systematic ones if the propensity score is estimate properly. Gary King & Richard Nielsen, *Why Propensity Scores Should Not Be Used for Matching*, 27 POL. ANALYSIS 435, 450 (2019) (noting that when used without appropriate caution, propensity score can increase rather than decrease bias in causal analysis).

As noted earlier, interpolation bias is present when the distributions of the covariates, which are summarized by the propensity score, substantially differ between the law school tiers within the region of interpolation. In extreme cases, the propensity scores will be concentrated close to 0 or 1, indicating that there is very little overlap in the distribution of covariates. That is, the more dissimilar the curves, the stronger the evidence of nonoverlap or very weak overlap.[253] Figures 10, 11, and 12 display the distribution of the probabilities of a black student attending a law school in the first (top) tier compared to lower tiers, after controlling for UGPA, LSAT, gender, undergraduate major, family income, and intention to attend law school part-time.[254] These figures clearly show that shape of the distributions of covariates are more similar when comparing the top tier (Tier 1) to the next highest (Tier 2); when comparing nonadjacent tiers, however, there is much less covariate balance. In fact, the covariate imbalance becomes quite drastic when comparing Tier 1 to Tier 3. This is especially telling because Sander and Williams advocate comparing Tier 1 to much lower tiers (Tiers 5 and 6) and the mismatch effects they report with respect to first-time bar passage are limited to those distant tier comparisons. Figures 13 and 14 illustrate distribution of probabilities of attending a law school in a particular tier compared to an adjacent tier. Specifically, comparisons are for Tier 2/Tier 3, Tier 3/Tier 4, Tier 4/Tier 5, and Tier 5/Tier 6. The comparison for Tier 1/Tier 2 is depicted in Figure 10. These figures reveal that the shapes of the multivariate distributions are much more similar for adjacent tiers, so counterfactual inferences positing the effect of attending an adjacent tier are much more likely to be supported by the actual data rather than merely by the assumptions of the model. By contrast, inferences about the effect of attending schools of different tiers are much less likely to be supported by the data when comparing students attending schools in nonadjacent tiers.

It must be emphasized that the bulk of the imbalance has been demonstrated within the region of interpolation. In the multivariate context, because the propensity score provides a summary of the covariate distribution, the interpolation region is the region where the range of the

253. Relatedly, one can examine these densities after balancing the data to determine whether sufficient overlap between the treatment and control has been achieved. Inferences based on balanced data are much more credible that unbalanced data because they are less model-dependent, and therefore less likely to suffer from bias (or as much bias as analyses based on unbalanced data). *Id.*

254. I experimented with various combinations of variables to estimate the propensity score—including, for example, parent's level of education, having a lawyer in the family, marital status, and number of children. The results were very similar across these various combinations.

propensity scores overlap. That is, there are participants in the control group tier with propensity scores that lie within the range of the propensity scores for participants in the treatment group tier, even though there may be very few participants in either that satisfy that condition. The figures also reveal that it is often the case with the BPS data that there are no comparables (or almost no comparables) across treatment and control groups within the overlapping range of propensity scores. Participants with propensity scores outside of this overlapping region are depicted by the hollow bar (see discussion of extrapolation bias in Subpart II.E).

The importance of employing nonparametric matching methods[255] to achieve covariate balance when conducting research on mismatch effects in educational settings outside of law schools has been advocated by various scholars because this method "makes no assumptions about the functional form of the dependence between the outcome of interest and [the explanatory variables]."[256] But even if the nonparametric approach is rejected, Ho and others suggest that Sander should have examined his results' sensitivity to the linearity and nonadditivity assumptions for the impact of GPA, LSAT, and tier location.[257] It is also unclear why Sander's models include inconsistent functional form assumptions (such as nonlinear or interaction effects in some models but not others).[258] Jesse Rothstein and Albert Yoon, for example, find

255. Propensity score matching is classified as a nonparametric matching method, although a parametric regression model (logit, probit, or complementary log-log) is used to estimate the propensity score.

256. Sigal Alon & Marta Tienda, *Assessing the "Mismatch" Hypothesis: Differences in College Graduation Rates By Institutional Selectivity*, 78 SOCIO. EDUC. 294, 299 (2005); *see also* King & Zeng, *supra* note 156, at 149 ("Interpolation bias can also be adjusted for without a specified functional form via matching, inverse propensity score weighting, or nonparametric smoothing.").

257. Camilli & Jackson, *supra* note 183, at 207 (noting that "regression analyses of the kind conducted by Sander are incapable of producing credible estimates of causal effects," including Sander's "strong assumptions of linearity"); Ho, Evaluating Affirmative Action, *supra* note 23, at 4 ("The logistic model, employed by Sander, makes particular functional assumption to identify a causal effect. . . . This implies that the average treatment effect and covariates enter the logistic link linearly and additively."); *see also* Katherine Y. Barnes, *Is Affirmative Action Responsible for the Achievement Gap Between Black and White Law Students*?, 101 NW. U. L. REV. 1759, 1774–75 (2007) ("Allowing a flexible function for the measurement of credentials means that I do not assume the relationship between student credentials and LSAT or UGPA is linear. As the LSAT is explicitly set on a nonlinear curve and many undergraduate institutions do no grade linearly, this is necessary to model student credentials accurately.").

258. One might be struck by the inconsistency in which Sander's models include (or exclude) nonlinear and nonadditive effects for not only UGPA and LSAT, but for other effects as well (such as interaction between race and other key explanatory variables). Perhaps these omissions are justified, but in most cases, Sander provides insufficient information

statistically significant nonlinear effects for UGPA and LSAT, and interactions between the two in many of their specifications. Interestingly, in Sander's reply to his initial critics, he notes the importance of curvilinear relationships between LGPA and graduation rates, as well as LGPA and bar passage rates, but to the best of my knowledge, there is no mention of this concern in *Mismatch* and other related work.[259] For a discussion of additional ways in which Sander's implausible functional form assumptions undermine his analyses, see Appendix C.

E. Extrapolation Bias

Extrapolation bias also impacts Sander's analysis.[260] In instances of interpolation and extrapolation, the model basically pretends that there are data values that do not actually appear in the sample (or are extremely sparse in the sample), and often minor changes in the specification of the model may substantially alter the results.[261] Consequently, regression analyses may

to assist the reader in making this determination. *Compare, e.g.*, Sander, *Systemic Analysis*, *supra* note 7 (failing to provide information about specification tests for nonlinear and nonadditive effects), Sander, *Racial Paradox*, *supra* note 113, at 1755 (same), *and* Richard Sander & Robert Steinbuch, *Mismatch and Bar Passage: A School Specific Analysis* (UCLA Sch. of Law Pub. Law & Legal Theory, Research Paper No. 17-40, 2017), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3054208 [https://perma.cc/M9XN-3C7N] (same), *with* Sander, *Mismeasuring*, *supra* note 24 (failing to explain the omission of curvilinear and interaction effects for his main model), Sander, *Reply to Critics*, *supra* note 10, at 2004 (examining curvilinear effects but not interaction effects), Sander & Doherty, *supra* note 80 (examining interaction effects but not curvilinear effects), *and* Sander & Bambauer, *supra* note 77 (examining some curvilinear and interaction effects, but failing to address the exclusion of others).

259. *See, e.g.*, Sander, *Replication of Mismatch*, *supra* note 4 (failing to mention curvilinear effects of grades and LSAT).

260. *See* Ho, Evaluating Affirmative Action, *supra* note 23 (highlighting extrapolation bias in Sander's analyses of mismatch); *accord* Camilli & Jackson, *supra* note 183; Rothstein & Yoon, *supra* note 21.

261. Minor specification changes typically entail adding or removing a variable, focusing on a subset of observations—including an interaction effect—or including a nonlinear (as opposed to linear) relationship. FRANK E. HARRELL, JR., REGRESSION MODELING STRATEGIES: WITH APPLICATIONS TO LINEAR MODELS, LOGISTIC AND ORDINAL REGRESSION, AND SURVIVAL ANALYSIS 109 (2d ed. 2015) ("Three major causes of failure of the model to validate are overfitting, changes in measurement methods/changes in definition of categorical variables, and major changes in subject inclusion criteria."); *see generally* King & Zeng, *supra* note 156; *see also* IMBENS & RUBIN, *supra* note 214, at 337 (describing problems created by lack of covariate balance between the treatment and control groups); Heckman et al., *supra* note 245, at 1072–73 ("Lack of common support—comparing the incomparable—is a major source of selection bias as it is conventionally measured. . . . Using a common support . . . goes a long way toward improving the performance of any econometric evaluation estimator.").

increase bias in estimates of the causal effect when there are large differences between the treated and untreated groups with respect to the average values and interrelationships of the noncausal variables.[262] Model dependence, then, is a function of the distance from the counterfactual to the observed data.[263] That is, the parametric assumptions of the model, rather than actual data, drive the results; therefore, the less justified the parametric assumptions, the more untethered the counterfactual. Sander's models extrapolate from the bounds of the data, so the parametric assumptions are less constrained by the data than in the context of interpolation because interpolation relies on data in the observed range of the treatment levels, and assume values for the nontreatment group that lies within the range.[264]

More concretely, extrapolation is required in order to, without adequate information, estimate how black students with a particular academic index score would perform on the bar exam relative to other black (or white) students with the same academic index score in a different tier without having any actual comparables in the relative tier that lie within the range of the overlapping academic index scores for both tiers (in other words, outside the region of interpolation). King and Zeng explain:

> One way to look at this is that the same level of smoothness [the assumed functional form] constrains the interpolated value more than the extrapolated value, as for interpolation any change in direction must be accompanied by a change back to intersect the other observed point. With extrapolation, one change need not be matched with a change in the opposite direction, as there exists no observed point on the other side of the counterfactual being estimated.[265]

Extrapolation bias can be addressed by restricting the analysis to cases in which there is "common empirical support" by focusing on cases in which the range of values across all of the variables in the model overlap for both the

262. Stuart, *supra* note 235, at 3.
263. *See* King & Zeng, *supra* note 156, at 135.
264. *See* Ho, Evaluating Affirmative Action, *supra* note 23, at 4 ("If observable covariates are unbalanced between treatment and control cases, these functional form assumptions may loom large, extrapolating from the bounds of the data."); *see also* Heckman et al., *supra* note 245, at 1025 ("The supports of the distributions of X may be different in the two groups and the shapes of the distributions may be different over regions of common support."); King & Zeng, *supra* note 156, at 150 ("If we use the data outside the region of common support, we must extrapolate and will therefore have some degree of model dependence and thus risk some bias for almost any model chosen.").
265. King & Zeng, *supra* note 230, at 189.

treatment and the control group.[266] The nonparametric matching approach described above restricts the matched data to areas of common empirical support, thereby "remov[ing] the possibility of difficult-to-justify extrapolations of the causal effect that end up being heavily model dependent."[267] The significant problems that extrapolation bias presents for causal inference have been underscored in econometrics for quite some time. Internationally acclaimed econometrician James Heckman explained that "[m]atching methods that impose the condition of pointwise common support . . . necessarily eliminate the bias arising from regions of nonoverlapping support."[268] Marta Tienda noted that the matching approaches described in the previous Part eliminate extrapolation bias because "matching ensures that the [values of the explanatory variables] in the treatment group are similar (matched) to those in the control group."[269] Simply stated, matching makes it plain whether or not comparable untreated observations are available for the treated observation.

Given the small number of black students in some of Sander's models (fewer than three hundred for some models), there is a strong risk of extrapolation bias. In fact, Jesse Rothstein and Albert Yoon find that there is very little overlap in credentials between black and white students in the lower tiers.[270] When this is taken into account, both Ho and Jesse Rothstein and

266. *See* Ho, Evaluating Affirmative Action, *supra* note 23, at 5 ("The central idea is to identify students that are similar in observable covariates, since the only way to identify the causal effect of attending a high-tier school is by examining comparable students in lower-tier schools.").

267. Matthew Blackwell, Stefano Iacus, Gary King & Giuseppe Porro, *CEM: Coarsened Exact Matching in Stata*, 9 STATA J. 524, 528 (2009).

268. Heckman et al., *supra* note 245, at 1031.

269. Alon & Tienda, *supra* note 256, at 299.

270. Jesse Rothstein and Albert Yoon note that approximately three-quarters of black students in the BPS sample have entering credentials in the bottom quintile of the credential distribution for all students, and black students' underperformance is attributable to the poor outcomes of students in this bottom quintile. Rothstein & Yoon, *supra* note 21, at 3, 19. These black students do not attend highly selective law schools, even with race-conscious affirmative action policies. *Id.* Among students with entering credentials placing them in the top four quintiles, black and white students with similar credentials graduate from law school and pass the bar exam at similar rates. *Id.* It is also difficult to reliably compare black and white students within the bottom quintile because there are so very few similarly situated students from the two groups due to the fact that "poorly [credentialed] white applicants have much greater difficulty gaining admission than do similarly [credentialed] blacks," *id.* 3–4, and black students in the lower tail of the credential distribution were twice as likely as similarly credentialed white students to be admitted to law school. *Id.* at 17. This is likely attributable to the fact that even the least selective schools apply lower thresholds for admission to black than to white applicants."

Albert Yoon find no support for mismatch effects.[271] According to the BPS, the median (50th percentile) academic index score in the fifth tier is 625; 54.9 percent (1112 of 2023) of white students in the fifth tier had an index score over 625, compared to 9.3 percent (10 of 107) of black students (I exclude the sixth tier, which is exclusively compromised of HBLSs, for reasons explained earlier).[272] Examining the high and low ends of the index score distribution in the fifth tier is even more illuminating. Whereas 28.4 percent (575 of 2023) of white students in the fifth tier have an academic index score of 680 or higher (75th percentile of index scores in the fifth tier), only 4.7 percent (5 of 107) of black students in the fifth tier have a similar score. Slightly over 12 percent (249 of 2203) of white students in the fifth tier have an index score exceeding 740 (roughly the 90th percentile), but only 2.8 percent (3 of 107) of black students in the fifth tier have a score above 740. When exploring the bottom end of the index score distribution in the fifth tier, one discovers that 1.4 percent of white students (29 of 2203) have an index score under 480 (the 5th percentile of the academic index score), compared to 64.5 percent (69 of 107) of black students in the fifth tier. When examining all tiers in the BPS, there are 1010 students without a crossracial comparables in the data with respect to their index score.[273] And even when there is some overlap across black and white students with respect to their index scores, a significant portion of the overlap is weak, which means there are very few cases to compare and, as a result, inferences based on those comparisons are unreliable. Figure 7 illustrates the degree of (non)overlap in academic index scores for all black and white students who participated in the BPS study

Id. Only 57 percent of the 90,335 law school applicants in the BPS cohort were admitted to any law school. Wightman, *supra* note 161, at 4.

271. The imbalance between top tier and non–top tier students in the raw BPS data on UGPA, LSAT, gender, and race is reduced via nonparametric matching, resulting in an effective sample size of $n = 3212$ (from $n = 20{,}827$). Ho, Evaluating Affirmative Action, *supra* note 23, at 8 tbl.1. Ho examined a much larger list of potential confounding variables and presented balance statistics on sixty-two variables for which there were statistically significant differences between the treatment and control groups with t-statistics greater than 4 (a t-statistic of 4 is significant at $p < 0.0001$ level [two-tailed]). *Id.* at 11 tbl.2.

272. *See infra* note 320 and accompanying text (explaining that students attending HBLS are invalid comparables for the purposes of Sander's analysis).

273. There are sixteen black students with index scores below any white students, and 1233 white students with index scores above any black students. For black students, the range is [225.61, 907.72] for academic index score and [11, 48] for LSAT score. For white students, the ranges are [302.11, 1000] for academic index and [17, 48] for LSAT score. Consequently, there are no comparable black and white students with academic index scores under 302.11 and over 907.72. Similarly, there are no comparable black and white students with LSAT scores under 17. *See supra* notes 270–271 and accompanying text.

across the six law school tiers. Figure 8 shows the same index score distributions, but the data are limited to black and white students who ultimately sat for the bar exam. Both figures underscore Jesse Rothstein and Albert Yoon's observation: there are virtually no black-white comparables in the lower tiers (and very few comparables at the very top of the index score distribution).

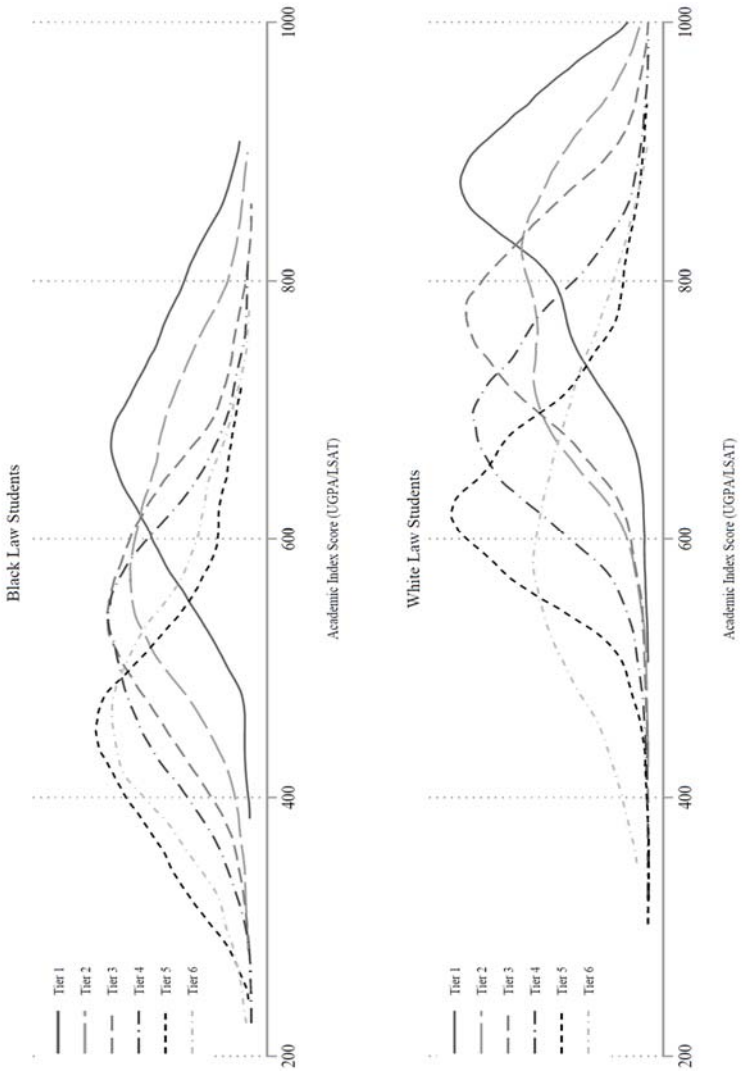


Figure 7: Index Score Distribution (All Enrolled Students)

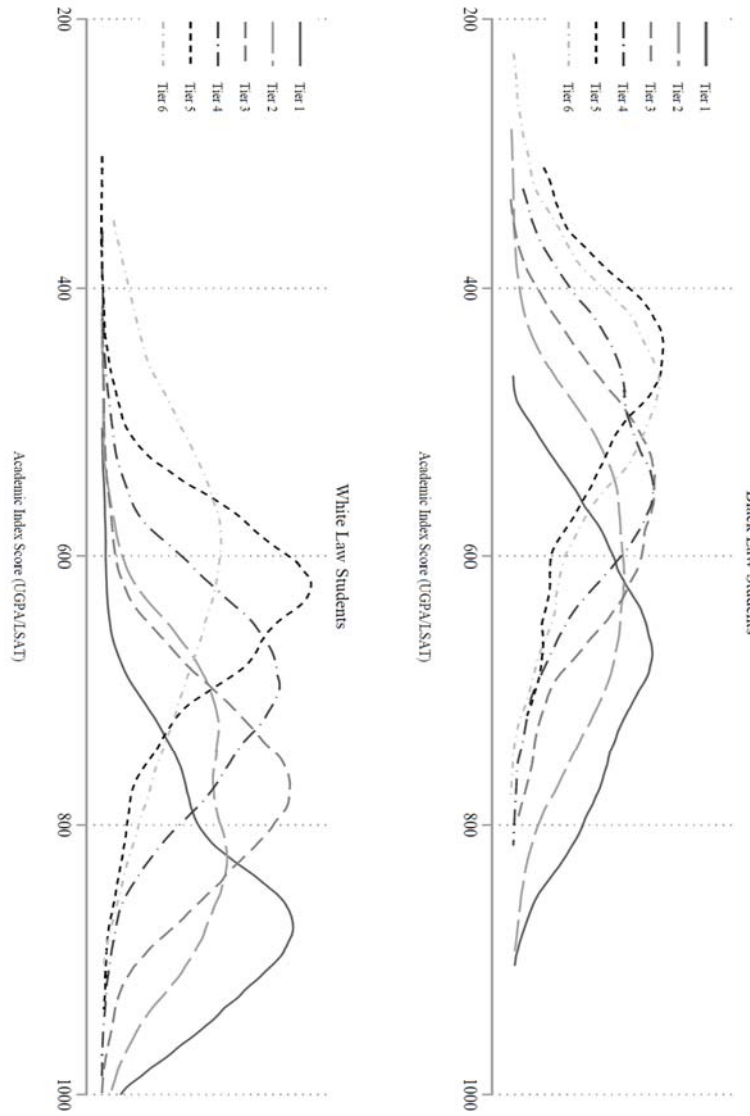


Figure 8: Index Score Distribution (All Bar Takers)

The extrapolation bias problem is not alleviated by focusing exclusively on black students. The median index score (50th percentile) for black students in the BPS is 540. For Tiers 1–6, the percentage of black students with an index score above 540 are, respectively, 98.6 percent, 73.4 percent, 55

percent, 44.9 percent, 17.8 percent, and 23.8 percent.[274] For black students with index scores that are above the 75th percentile (620), the distribution across Tiers 1–6 are, respectively, 77.2 percent, 43.8 percent, 24.1 percent, 16 percent, 10.3 percent, and 13.8 percent. Black students with index scores, approximately, at or above the 90th percentile (690) across Tiers 1–6 are 46.2 percent, 23.7 percent, 7.7 percent, 6 percent, 4.8 percent, and 5.5 percent.

As these percentages reveal, there can be very few within-race comparables for black students across tiers, so causal inference is heavily based on parametric assumptions of the model that are difficult to justify. Figure 9 graphically depicts the degree of overlap (and nonoverlap) in academic index scores for black students across the six law school tiers (for both enrolled students and bar takers). The curve farthest to the right represents the distribution of academic index scores in the top tier (Tier 1), whereas the curve farthest to the left displays the distribution of the academic index scores for Tier 5. The tier second farthest to the left consists entirely of HBLS (Tier 6). This graphic clearly illustrates that the bulk of black students with high academic index scores are concentrated in the top two tiers (Tiers 1 and 2), and as a result, there are very few comparable black students in the lower tiers to draw empirically-driven inferences about the impact of tier location on LGPA and bar passage.

274. Again, black students in the sixth tier, which is solely composed of HBLS, have higher academic index scores than black students in the fifth tier.

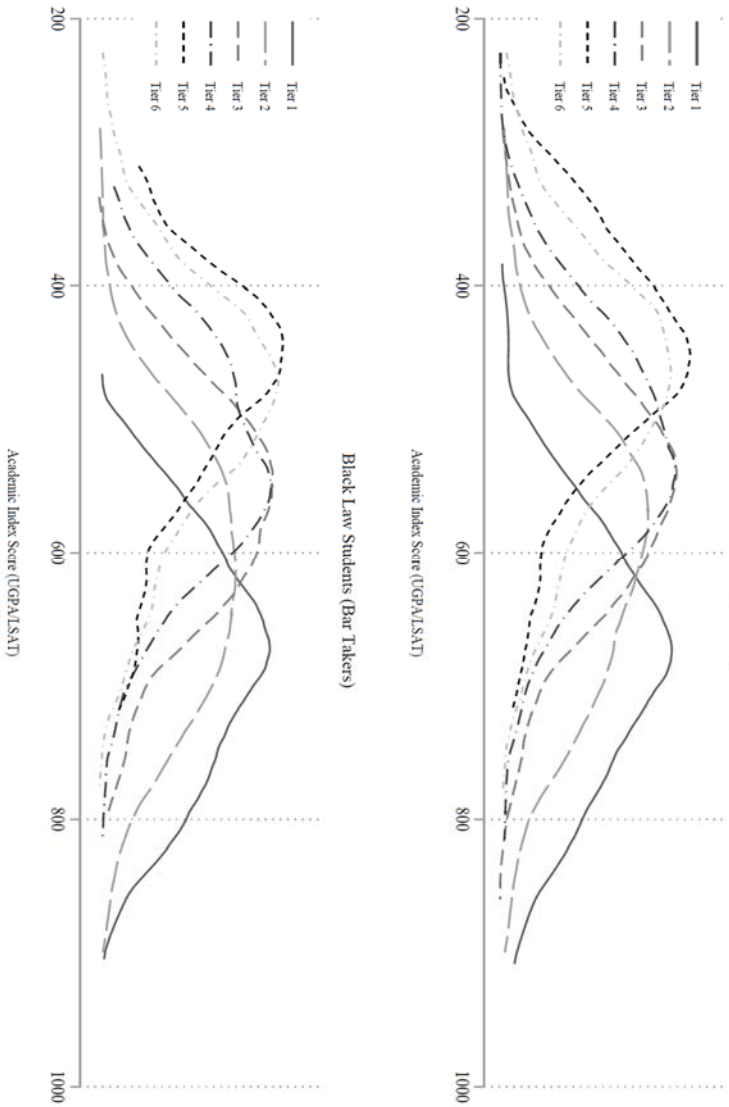


Figure 9: Index Score Distribution (Enrolled & Bar Takers)

It must also be emphasized that I focus on a single variable, academic index score, in the above analysis. The inclusion of UGPA and LSAT separately, or the consideration of additional covariates, such as gender, fulltime status, family income, would further reduce the number of comparables across the tiers. Although the problems with employing an analysis that attempts to compare law students across nonadjacent law school tiers have been clearly articulated by scholars examining the impact of

affirmative action on bar passage rates,[275] Sander continues to ignore this admonition in his recent work that purports to address Ho's criticisms of his work.[276]

The plausibility of the overlap assumption in the multivariate context can also be explored graphically by calculating the relative densities (or frequencies) of the probability of receiving each treatment level based on the values of a collection of covariates.[277] The overlap assumption is satisfied when there is a strong likelihood of witnessing observations in both the treatment and control groups (such as different law school tiers) at each combination of the covariate values. There is evidence that the overlap assumption is violated when an estimated density has too much mass around 0 or 1 because the estimated densities have relatively little mass in the regions where they overlap. That is, the less the distributions overlap on the graph, the stronger the evidence of the implausibility of the overlap assumption or very weak overlap. Relatedly, one can examine these densities after balancing the data to determine whether sufficient overlap between the treatment and control has been achieved.

The relative densities of the probability of a black student attending a law school in the first tier compared to lower tiers, after controlling for a host of relevant variables,[278] are depicted in Figures 10, 11, and 12, while Figures 13 and 14 display the relative densities for adjacent tiers (exclusive of Tier 1/Tier 2 comparison, which is depicted in Figure 10). The overlap assumption appears plausible when comparing adjacent tiers. When comparing nonadjacent tiers, however, there is much less covariate balance. The hollow bars in these figures represent the proportion of students in Tier 1 without comparables in the relative tier, based on the variables included in the model. In fact, a comparison of Tier 1/Tier 3 (Figure 10, right column) reveals both weak overlap and a substantial nonoverlapping region, and similar nonoverlap is observed when comparing to Tier 1/Tier 4. Comparisons of Tier 1/Tier 5 (Figure 11) and Tier1/Tier 6 (Figure 12) highlight an extremely large nonoverlapping region that reveals a sizable portion of the students in Tier 1 lack any comparabes in the data. Collectively, these figures demonstrate that inferences concerning similarly situated students who could have been admitted to a Tier 1 law school, but elected to attend a law school in the third (or lower) tier are not informed by the data. By contrast, Figures

275. *See* Ho, *supra* note 7, at 2003; Kidder & Lempert, *Mismatch Myth*, *supra* note 22, at 110.
276. Sander, *Replication of Mismatch*, *supra* note 4, at 87 (finding no difference in first-time bar passage rates for black students when comparing adjacent tiers).
277. IMBENS & RUBIN, *supra* note 214, at 319.
278. See *supra* note 254 and accompanying text for the description of the variables included in the propensity score model.

10 (left column), 13, and 14, which provide comparisons for adjacent tiers (Tier 1/Tier2, Tier 2/Tier 3, Tier 3/Tier 4, Tier 4/Tier 5, and Tier 5/Tier 6), show that the covariate overlap is much greater, and consequently, counterfactual inferences positing the effect of attending an adjacent tier are much more likely to be supported by the actual data rather than merely by the assumptions of the model.

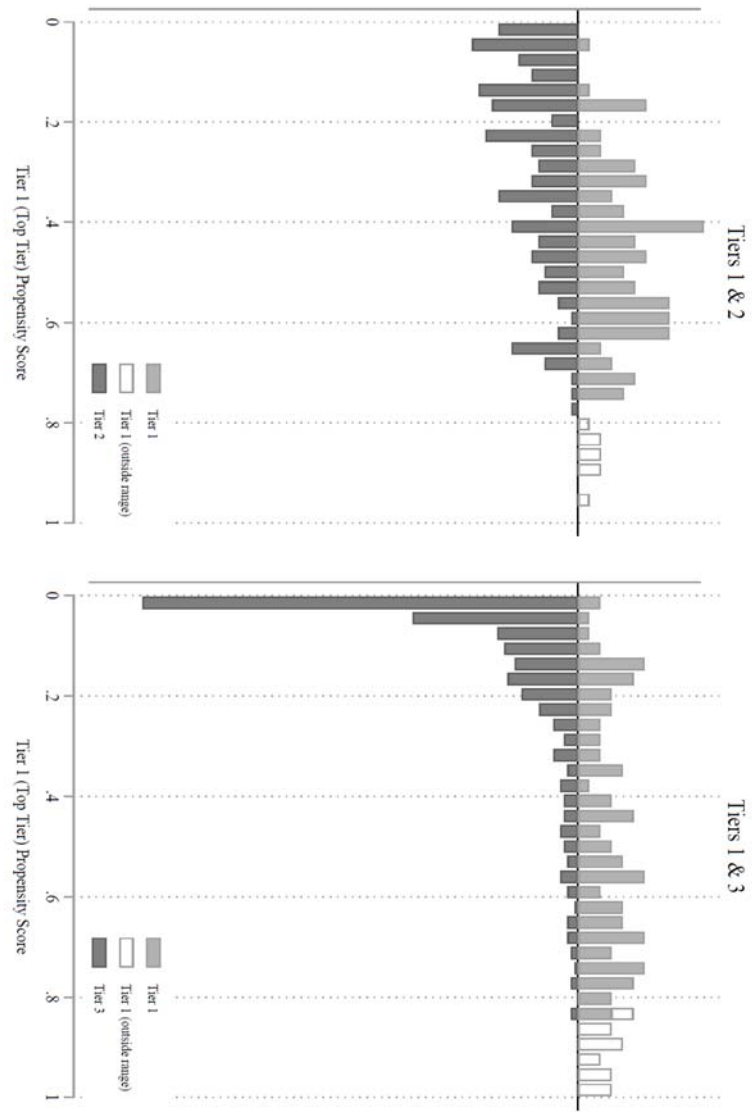
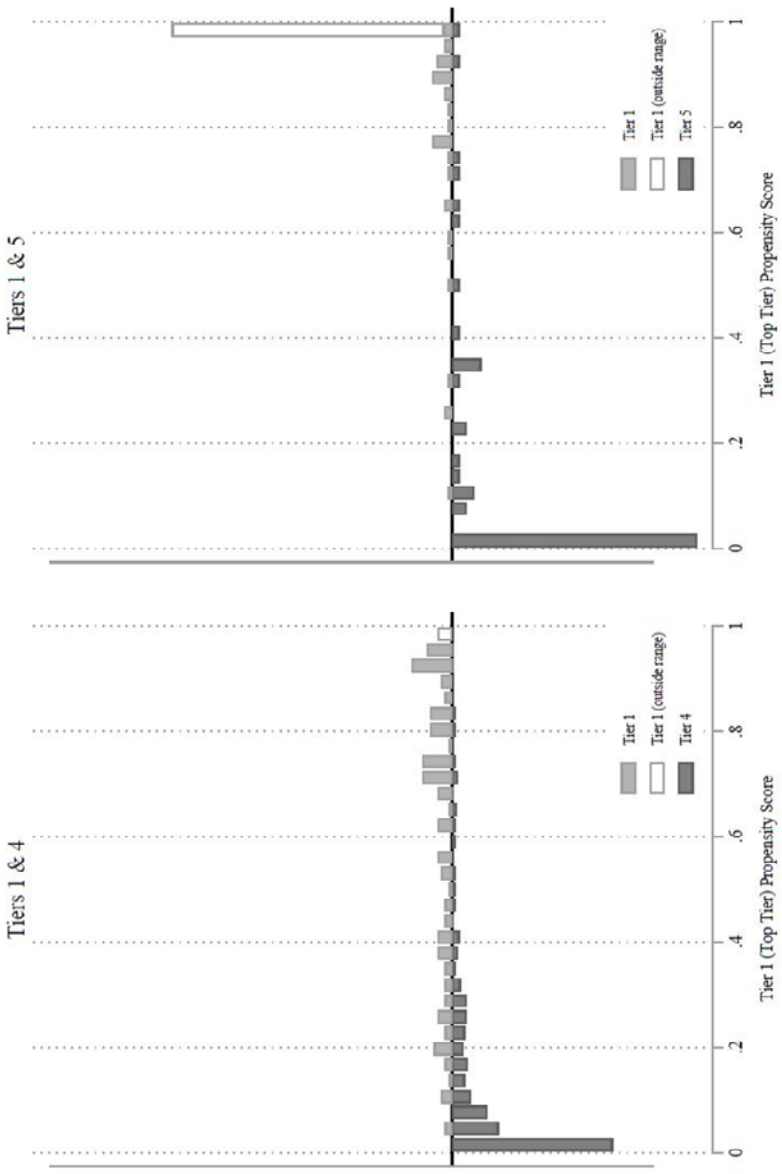


Figure 10: Distribution of the Probabilities of Attending a Law School in the Relevant Tier

Figure 11: Distribution of the Probabilities of Attending a Law School in the Relevant Tier



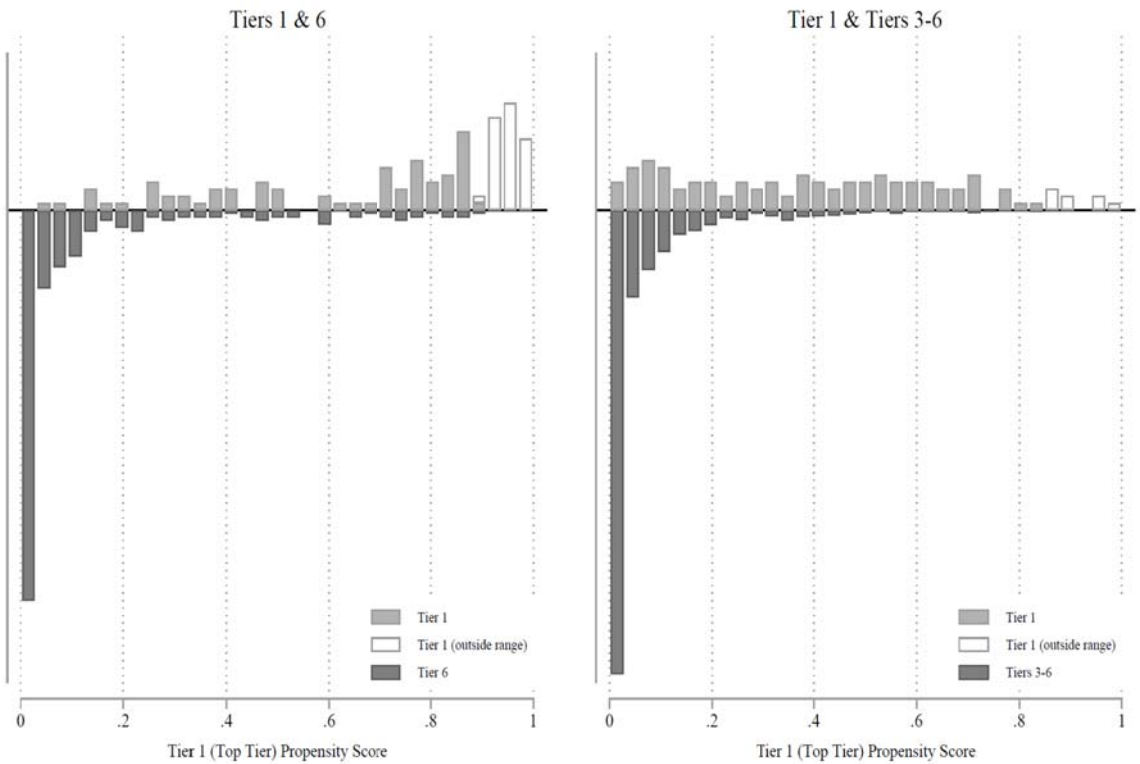


Figure 12: Distribution of the Probabilities of Attending a Law School in the Relevant Tier

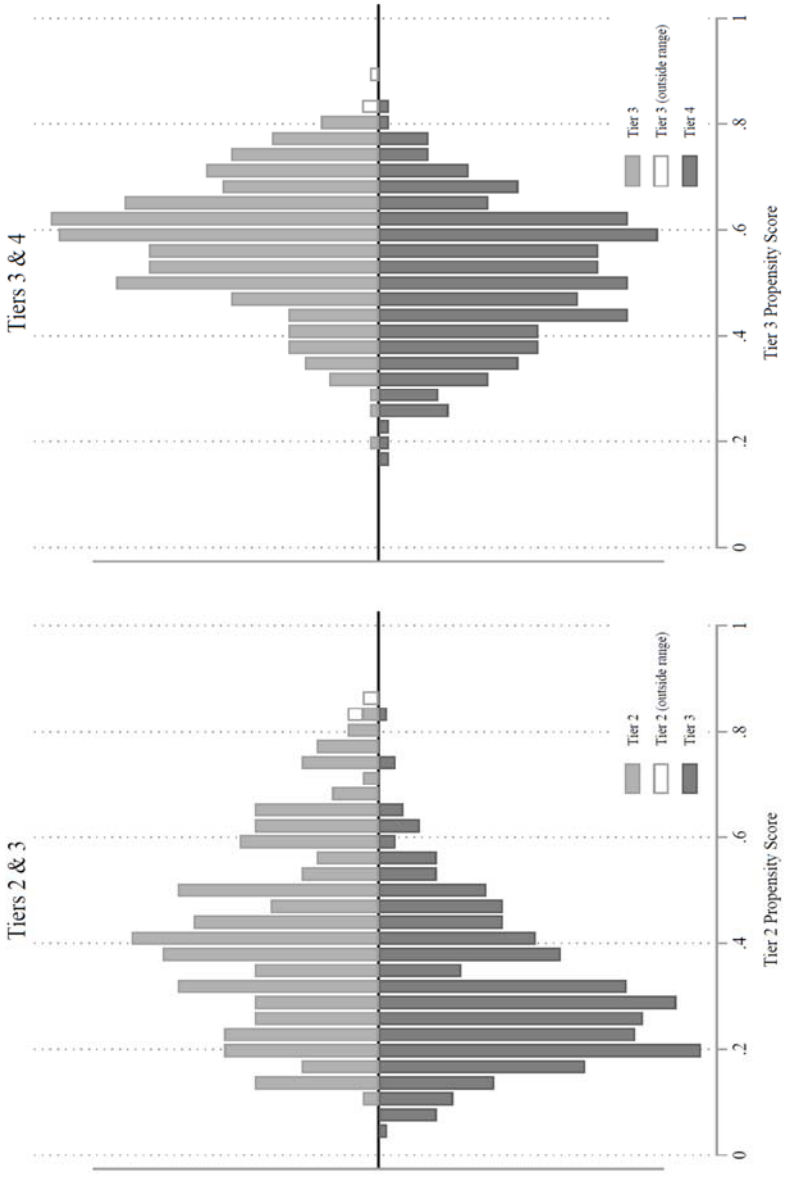


Figure 13: Distribution of the Probabilities of Attending a Law School in the Relevant Tier

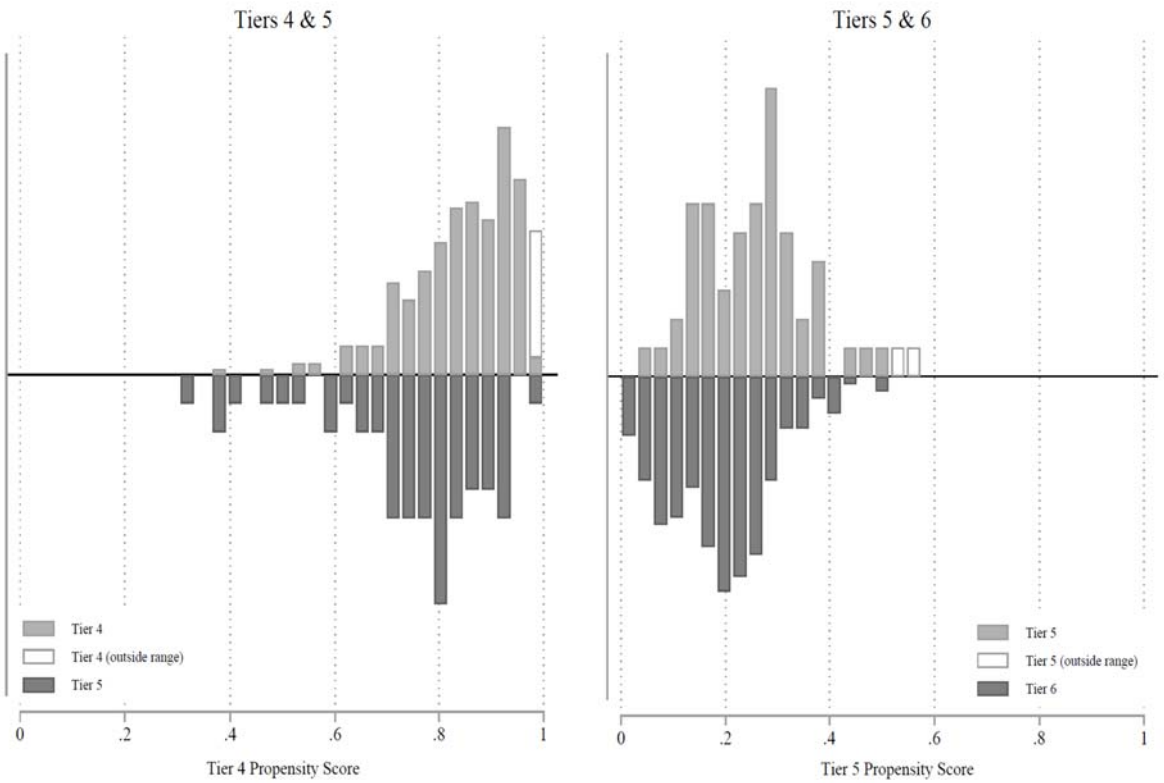


Figure 14: Distribution of the Probabilities of Attending a Law School in the Relevant Tier

Extrapolation bias, among other issues, negatively impacts the peer-reviewed paper by Williams. He purportedly compares black students in the first/second tiers to black students in the fifth/sixth tiers, holding index scores constant, claiming that by eliminating the middle two tiers, he reduces measurement error caused by the fuzzy-boundaries of the tiers.[279] Not only does Williams's choice come at the steep price of external validity (generalizability) because 100 of the 163 law schools examined in the BPS are in those tiers—as are 53 percent of the BPS sample of black law students—but it also increases the likelihood of extrapolation bias because, as noted above, black students in the combined first/second tier have very few comparable black students (or white students) in the combined fifth/sixth tiers. Within-race comparisons across tiers or choice are heavily reliant on parametric assumptions of the model, so more attention must be given to "common support" across a wider range of observed variables in order to have confidence in the results. Camilli and Kevin G. Welner explain that Williams's decision to eliminate the third and fourth tiers "removes from the analysis the most convincing counterfactual students"[280] and "also raises the question of whether this comparison has many real-world (as opposed to modeled) examples. . . . [Further, t]he comparison only to lower-tier law schools also raises a related methodological question: whether the study is comparing applicants so substantially different that it is beyond the capacity of parametric regression models to control for those differences."[281] Many of the statistically significant mismatch effects are only found when Williams removes the middle two tiers of law schools (second tier public and second tier private) from the analysis. Consequently, Williams claims to find support for mismatch theory under the implausible assumption that "the counterfactual to elite law school attendance was that the student would attend a very non-elite school."[282] Students with similar

279. Williams, *supra* note 22, at 174 n.10, recognizes that others using a matching procedure to minimize model dependence have not found support for mismatch or found support for reverse mismatch, *see* Rothstein & Yoon, *supra* note 21; Camilli & Jackson, *supra* note 183, and partially attributes this to the fact that the prior studies analyze a much smaller sample size based on their matching algorithm. The smaller sample is based, in part, on the fact that prior research has matched on a richer set of variables than Williams's research. In other words, Williams admits that his results are based, at least in part, on much coarser comparisons that increase the likelihood of inappropriate comparisons and model dependence. *Cf.* Iacus et al., *supra* note 237, at 23 (acknowledging that properly adjusting for covariate imbalance across treatment and control groups will produce datasets with "fewer observations than we might have otherwise, with the result being less covariate imbalance, less model dependence, and less resulting statistical bias").

280. Camilli & Welner, *supra* note 9, at 517.

281. *Id.*

282. *Id.*

entering credentials who chose to attend law schools with vastly different positions in the law school hierarchy are probably not comparable because they are very likely to differ along important dimensions that are related to graduation and bar passage but that Williams did not take into account in his analysis. Ho also strongly advocated comparing students in adjacent tiers to reduce bias on unobservables.[283]

In addition to the concern over comparing cases that differ along important unobserved dimensions, there is also the issue of the availability of comparable cases along observed dimensions in the data. At minimum, basic disaggregated descriptive statistics across racial groups by tiers and within racial groups across choice should be carefully examined and reported.[284] A careful examination of the number of students included in Williams analysis is illuminating. The median academic index score for black students in the first, second, fifth, and sixth tiers—the tiers that Williams compared—is 542. Whereas 20.3 percent (85 of 419) of black students in the combined first/second tiers scored below the median, 77.7 percent (366 of 432) of black students in the combined fifth/sixth tiers scored below the median. Forty-seven percent (196 of 419) of black students in the combined first/second tiers had an index score above a 646 (75th percentile), compared to 9.2 percent (40 of 432) of black students in the combined fifth/sixth tiers. Black students with index scores at or above 727 (the 90th percentile) composed 22.2 percent (93 of 419) of black students in the first/second tiers and 3.9 percent (17 of 432) of black students in the fifth/sixth tiers. These descriptive statistics make it abundantly clear that Williams's analysis hinges on a very small number of comparisons. Furthermore, the comparisons are especially fragile because of the strong likelihood of unmeasured confounding variables. Sander's recent analysis of mismatch effects with the BPS data, that claims to find strong support for the theory, continues to disregard Ho's and Camilli and Welner's admonition concerning the erroneous comparisons of students from nonadjacent tiers, and thereby reproduces a similar error.[285]

283. Ho, *supra* note 7, at 2002 n.25.

284. King and Lengche Zeng explain that extrapolation bias can exacerbate interpolation bias because using data outside the region of common support induces some degree of model dependence and increases the risk of bias for nearly any model chosen. King & Zeng, *supra* note 156, at 146–51; *see also* Ho, Evaluating Affirmative Action, *supra* note 23.

285. *See* Sander, *Replication of Mismatch*, *supra* note 4, at 87 ("Once again, the differences in outcomes for adjacent tiers are small and not statistically significant. Four of the five 'adjacent' comparisons are positive. But when we compare tiers that are at least three places apart . . . five of the six coefficients are negative—consistent with mismatch—and two of these are at least weakly statistically significant.").

As I previously stated, my discussion of interpolation and extrapolation concludes with a visual example that should help concretize these ideas. The implausibility of Williams's counterfactual comparisons can be shown graphically for his two key explanatory variables, UGPA and LSAT, using a convex hull. A convex hull is a polygon with extreme data points as vertices and represents the smallest convex set that contains the data.[286] A convex set is a collection of points such that, given any two points in the set, the line joining those points lies entirely within that set.[287] In other words, the set is connected such that it is possible to travel between any two points without leaving the set, and the polygon encompassing the set has no edge that is bent inwards. A every intuitive way to describe a convex hull is to imagine the points in a set as tacks on corkboard and placing a rubber band around the perimeter of the points. This rubber band represents the convex hull of all of the points contain therein. The convex hull framework is appealing because it reveals whether model dependence is present without having to examine alternative models.

Figure 15 displays a scatterplot with LSAT scores on the horizontal axis and UGPA on the vertical axis for black law students in the BPS study. The plus signs (+) and the hollow circles (O) represent—according to Williams's analytical framework—students attending, respectively, elite law schools (Tiers 1 and 2) and nonelite law schools (Tiers 5 and 6). The two shaded polygons depict two convex hulls: a lighter shaded hull with +'s at the vertices for the elite tier and a darker shaded hull with O's at the vertices for the nonelite tier. Recall that Williams's counterfactual analysis asks, essentially, how might we expect otherwise similarly situated black students' outcomes to change if the +'s became O's (or, alternatively, if the O's became +'s). Empirically supported counterfactual matches are represented by $\oplus$, indicating that + and O share an identical UGPA and LSAT.

286. King & Zeng, *supra* note 156, at 138. The convex hull is distinct from the bounding box of the data. The bounding box is the lower and upper limits of each of the variables in the model, but the extrapolation can often occur when attempting counterfactual inference within the bounding box because the convex hull is usually smaller than the bounding box. Raphael T. Haftka, *Design of Experiments*, *in* EXPERIMENTAL OPTIMUM ENGINEERING DESIGN: COURSE NOTES 49 (2004), https://mae.ufl.edu/haftka/eoed/Chapter4.pdf [https://perma.cc/3T2T-MKWH].

287. Haftka, *supra* note 286.

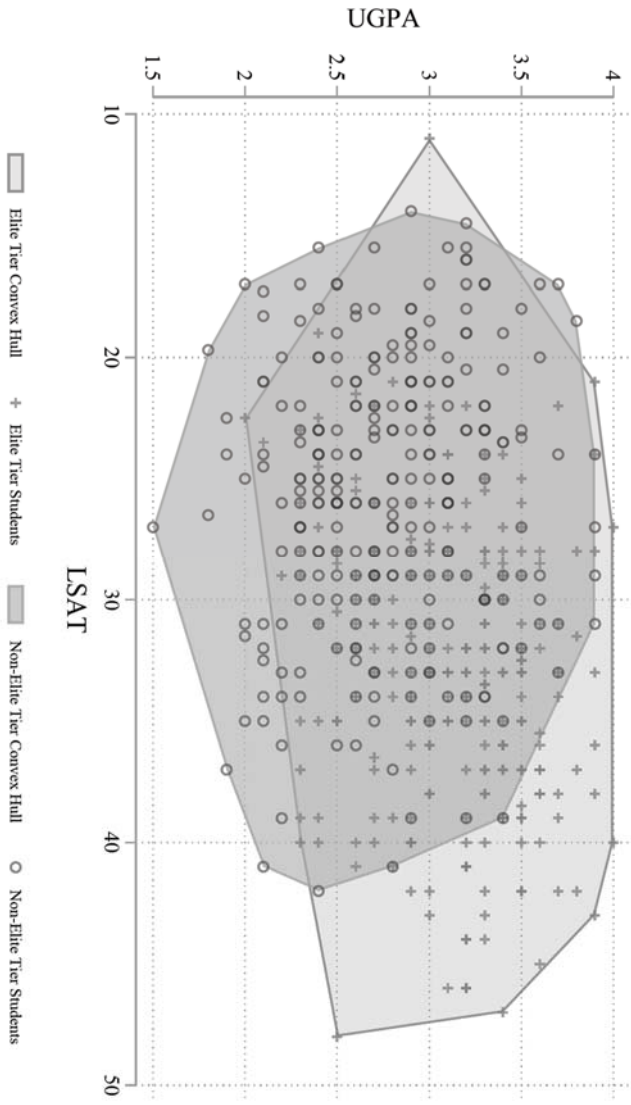


Figure 15: Distribution of Counterfactuals in Williams (2013)

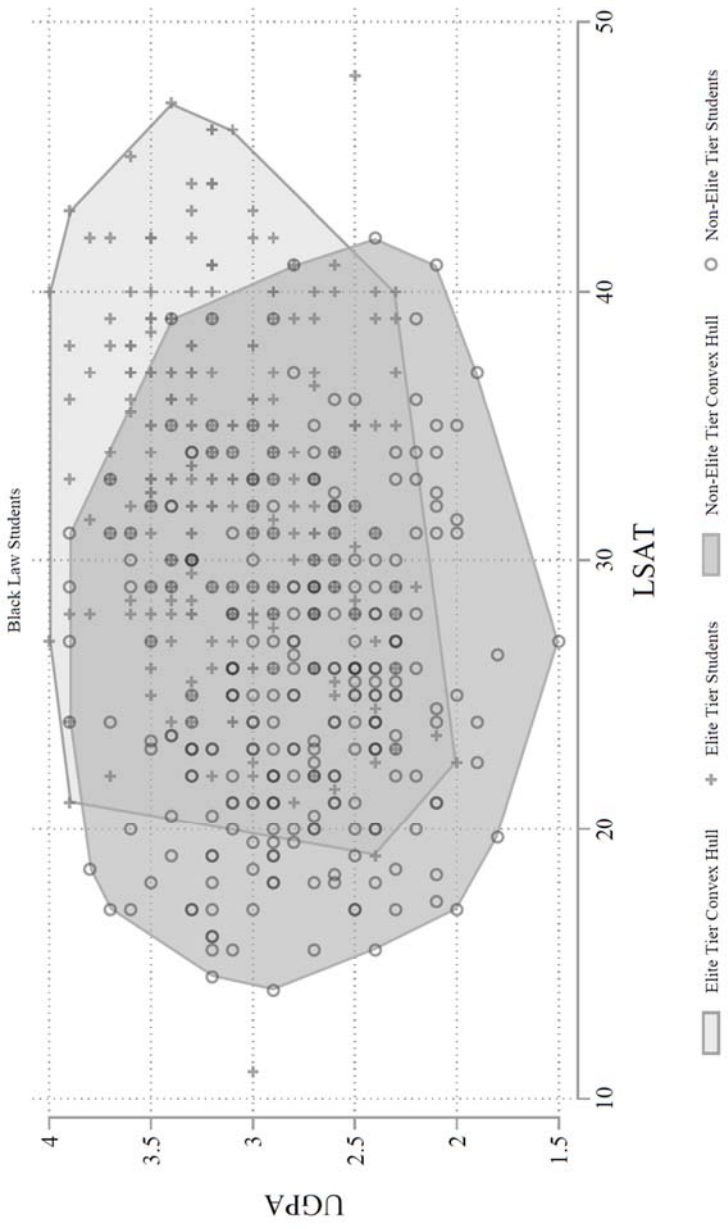


Figure 16: Distribution of Counterfactuals in Williams (2013) [with Outliers Removed]

The shape of a convex hull is sensitive to outliers. It is advisable to remove observations that are very distant from the rest of the observations composing the set used to construct the convex hull. This removal reduces the overall area of the convex hull and helps illuminate the underlying structure of the data.[288] Figure 16 depicts the newly constructed elite tier convex hull when the two extreme outliers in the elite tier—3.0 UGPA/11 LSAT and 2.5 UGPA/48 LSAT—are removed. The shape of the elite tier convex hull changes considerably. Figure 16 also includes the two elite tier observations (+) that now are located outside of the convex hull. It should be obvious that these two observations are very dissimilar to the elite tier students, but it is the 3.0 UGPA/11 LSAT outlier that has the most pronounced impact on the shape of the convex hull. The practical effect of this change is that the minimum LSAT value for the elite tier observations significantly increases, from 11 to 19. The impact on the maximum value of the LSAT is negligible, decreasing from 48 to 47. More importantly, 38 observations in the nonelite tier are now outside of convex hull of the elite tier, and therefore should not be used for counterfactual analysis.

Figures 15 and 16 help highlight the shortcomings of Williams's analysis. A significant portion of the black students in the elite tier are observed in the top right corners of Figures 15 and 16 where students have high UGPA and LSAT scores. It is also apparent that these +'s are entirely outside the convex hull for the nonelite tiers, which is to say that they have no counterfactuals, O, in the data. Any "what if" questions involving these students require extrapolation and are highly dependent on the assumptions of the model. The problem with Williams's analysis does not end there. When exploring the region where the two polygons overlap (intersection of convex hulls), many of the elite tier students who are located within the nonelite tier convex hull (the darker shaded polygon) near its right edge have almost no nearby counterfactuals. In fact, +'s inside the nonelite convex hull with high UGPA or high LSAT scores have extremely few similarly situated O's. The paucity of nearby nonelite counterfactuals, O, requires interpolation to answer "what if" questions. This interpolation is based on comparisons to other students who are quite distant from the counterfactual of interest with respect to UGPA and LSAT. The overwhelming majority of the empirical-based counterfactuals, ⊕, are located near the center of the intersection of the

288. Rossen Atanassov et al., *Algorithms for Optimal Outlier Removal*, 7 J. DISCRETE ALGORITHMS 239, 239 (2009) (noting that outliers can distort statistical results and describing approaches to remove outliers from convex hulls).

convex hulls and, consequently, they are the most appropriate counterfactual comparisons because they require no extrapolation and seemingly modest (if any) interpolation.[289] The number of plausible counterfactuals substantially decreases when additional variables are included in the analysis, but the high-dimensionality of the model precludes a visual representation with convex hulls. In addition to UGPA and LSAT, Williams includes statistical controls for gender, physical/learning impairment, family income, parental education, and whether the student is a native English speaker. When approximate matching is performed on the continuous variables[290] and exact matching is performed on the binary variables,[291] there were—depending on the matching algorithm used for the continuous variables—between zero and six black students in the elite tier with observable counterfactuals in the nonelite tier. The matching methods previously discussed in this Article facilitate the identification of the most appropriate counterfactuals, making causal inferences more credible and disciplined by the data.[292]

289. *Id*. at 137 (describing a metric to assess the fraction of the observed data near the counterfactual).

290. The continuous variables were UGPA, LSAT, family income, and parental education. A coarsened exact matching algorithm was used to maximize the number of comparable observations. *See* Iacus et al., *supra* note 242, at 1 (describing the utility of coarsened matching in causal inference).

291. The binary variables were gender, physical/learning impairment, and native English speaker.

292. Williams purportedly tests the sensitivity of his results with matching: "To assure that these [parametric] results are not driven by a lack of common support, I also estimated a matching model using the matching estimator suggested by Abadie et al. (2001) [sic]. The results from the matching model are consistent with the regression results." Williams, *supra* note 22, at 188 n.24 (citing Alberto Abadie et al., *Implementing Matching Estimators for Average Treatment Effects in Stata*, 4 STATA J. 290 (2004)). Abadie and colleagues describe a nearest neighbor matching algorithm that "find[s] other individuals in the data whose covariates are similar but who were exposed to the other treatment." Abadie et al., *supra*, at 292. But the credibility of Williams's matching model is significantly impacted by (1) the actual level of balance achieved between treatment and control groups for the other explanatory variables in the model; (2) the degree of dissimilarity permitted between matched subjects (called the caliper); and (3) the number of plausible comparisons available in the data. Gary King, Christopher Lucas & Richard A. Nielsen, *The Balance-Sample Size Frontier in Matching Methods for Causal Inference*, 61 AM. J. POL. SCI. 473, 477–79 (2017). The potential bias in the matching estimator will be most pronounced when there is "a relatively small reservoir of potential matches and strong confounder exposure association." Mark Lunt, *Selecting an Appropriate Caliper Can Be Essential for Achieving Good Balance With Propensity Score Matching*, 179 AM. J. EPIDEMIOLOGY 226, 226 (2014) (explaining that "the quality of the matches can be affected by decisions made during the matching process").

Williams does not report any diagnostics of his matching analysis to demonstrate that his results are not driven by extreme covariate imbalance, excessive dissimilarity between matched comparisons, and a very small number of potential matches. *But cf.* King & Nielsen, *supra* note 252, at 450–51 (strongly advising researchers using matching methods to evaluate and document the (in)effectiveness of the chosen matching

F. Measurement Error Bias

The final problem with Sander's analysis is measurement error bias. Simply stated, measurement involves the assignment of numbers to observations in order to quantify phenomena.[293] In the social sciences, many

algorithm). In a prepublication draft of his article, Williams provides more information about his matching analysis, but does not report any of the standard diagnostics that are crucial for evaluating the plausibility of this results. Doug Williams, Does Affirmative Action Create Educational Mismatches in Law Schools? 30–31 & tbl.9 (Apr. 13, 2009) (unpublished manuscript) [hereinafter Williams, Whitepaper] (available at http://public.econ.duke.edu/~hf14/ERID/Williams.pdf [https://perma.cc/TW3Z-8SHP]). I implemented the matching procedure Williams describes in this version of the manuscript, yet I was unable to reproduce several of the results he presents in the manuscript. More importantly, my reexamination of his matching model reveals all of the aforementioned problems that bias causal inference when using matching estimators. First, many covariates in the model remained highly imbalanced after restricting comparison to the region of common support. When better balance is achieved, particularly with respect to UGPA and LSAT, any mismatch effect disappears. Second, when very modest caliper restrictions are imposed to reduce the degree of dissimilarity between matched comparisons, the mismatch effect disappears. It is difficult to determine, *a priori*, what choice of caliper is reasonable. PAUL R. ROSENBAUM, DESIGN OF OBSERVATIONAL STUDIES 168–72 (2010) (discussing caliper restrictions when matching counterfactuals); Lunt, *supra*, at 229–31. Nevertheless, even the slightest restriction imposed in my analysis eliminated any mismatch effect. When I experimented with caliper restrictions typically suggested in the literature, *see* ROSENBAUM, *supra*, at 242–43 (noting calipers of 0.1 to 0.25 are appropriate in most cases), which are much more conservative and result in the reduction of many possible matches, there are no plausible counterfactual comparisons. Caliper matching is, in fact, one form of imposing the common support condition that reduces the bias resulting from comparing cases that are technically within the region of common support, but are too dissimilar to constitute reasonable counterfactuals comparisons. As Austin explains, "nearest neighbor matching within specified caliper widths . . . result[s] in less biased estimates compared with the other matching algorithms." Peter C. Austin, *A Comparison of 12 Algorithms for Matching on the Propensity Score*, 33 STAT. MED. 1057, 1067 (2014).

Perhaps the most troubling aspect of Williams's matching analysis is the extremely limited set of variables he includes to balance the treatment and control groups. As noted, *see supra* note 279, Williams expressly recognizes that prior studies using the BPS data fail to find any mismatch effects, and sometimes report finding a reverse mismatch effect, when adopting a rigorous matching procedure that includes a wider range of relevant variables. Given the centrality of the unconfoundedness assumption for causal inference, Williams's failure to include a richer set of matching variables in light of these prior students ultimately negates any of the causal attributions he attempts to make. Williams's matching analysis is highly suspect even on its own terms. Although he claims that his matching results support his regression results, he uses less variables in the matching model than in the regression analysis, and consequently, the results are not directly comparable.

293. Generally speaking, there are four levels of measurement: *nominal* (mutually exclusive unordered categories, such as state of residence); *ordinal* (ranked categories that do not specify the distance between the ranks, such as dislike, neutral, like); *interval* (ordering that does specify distances between ranks, such as LSAT score); and *ratio* (ordering that

of these phenomena are abstract concepts, and assessments are only valid when they measure what they purport to measure. Measurement error bias occurs when the association between variables is distorted as a result of the process by which the data are measured. The effects of measurement error "can range from the simple attenuation [weakening the true effect of the association] to situations where (a) real effects are hidden; (b) observed data exhibit relationships that are not present in the error-free data; and (c) even the signs ($\pm$) of estimated coefficients are reversed relative to the case with no measurement error."[294] Sander states that measurement error issues in the BPS data would bias the results against finding a mismatch effect.[295] His claim, however, is only true for "simple types of random measurement error with a single explanatory variable."[296] In fact, "the conclusion that the effect of measurement error is to bias the slope estimate in the direction of zero . . . must be qualified, because it depends on the relationship between the [mismeasured predictor], and the true predictor . . . and possibly other variables in the regression model as well."[297]

I forgo a detailed description of biases resulting from measurement error when there are multiple variables in the model because the discussion can become quite complicated by the various measurement error structures, and

both specifies distances between ranks and has a meaningful zero value, such as income). DICTIONARY OF STATISTICS & METHODOLOGY: A NONTECHNICAL GUIDE FOR THE SOCIAL SCIENCES (W. Paul Vogt ed., 3d ed. 2005).

294. RAYMOND J. CARROLL, DAVID RUPPERT, LEONARD A. STEFANSKI & CIPRIAN M. CRAINICEANU, MEASUREMENT ERROR IN NONLINEAR MODELS: A MODERN PERSPECTIVE 41 (2d ed. 2006).

295. Sander, *Systemic Analysis*, *supra* note 7, at 466 ("Measurement error always has the effect of weakening the explanatory power of a variable, since there is more 'noise' in the measure."). Williams repeats this erroneous statement when he claims that "measurement-error bias . . . will further bias the coefficients on distance toward zero [and] measurement-error bias will make it more difficult to discern a mismatch effect if it exists." Williams, *supra* note 22, at 184.

296. Matthew Blackwell, James Honaker & Gary King, *A Unified Approach to Measurement Error and Missing Data: Overview and Applications*, 46 SOCIO. METHODS & RSCH. 303, 304 (2017).

297. CARROLL ET AL., *supra* note 294, at 41 ("[T]he effects of measurement error vary depending on (i) the regression model, be it simple or multiple regression; (ii) whether or not the predictor measured with error is univariate or multivariate; and (iii) the presence of bias in the measurement."); *accord* Kibrom A. Abay, Gashaw T. Abate, Christopher B. Barrett & Tanguy Bernard, *Correlated Non-Classical Measurement Errors, 'Second Best' Policy Inference and the Inverse Size-Productivity Relationship in Agriculture*, 139 J. DEV. ECON. 171, 179 (2019) ("The signs and magnitude of resulting biases in estimates of a key parameter are analytically ambiguous and depend on several parameters characterizing measurement errors in these variables as well as the relationship under investigation.").

different manners in which each contributes to bias.[298] The important takeaway is the assumption that no measurement error bias exists is unlikely to hold for the BPS data because, as numerous scholars have emphasized, the key variables examined by Sander are all inherently error-prone measures.[299] Measurement error bias is of particular relevance in causal inference because using error-prone measurements will either misidentify the treatment and control groups,[300] misjudge the treatment dosage,[301] or fail to balance the treatment and control groups with respect to the confounding variables.[302] In fact, Sander expressly acknowledged the problems that have been identified with using LGPA as an indicator of legal learning, especially as it relates to the predictability of LGPA using LSAT scores, but he downplays the implications of this line of research in assessing the causes and consequences of racial disparities in law school and bar exam performance.[303]

298. *See* Tyler J. VanderWeele & Miguel A. Hernán, *Results on Differential and Dependent Measurement Error of the Exposure and the Outcome Using Signed Directed Acyclic Graphs*, 175 AM. J. EPIDEMIOLOGY 1303, 1305 (2012) (describing the classification of measurement errors and the various methods that can be used to correct for them). For more detailed treatments of measurement error, see generally CARROLL ET AL., *supra* note 294, and JOHN P. BUONACCORSI, MEASUREMENT ERROR: MODELS, METHODS, AND APPLICATIONS (2010).

299. WIGHTMAN & MULLER, *supra* note 240, at 25 (noting that entering credentials overpredict law school performance for minority students when white students are used as the baseline); Kevin R. Johnson & Angela Onwuachi-Willig, *Cry Me a River: The Limits of "A Systemic Analysis of Affirmative Action in American Law Schools"*, 7 AFR.-AM. L. & POL'Y REP. 1, 2 n.6 (2005) ("Sander wholly ignores the criticism of undue reliance on the [LSAT] in the admissions process."); *see also* William D. Henderson, *The LSAT, Law School Exams, and Meritocracy: The Surprising and Undertheorized Role of Test-Taking Speed*, 82 TEX. L. REV. 975, 1044 (2004) (discovering that the heavy reliance on time-pressured law school exams has the effect of increasing the predictive validity of the LSAT and advantages fast-rate test takers over lower-rate test takers, irrespective of their level of reasoning ability); William C. Kidder, *Does the LSAT Mirror or Magnify Racial and Ethnic Differences in Educational Attainment?: A Study of Equally Achieving "Elite" College Students*, 89 CALIF. L. REV. 1055, 1074–79 (2001) (noting that the racial gap in LSAT scores that favors white law applicants relative to all other racial groups remains after taking into account applicant's age, undergraduate institution, college major, UGPA, and grade inflation).

300. *See infra* Subparts II.F.1.a, II.F.1.c.

301. *See infra* Subpart II.F.1.b.

302. *See* GRACE Y. YI, STATISTICAL ANALYSIS WITH MEASUREMENT ERROR OR MISCLASSIFICATION: STRATEGY, METHOD AND APPLICATION 49 (2017) (explaining that measurement error may impact, *inter alia*, parameter estimation, hypothesis testing, prediction, and model selection); VanderWeele & Hernán, *supra* note 298; *infra* Subpart II.F.2.

303. *Compare* Sander, *Systemic Analysis*, *supra* note 7, at 434 n.182 (positing that in-class exams do not explain black-white differentials in law school grades), *with* Henderson, *supra* note 299, at 1029 (noting that black students had the highest differentials, by far, between in-class test performance and other methods at a national law school).

1. Instability Bias

Sander's analysis violates the "stable unit treatment value assumption" (SUTVA) of causal inference. Also referred to as the "stability assumption," "consistency assumption," "no multiple versions of treatment assumption," "no treatment diffusion assumption," and the "no inference between units assumption," SUTVA requires that the treatment is defined unambiguously. The assumption "rel[ies] on external, substantive information" and is necessary so that the potential outcomes for each individual under each possible treatment are well identified and take on a single value.[304] One must be able to identify, with precision, which units actually receive the treatment, as well as the quantity of the treatment received.[305] "Causal inference is generally impossible without such assumptions [about stability], and thus it is critical to be explicit about their content and their justifications."[306] When treatment assignment does not accurately reflect which unit received the treatment or when the treatment dosage is unstable across units, one cannot meaningfully assess the treatment effect. At its core, the SUTVA is about measurement error in treatment assignment and dosage (or, more precisely, the absence thereof). The treatment and control groups must be defined properly so the analyst can ascertain, for example, how a student's academic performance in law school, likelihood of graduation, likelihood of passing the bar examination, and post–law school earning would have been different if that student attended a less selective (and less academically competitive) law school. The problems in Sander's analysis resulting from the violation of the three components of the SUTVA—unambiguous treatment assignment, unambiguous treatment dosage, and no interference between units—are discussed below.

304. IMBENS & RUBIN, *supra* note 214, at 10. Under the potential outcomes framework, each unit has two possible outcomes: (1) the outcome observed if the unit received the treatment, and (2) the outcome observed if the unit did not receive the treatment. Obviously, we can only observe one of these two outcomes. The consistency requirement is necessary so that each unit's potential outcome under the unit's observed treatment history is precisely the unit's observed outcome. The consistency assumption may be difficult to satisfy in observational studies with treatments "for which manipulation is difficult to conceive." Stephen R. Cole & Constantine E. Frangakis, *The Consistency Statement in Causal Inference: A Definition or an Assumption?*, 20 EPIDEMIOLOGY 3, 3 (2009).

305. *See* MORGAN & WINSHIP, *supra* note 58, at 37; Ho & Rubin, *supra* note 218, at 21.

306. IMBENS & RUBIN, *supra* note 214, at 10.

a. Bias From Ambiguous Treatment Assignment

Sander assumes that the different tiers in the data he examines from the BPS are meaningfully distinguishable based on the average UGPA and LSAT scores of students attending those tiers. Consequently, the tier hierarchy should roughly reflect the selectivity of the schools and the academic strength of the students attending those schools in the respective tiers, so holding a respondent's UGPA and LSAT constant, the tier assignment of the respondent's school will capture how much (or how little) that student is "mismatched" relative to their classmates.[307] But this assumption has been proven to be incorrect. Table 2 (below) presents the median values of the seven variables used to create the six law school tiers: tuition, enrollment, selectivity (percent accepted), percent minority, student/faculty ratio, LSAT, and UGPA.[308] The theoretical LSAT range is 10–48, whereas the UGPA range is 1.0–4.0.[309] In the 1990–1991 LSAT test administrations, law school applicants in Tiers 1–6 ranked, respectively, in the 88th, 80th, 71st, 60th, 38th, and 24th percentiles based on the median LSAT scores of the students in each tier.[310] The minimal differences between the second and third tier schools with respect to UGPA (0.05), LSAT (1.88), and school selectivity (0.02) are

307. Sander, *Mismeasuring*, *supra* note 24, at 2009.

308. The information provided in Table 2 was originally presented in WIGHTMAN, BAR PASSAGE STUDY, *supra* note 36, at 9 n.20. *See also id.* ("[C]lassification of schools by type of control (public or private) was considered as a potential clustering variable, but it was not included because it was so highly correlated with tuition."). In the original table, the bottom row ("Number of Schools") is incorrect. This error is obvious when examining Figure 3 and Table 7 in LINDA F. WIGHTMAN, LAW SCH. ADMISSION COUNCIL, RESEARCH REPORT NO. 93-04, CLUSTERING U.S. LAW SCHOOLS USING VARIABLES THAT DESCRIBE SIZE, COST, SELECTIVITY, AND STUDENT BODY CHARACTERISTICS 26 fig.3, 31 tbl.7 (1993) [hereinafter WIGHTMAN, CLUSTERING].

309. The scale used to score the LSAT has changed over time. The original LSAT score of 200–800 remained from 1948 until 1982, but was changed to the 10–48 scale, in part, because of a concern that this scale gave the impression of too much precision. The LSAT was converted to the current scale, 120–180, in June 1991. LAURA A. LAUTH, ANDREA THORNTON SWEENEY, CHRISTIAN FOX & LYNDA M. REESE, LAW SCH. ADMISSION COUNCIL, LSAT TECHNICAL REPORT NO. 14-01, THE PERFORMANCE OF REPEAT TEST TAKERS ON THE LAW SCHOOL ADMISSIONS TEST: 2006–2007 THROUGH 2012–2013 TESTING YEARS 3 (2014) (describing the evolution of the LSAT scale from 1948 to 1991).

310. *See* LINDA F. WIGHTMAN, LAW SCH. ADMISSION COUNCIL, RESEARCH REPORT NO. 94-02, ANALYSIS OF LSAT PERFORMANCE AND PATTERNS OF APPLICATION FOR MALE AND FEMALE LAW SCHOOL APPLICANTS 13 (1994) [hereinafter WIGHTMAN, ANALYSIS OF LSAT PERFORMANCE]. The LSAT score distributions were disaggregated by gender, so the percentiles were calculated using a weighted average of the scores of men and women. The weights were derived from the proportion of men and women applicants. The average LSAT scores for men (33.6) and women (32.78) are nearly identical, so the nonweighted averages are virtually indistinguishable.

illuminating. Attending a third tier school rather than a second tier school results in competing with classmates who are, on average, only 1.5 percent lower with respect to UGPA and only 4.76 percent lower in LSAT (and 8 percentile points).

Table 2: Mean Scores for Each Clustering Variable

Cluster	1	2	3	4	5	6
Tuition	13,659	11,153	3,481	11,428	6,141	3,136
Enrollment	704	1,466	606	797	516	347
Selectivity	0.17	0.26	0.28	0.34	0.5	0.33
Percent minority	0.2	0.19	0.15	0.12	0.08	0.58
Student/faculty ratio	22.04	28.14	21.14	24.73	21.64	17.77
LSAT	42.06	39.53	37.65	35.51	32.29	29.25
UGPA	3.50	3.34	3.29	3.09	3.05	2.86
Percent Private	88	60	4	98	56	29
Number of Schools	18	19	52	53	21	8

Table 3 (below) lists the standardized scores for the clustering variables indicating how far each tier differs, on average, from the overall means of all of the law schools in standard deviation units.[311] The average LSAT scores of students attending schools in the second tier and third tiers diverge by less than a half a standard deviation (0.47), the average UGPA only differs by a quarter of a standard deviation (0.25), and the average selectivity score varies by one-sixth of a standard deviation (0.17).[312] A report released in 1993 by the

311. The scores presented in Table 3 were originally presented in WIGHTMAN, CLUSTERING, *supra* note 308, at 31 tbl.7. The overall means/standard deviations of the seven clustering variables for the 171 American Bar Association (ABA) accredited law schools are: Tuition (8179.16/4808.09); Enrollment (748.39/375.66); Selectivity (0.32/0.11); Percent Minority (0.16/0.12); Student/Faculty Ratio (23.03/4.34); LSAT (36.61/3.98); GPA (3.20/0.23). *See id.* at 5 tbl.2.

312. *See id.* at 32 tbl.8 (Tier 2 is "Cluster 4" and Tier 3 is "Cluster 1"); *see also* Kidder & Lempert, *Mismatch Myth*, *supra* note 22, at 110 (noting that the near identical UGPA and LSAT scores in the second and third tiers is something that has been overlooked by most scholars); *cf.* Brief Amicus Curiae for Richard Lempert in Support of Respondents at 6, Fisher v. Univ. of Tex. at Austin (*Fisher II*), 136 S. Ct. 2198 (2016) (No. 14-981) [hereinafter Lempert Brief] ("Sander's law school 'mismatch' work further suffers because it rests on a mistaken assumption unremarked by Sander and others (myself included) who have analyzed the BPS data. . . . [T]he assumption of a reliable index score hierarchy fails in two important instances. The mean index score of schools in tier 3 is not significantly below the mean of the tier 2 schools, and there is almost no difference between each tier's typical (or centroid) school.").

Law School Admission Counsel (LSAC), which administered the BPS, expressly noted that there were no statistically significant differences between the second and third tiers in terms of UGPA, LSAT, selectivity, and percent minority.[313] There were also no statistically significant differences in UGPA between the fourth and fifth tiers, and the fifth and sixth tiers.[314]

With respect to selectivity (percent admitted), there were no statistically significant differences between the second and sixth tiers, third and sixth tiers, and the fourth and sixth tiers. The only statistically significant differences in selectivity involving comparisons with sixth tier involve the first and fifth tiers. The first tier is significantly more selective than the sixth tier (17 percent vs. 33 percent), while the fifth tier is significantly less selective (50 percent vs. 33 percent). The selectivity differences (or more accurately, lack thereof) between the sixth tier and most of the other tiers merit additional discussion because the sixth tier ranks last in terms of average UGPA and LSAT. The sixth tier consists entirely of HBLS and is the only tier that averages more than 20 percent minority enrollment (58 percent). Schools in the sixth tier have, on average, the lowest tuition, smallest class sizes, and lowest student-faculty ratio.

Table 3: Standardized Scores for Each Clustering Variable

Cluster	1	2	3	4	5	6
Tuition	1.139	0.618	-0.977	0.675	-0.423	-1.048
Enrollment	-0.118	1.912	-0.377	0.131	-0.618	-1.066
Selectivity	-1.302	-0.489	-0.321	0.247	1.687	0.112
Percent minority	0.328	0.255	-0.099	-0.359	-0.660	3.415
Student/faculty ratio	-0.228	1.178	-0.434	0.392	-0.321	-1.213
LSAT	1.370	0.734	0.263	-0.276	-1.087	-1.850
UGPA	1.291	0.609	0.359	-0.500	-0.691	-1.517

Table 4 (below) provides information about the typical law school ("centroid").[315] The second and third tiers are revealing: The two tiers have identical LSAT scores (39; 76th percentile), and the third tier centroid has a higher UGPA (second tier: 3.30; third tier: 3.33) and is more selective (second tier: 0.31; third tier: 0.27) than the second tier.[316] In fact, second tier and third

313. *See* WIGHTMAN, CLUSTERING, *supra* note 308, at 32 tbl.8, 33.
314. *See id.* (Tier 4 is "Cluster 3," Tier 5 is "Cluster 2," and Tier 6 is "Cluster 6").
315. The statistics in Table 4 were calculated from data provided in Wightman's clustering analysis. *Id.* at 5 tbl.2, 35 tbl.9.
316. *Id.* at 35. The selectivity measure captures the percent of applicants admitted to the law school, so a lower admissions rate corresponds to a more selective law school. The LSAT

tier schools were differentiated in the BPS because of reasons other than the academic potential of students—namely, tuition (second tier: $13,314; third tier: $3,332), student/faculty ratio (second tier: 30.02; third tier: 18.36), and class size (second tier: 1334; third tier: 648).

Table 4 (below) provides information about the typical law school ("centroid").[317] The second and third tiers are revealing: The two tiers have identical LSAT scores (39; 76th percentile), and the third tier centroid has a higher UGPA (second tier: 3.30; third tier: 3.33) and is more selective (second tier: 0.31; third tier: 0.27) than the second tier.[318] In fact, second tier and third tier schools were differentiated in the BPS because of reasons other than the academic potential of students—namely, tuition (second tier: $13,314; third tier: $3,332), student/faculty ratio (second tier: 30.02; third tier: 18.36), and class size (second tier: 1334; third tier: 648).

Table 4: Scores for Cluster Centroid for Each Clustering Variable

Cluster	1	2	3	4	5	6
Tuition	15,900	13,314	3,332	10,698	4,255	1,019
Enrollment	606	1,334	648	685	464	319
Selectivity	0.192	0.312	0.272	0.372	0.472	0.362
Percent minority	0.237	0.208	0.131	0.131	0.141	0.519
Student/faculty ratio	22.444	30.021	18.361	26.351	20.169	20.903
LSAT	42	39	39	35	32	28
UGPA	3.50	3.30	3.33	3.00	3.09	2.91

Table 5 (below) lists the standardized scores for the clustering variables indicating how far each centroid differs from the overall means of all of the law schools (in standard deviation units). The LSAT scores of the typical student attending the median school ("centroid") in the second tier and third tiers diverge from the overall mean by an identical amount (0.601), the divergence from the mean UGPA only differs by less than one-seventh of a

for the centroid law school for Tiers 1–6 placed students in, respectively, the 88th, 76th, 76th, 55th, 38th, and 20th percentiles. *See* WIGHTMAN, ANALYSIS OF LSAT PERFORMANCE, *supra* note 310, at 13.

317. The statistics in Table 4 were calculated from data provided in Wightman's clustering analysis. *Id.* at 5 tbl.2, 35 tbl.9.

318. *Id.* at 35. The selectivity measure captures the percent of applicants admitted to the law school, so a lower admissions rate corresponds to a more selective law school. The LSAT for the centroid law school for Tiers 1–6 placed students in, respectively, the 88th, 76th, 76th, 55th, 38th, and 20th percentiles. *See* WIGHTMAN, ANALYSIS OF LSAT PERFORMANCE, *supra* note 310, at 13.

standard deviation (0.13), between the second and third tier schools, and the divergence average selectivity score varies by approximately two-fifths of a standard deviation (0.4).

Table 5: Standardized Scores for Cluster Centroid for Each Clustering Variable

Cluster	1	2	3	4	5	6
Tuition	1.606	1.068	-1.008	0.524	-0.816	-1.489
Enrollment	-0.379	1.561	-0.267	-0.168	-0.757	-1.143
Selectivity	-1.156	-0.068	-0.431	0.474	1.380	0.384
Percent minority	0.642	0.400	-0.247	-0.247	-0.166	2.990
Student/faculty ratio	-0.135	1.611	-1.076	0.765	-0.659	-0.490
LSAT	1.356	0.601	0.601	-0.404	-1.159	-2.165
UGPA	1.308	0.425	0.558	-0.855	-0.457	-1.252

These very modest standardized differences indicate that not all of the clusters (or their centroids) are statistically distinct on the variables used to capture mismatch.[319] The most plausible explanation is that the differential performance of students in the second and third tiers with respect to law school grades, graduation, and bar passage (holding the students' UGPA and LSAT constant) is attributable to lower tuition costs, smaller enrollments, far better student-faculty ratios, or some other factors unrelated to mismatch. Sander, himself, notes that the average academic index score (a combination of UGPA and LSAT) is higher for black students in the sixth tier (524.16) compared to those in the fifth tier (501.03), and this is especially noteworthy because, as I previously noted, the sixth tier includes only HBLS. Black students at HBLS obtain better outcomes with respect to graduation and bar passage than their tier ranking would predict, and this is likely attributable to the fact that HBLS have far more black students than the number needed to constitute a critical mass, are very low cost (except for Howard University), offer plenty of same-race role models, and their graduates tend to take the bar in states with relatively lower bar exam passing standards.[320]

319. WIGHTMAN, CLUSTERING, *supra* note 308, at 32 tbl.8, 33.

320. *See* Kidder & Lempert, *Mismatch Myth*, *supra* note 22, at 111. Law students attending schools housed at historically black colleges and universities (HBCUs) are substantially more likely to have black professors than law students attending predominately white law schools. At primarily white institutions, approximately 5 percent of all tenure-track faculty are black, compared to 12 percent of students. THOMAS D. SNYDER, CRISTOBAL DE BREY & SALLY A. DILLOW, U.S. DEP'T OF EDUC., DIGEST OF EDUCATION STATISTICS 2016, at 515 tbl.315.20 (2018). Twenty percent of all black, tenure-track college and university

faculty are employed by historically black four-year colleges and universities (HBCUs), yet these institutions account for just 1.7 percent of all tenure-track faculty nationwide. Matt Krupnick, *After Colleges Promised to Increase It, Hiring of Black Faculty Declined*, HECHINGER REP. (Oct. 2, 2018), https://hechingerreport.org/after-colleges-promised-to-increase-it-hiring-of-black-faculty-declined [https://perma.cc/6J84-KE9L]. In 2016, there were 102 HBCUs located in nineteen states, the District of Columbia, and the U.S. Virgin Islands. SNYDER ET AL., *supra*, at 503. Moreover, from 2007–2016, the percent of annual black fulltime faculty hires at predominately white four-year public and nonprofit colleges and universities dropped from 7 percent to 6.6 percent, nationally. Krupnick, *supra*. According to the U.S. Department of Education, black professors composed 61 percent of the fulltime faculty at Howard University and 58 percent at the University of the District of Columbia (both historically black colleges), but 4 percent at Georgetown University, 5 percent at American University, and 6 percent at George Washington University (all of these institutions have law schools and are located in the District of Columbia). *Integrated Postsecondary Education Data System*, NAT'L CTR. FOR EDUC. STAT., https://nces.ed.gov/ipeds/ [https://perma.cc/A9ZB-22GM] (last visited Sept. 12, 2020). North Carolina Central University (an HBCU), Duke University, and University of North Carolina–Chapel Hill all have law schools and are in the Raleigh-Durham area, and their percentage of black faculty are, respectively, 57 percent, 5 percent, and 6 percent. *Id.* In Houston, Texas Southern University (an HBCU), which houses Thurgood Marshall School of law, has 68 percent black faculty members, whereas the University of Houston, which also houses a law school, has 4 percent black faculty members. *Id.* Similarly, 66 percent of faculty were black at Florida A&M University (an HBCU), whereas only 4 percent at Florida State University. Both universities are located in Tallahassee, Florida and have law schools. *Id.* Louisiana State University and Southern University (an HBCU), both with law schools and located in Baton Rouge, significantly differ in their percentage of black faculty: 4 percent and 70 percent, respectively. *Id.* California does not have any historically black colleges, but the percentage of fulltime black faculty members at its premier universities with law schools are: 3 percent at University of California, Berkeley (UC Berkeley), 3 percent at University of California, Los Angeles (UCLA), 3 percent at University of California, Davis (UC Davis), 3 percent at University of California, Irvine (UCI), 3 percent at University of Southern California (USC), and 2 percent at Stanford University (Stanford). *Id.*

Ho finds evidence that racial context matters for white students, as well. Specifically, white students, on average, have a 10 percent increase in bar passage probability if they attend a fifth tier school rather than a sixth tier school—a finding inconsistent with Sander's mismatch hypothesis. This might result from the fact that the sixth tier consists entirely of HBLS, and white students actually do better in environments that are predominately white. *See* Ho, *supra* note 7, at 2004.

Lempert makes a related point when comparing the effect of college selectiveness on minority graduation rates for science majors. Minorities at UCLA with SAT scores in the next to bottom quartile (25–50th percentile) are nearly twice as likely to graduate within five years with science degrees than students at UC Berkeley, although the schools are equally selective. Richard Lempert, *Mismatch and Science Desistance: Failed Arguments Against Affirmative Action*, 64 UCLA L. REV. DISC. 136, 159 tbl.1, 160–62 (2016). Lempert asks:

> What then can explain the difference between the way second quartile intended science majors perform at the two schools? The answer is likely to lie in the way the sciences are taught at the two schools and/or in the cultures of the two schools and the comfort that minorities feel on campus. Sheer numbers may also play a role. UCLA has considerably more minorities who, when they enter, intend to complete a science major (973

The inability of the tier location variable to accurately capture differences in law school competitiveness/selectivity is especially important for analyses of the BPS data that combine different tiers and compares them to others. Sander places significant emphasis on Williams's peer-reviewed article that purportedly finds support for the mismatch hypothesis.[321] But on closer inspection, Williams's inattention to the fact that the second tier and third tier

> compared to 684). Hence UCLA's introductory science classes are likely to contain larger numbers of minority students. The relevance is that if large persistence to degree differences can arise between similar schools like UCLA and Berkeley that do not differ substantially in the degree to which the academic credentials of their science-interested minorities fall short of their schools' white intended science majors, then explanations other than so-called science mismatch, may explain any advantage in graduating scientists that minimally selective schools have over highly selective ones.

Lempert, *supra*, at 160–62. *See also* Valerie K. Bostwick & Bruce A. Weinberg, *Nevertheless She Persisted? Gender Peer Effects in Doctoral STEM Programs* 9 (Nat'l Bureau of Econ. Rsch., Working Paper No. 25028, 2018) (discovering that the number of women in science, technology, engineering, and mathematics doctoral programs significantly impacted the probability of women graduating within six years).

Medical school graduates are also instructive. An examination of the impact of affirmative action on outcomes of 1784 students admitted to the UC Davis Medical School over a 20 year period (1968–1987) revealed that although students admitted via the school's affirmative action program (53.5 percent racial minorities) earned lower grades and were more likely to repeat the National Board of Medical Examiners examination than non–affirmative action admits, there were no differences between the two groups with respect to the likelihood of graduation, completion of medical residency, performance assessments by medical residency directors, and selection of primary care disciplines. Davidson & Lewis, *supra* note 196. The authors of the study concluded that race-based affirmative action yields powerful effects on the diversity of the medical student population and shows no evidence of diluting the quality of graduates. *Id.* at 1153–58. A study of a random sample of nearly three thousand physicians discovered that racial minorities are significantly more likely to express a strong interest in treating underserved populations before attending medical school and ultimately provide care to those underserved communities. Howard K. Rabinowitz, James J. Diamond, J. Jon Veloski & Julie A. Gayle, *The Impact of Multiple Predictors on Generalist Physicians' Care of Underserved Populations*, 90 AM. J. PUB. HEALTH 1225 (2000).

321. Williams, *supra* note 22. Sander has claimed that "none of the critics who argue that mismatch does not exist have successfully published such criticisms in a peer-reviewed journal (much less one with the stature of JELS)." Sander, Brief in *Fisher II*, *supra* note 40, at 21. This is an extremely selective presentation of the facts. As Lempert has explained:

> Sander seeks in his brief [in *Fisher II*] to bolster Williams' findings by noting that the study was published in JELS, a peer reviewed journal, without later rebuttal. In fact, I and a coauthor sought to reply but were told by the JELS, editor that he did not publish replies that simply documented flaws in published articles. Similarly misleading is Sander's attempt to deflate studies criticizing his work by noting they were not published in peer reviewed journals. Sander himself avoided peer review by publishing in law reviews, which established the forums for conversation.

Lempert Brief, *supra* note 312, at 9–10 n.5.

schools are indistinguishable on mismatch-relevant characteristics seriously biases his analysis.[322] Specifically, he combines first/second tier students, but ignores third tier students, thereby assuming that second tier students are distinguishable from third tier students. Although the problem of measurement error with respect to tier is present in nearly all studies of mismatch, Williams's analysis exacerbates the problem because of the way he aggregates the tiers. Table 2 and Table 3 reveal that first and second tier students are more dissimilar than second and third tier students. The more appropriate analysis would be to compare first tier students with the combined second/third tier students. Lempert explains that "[t]hese overlooked considerations affect all studies that use the BPS tier order as a proxy for academic selectivity, but no study's conclusions are rendered more suspect than the study by Williams on which Sander heavily relies."[323] Furthermore, the decision to combine the fifth and sixth tiers should have appeared obviously erroneous from the outset because sixth tier schools are HBLS and unhelpful in evaluating mismatch for the reasons I have previously discussed (including critical mass of black students, larger proportion of black faculty, and much cheaper tuition).[324] Seventy-five percent of the black students in the combined fifth/sixth tiers attend a HBLS, which is not surprising because "percent minority" is a factor used to construct the tiers. Williams combines tiers that are most dissimilar (first/second and fifth/sixth) and eliminates the large middle-range (third/fourth) to find support for mismatch, but "Williams never alerts his readers to the unusual locations of black students in the bottom two tiers or the plausible rival hypothesis that contaminates his results."[325]

322. *See* Kidder & Lempert, *Mismatch Myth*, *supra* note 22, at 109–11 (discussing biases in Williams's analyses resulting from the improper comparison of tiers and omitted variables). Williams's analysis suffers from other conceptual and methodological shortcomings as well. These include sample selection bias, nonresponse bias, interpolation bias, and extrapolation bias. *See* Rothstein & Yoon, *supra* note 21. In fact, two scholars have noted that some of the effect sizes reported by Williams "are so large as to not be plausible." Arcidiacono & Lovenheim, *supra* note 7, at 20 n.30.

323. Lempert Brief, *supra* note 312, at 8 ("It is hard to imagine a study better designed to take account of the idiosyncrasies of the BPS tier structure in order to find mismatch [than Williams' study].").

324. Kidder & Lempert, *Mismatch Myth*, *supra* note 22, at 111 (explaining why comparisons to students enrolled in HBLS "are of little if any use in evaluating the mismatch hypothesis"); *see also* LINDA F. WIGHTMAN, USER'S GUIDE: LSAC NATIONAL LONGITUDINAL DATA FILE 22 (1999) [hereinafter WIGHTMAN, USER'S GUIDE] (noting that a distinguishing feature of law schools in Tier 6 (HBLS) is their high proportion of minority students and comparing students from the Tier 6 with other tiers "may provide some insight into the benefits and impact for students of color when the law school environment provides a critical mass on non-majority students").

325. Lempert, *supra* note 67, at 19–20 n.43.

Sander has highlighted the fact that law school tier is mismeasured when criticizing the work of others, but refuses to apply similar scrutiny to, and exercise requisite caution in, his own analyses.[326] For example, in response to criticisms of his work by Ho, Sander writes:

> [Ho] assumes that the "tier" variable in the BPS data set is a perfect hierarchical measure of school prestige. That is, he behaves as though all Tier One schools are more elite than all Tier Two schools, all Tier Two schools are more elite than all Tier Three schools, and so on. This is false. As the BPS data manual notes, law schools were "clustered" into six broad groups based on seven characteristics: size, cost, selectivity, faculty-student ratio, percentage minority, median LSAT score, and median undergraduate GPA. Several of these characteristics correlate strongly with prestige, and the clustered tiers as a whole are a reasonable proxy for prestige, if one takes their limitations into account.[327]

Clearly Sander recognizes the limitations of the tier variable, but he still underemphasizes the problem with the variable because one cannot even say "as a whole" that the tiers adequately capture selectivity. As explained previously, not only are the average UGPA, LSAT, and selectivity variables for the second and third tiers statistically indistinguishable, but the median law schools (the centroid) in the second and third tiers have an identical average student LSAT score, and the third tier median school has a higher UGPA and selectivity score than median school in the second tier.

---

326. *See* Ho, *Reply to Sander*, *supra* note 23, at 2014 ("Sander claims that law school tier is mismeasured and hence that the results of my analysis do not prove much. Yet if law school tier, the key causal variable, is mismeasured, [Sander's] original analysis fails as well. Sander cannot have it both ways . . . . " (footnote omitted)).

    The California Court of Appeals made a similar observation when denying Sander's request for confidential information from California's bar admissions database, highlighting Sander's peculiar claim that trial court was obligated to independently formulate a viable plan that would allow the State Bar to provide some, but not all, of the requested data while still protecting bar applicants' privacy interests: "It seems odd for [Sander] to expect the trial court to succeed where [his] own experts in this highly technical field did not." Sander v. State Bar of Cal., 237 Cal. Rptr. 3d 276, 291–92 (Cal. Ct. App. 2018).

327. Sander, *Mismeasuring*, *supra* note 24, at 2009 (emphasis omitted) (footnote omitted). Interestingly, Williams, claimed the BPS made available an "unprecedented amount and quality of micro data," Williams, *supra* note 22, at 173, but then justified eliminating the middle two tiers (third and fourth tiers)—discarding more than three-fifths of the data—on the grounds that the tier location was too coarse a measure of school selectivity (and student quality) to reasonably compare students from adjacent tiers. *Id.* Williams was only able to find any support for the mismatch hypothesis by improperly discarding the middle two tiers, combining dissimilar tiers, and using inappropriate counterfactual students across the combined first/second tier and combined fifth/sixth tier. *Id.*

b. Bias From Ambiguous Treatment Dosage

The failure of Sander's analysis to satisfy the SUTVA, also known as the consistency assumption, is also evident in his analysis of the effect of first choice law school attendance on tier location and bar passage rates. Recall that the SUTVA requires accurate measurement of both exposure to the treatment and the "dosage" of the treatment. In "System Analysis," Sander opines that an ideal way to test the mismatch thesis would be to compare black students with similar UGPA and LSAT credentials admitted to multiple schools: one of the black students would attend the most elite school, while the other black student would attend a significantly less elite school.[328] This type of test, according to Sander, would likely identify black students who were very similar with respect to unobservable characteristics related to law school and bar performance. Sander lamented that the test would be infeasible because few black students pass up the opportunity to attend more elite schools.[329] Ayres and Brooks recognized that BPS data contained information on students who were admitted to their first choice law school but decided to attend a different law school. The BPS also asked students whether their decision to not attend their first choice school was because: (1) they were "not admitted"; (2) "it was too expensive given the financial aid made available to me"; or (3) "it was too distant from my family or for personal responsibilities or attachments."[330]

Ayres and Brooks's analysis "[a]ssum[es] that students on average rank more competitive schools higher among their choices, [and] if mismatch exists [they] ought to observe that relative to those students who attended their first choice schools, those who were admitted but attended their second- or lower-choice law school should perform better."[331] Ayres and Brooks note that their framework "provide[d] a *limited* test of the mismatch hypothesis,"[332] but they were also candid about the shortcomings of the approach. Specifically, Ayres and Brooks acknowledge two key limitations. The first limitation is that the BPS data contain information about the tier of the law school each respondent attended, but not about the tier location of respondent's preferred choice(s)

328. Sander, *Systemic Analysis*, *supra* note 7, at 453.
329. *Id.*
330. It is important to note that the financial aid question does not capture whether the second (or lower) choice school offered sufficient merit aid to attract the student from their first choice school (and the amount of the award required to do so). *See* Ayres & Brooks, *supra* note 12, at 1831–32 n.48.
331. *Id.* at 1831–32 (footnote omitted).
332. *Id.* at 1831 (emphasis added).

in the event that the respondent elected to attend her/his less preferred option.[333] So if a respondent chose their second choice, there is no data identifying the respondent's first choice. Similarly, if the respondent chose their third choice, there is no data identifying the respondent's first or second choice. This is a potentially fatal flaw because even assuming, *arguendo*, that the tier location variable meaningfully distinguished schools in terms of selectivity (indicating no ambiguous treatment assignment bias),[334] the choice analysis is incapable of determining which respondents elected to attend less selective schools, as opposed to equally (or more) selective schools. The inability of the BPS data to provide meaningful information concerning potential mismatch effects is exacerbated by another limitation that I discuss next.

The second limitation pertains to the manner in which respondents in the BPS view their first choice school. Respondents interpreted their first/second/third choice school in drastically different ways. Some students "interpreted 'first choice' to be first choice on the universal set of law schools, while others apparently had in mind first choice among the schools to which they *applied*, and still others interpreted it as first choice among schools to which they were *admitted*."[335] This variation in interpretation was evident from the fact that some students said their current school was their: (a) second or third choice, but they only applied to one school; (b) third or lower choice, but they applied to only two schools; (c) second, third, or lower choice school, but they were only admitted to one school; and (d) third or lower choice school, but they were admitted to only two schools. Consequently, a student may have not even applied, let alone have been admitted, to their first or second choice school. Ayres and Brooks's analysis adjusts for the number of applications a respondent submitted, but this does not specifically address the problem concerning how respondents interpreted their first or second choice school because even among respondents who submitted the same number of law school applications, it is unclear whether they applied or were admitted to their first or second choice school.

This limitation concerning how respondents interpreted their first, second, third, or lower choice school was so significant to Ayres and Brooks that they noted, "We go forward with this analysis nonetheless because it illustrates the type of design needed to assess academic mismatch and the generally poor quality of the extant data in this regard."[336]

333. *Id.* at 1833.
334. *See supra* Subpart II.F.1.a.
335. Ayres & Brooks, *supra* note 12, at 1832 n.50 (emphasis added).
336. *Id.*

Ayres and Brooks's admonition stands in stark contrast to Sander and Taylor's assessment of the choice analysis. According to Sander and Taylor, "the first choice/second choice model provides the soundest direct empirical test of mismatch with available data, and strongly confirms the [mismatch] hypothesis."[337] Similar to Ayres and Brooks, Sander assumes that individuals who elect to attend their second or third choice school, but were admitted to their first choice school, have enrolled in a less selective (and therefore, less academically competitive) school.[338] In contrast to Ayres and Brooks, he disaggregates non–first choice attendees into second choice and third choice, thereby implying that the choice hierarchy maps on to the selectivity hierarchy such that a second choice school is less selective than a first choice school, and a third choice school is less selective than a second choice school. In other words, he assumes the treatment dosage (mismatch) corresponds to the choice hierarchy with even greater granularity. The unsubstantiated choice hierarchy assumption is critical for an analysis of mismatch, and as noted earlier, Ayres and Brooks are much more candid than Sander about the potential deficiencies of this approach when assessing mismatch effects.

There is strong reason to believe, however, that the choice hierarchy assumption is unwarranted. It is very likely that many individuals' second (or third) choice schools are at least equally selective as their first choice (or second) schools because students often are admitted to similar "batches" of schools based on their entering credentials. So, for example, many of the students admitted to University of California, Los Angeles (UCLA) are also admitted to University of California, Berkeley (UC Berkeley), Georgetown University, Northwestern University, University of Texas, University of Southern California (USC), Vanderbilt University, and Washington University–St. Louis. For the most part, these schools are fungible with respect to average UGPA, LSAT, and

337. SANDER & TAYLOR, *supra* note 4, at 86.

338. Sander states:

> It is plausible that, for the most part, a student's "first choice" school is the most elite school to which he applied, since most law students believe that going to a more elite school is the best way to land a good job. That means the "second choice" school is probably less elite—and has lower student median credentials—and the "third choice" school is less elite still. If an African-American student would, at his first-choice school, have much lower-than-average credentials, and thus a large "mismatch" gap, the average gap should be smaller at the second-choice school and smaller still at the third choice-school.
>
> *If these assumptions are true, then these data can help us solve multiple problems in examining mismatch.*"

Sander, *Replication of Mismatch*, *supra* note 4, at 77 (emphasis added).

selectivity, so it is implausible to assume that attending one of these schools as a second (or third choice), as opposed to attending another first choice school, decreases mismatch.

To make ideas concrete, consider a common scenario of a student deciding between three schools to which they have been admitted: UC Berkeley, UCLA, and USC. Based on *U.S. News and World Report* in 2017, the respective rankings of these schools are eighth, seventeenth, and nineteenth, yet their respective median UGPA/LSAT numbers were nearly indistinguishable: 3.78/166 (UC Berkeley), 3.74/166 (UCLA), and 3.76/166 (USC).[339] Assuming, as Sander does, that students' first choice schools are higher ranked than their second choice school, and their second choice school is higher ranked than their third choice school, then UC Berkeley will be preferred over UCLA and USC, and UCLA would be preferred over USC. It should be obvious, however, that the "quality" of the typical students at each of these schools based on their UGPA and LSAT are nearly identical and the fact that a student at USC has a higher probability of passing the bar than a similarly situated student at UCLA, and an even higher probability of passing the bar than a student at UC Berkeley, tells us nothing about mismatch effects. Even if we assume that a student was also accepted to Stanford University (Stanford), ranked second by *U.S. News* in 2017, the median UGPA/LSAT at Stanford was 3.89/171. The difference in median UGPA between Stanford and UC Berkeley, UCLA, and USC is negligible, and the LSAT difference is six points. The six-point LSAT difference may sound significant, but on closer inspection, students with 171 and 166 LSAT scores are rather close in the overall LSAT distribution. From the LSAT administrations between June 2016 and February 2017, a 171 placed a student in the 98th percentile and a 166 placed a student in the 93rd percentile.[340] It is implausible that such a small difference would impact the likelihood of graduating and passing the bar on the first attempt.[341]

339. The acceptance rates for UC Berkeley, UCLA, and USC were, respectively, 21.1 percent, 29.7 percent, and 29.9 percent. Sarah Garvey, *2017 U.S. News & World Report Law School Rankings: Highs, Lows and Specialties*, LAW CROSSING, https://www.lawcrossing.com/ article/900046480/ 2017-U-S-News-World-Report-Law-School-Rankings-Highs-Lows-and-Specialties/ [https:// perma.cc/B9R9-6XQ5] (last visited Sept. 13, 2020).

340. Memorandum from Lisa Anthony, Senior Research Assoc., Law Sch. Admissions Council, to LSAT Score Recipients, June 2014–February 2017 LSAT Score Distributions (June 20, 2017), https://www.lsac.org/sites/default/files/legacy/docs/default-source/ data-%28lsac-resources%29-docs/lsat-score-distribution.pdf [https://perma.cc/GL4Q-QJYT].

341. While a difference between a 171 and 166 may appreciably impact where the student is admitted to law school, it has a very minor impact on law school grades and no meaningful impact on graduation or first-time bar passage.

Sander acknowledges that the Ayres and Brooks article does not attempt to empirically validate the assumption concerning first choice schools and he purports to test the assumption.[342] He presents findings showing that, relative to their first choice, students attending their second or third choice schools are located in lower tiers.[343] According to Sander:

> The estimates suggest that for blacks, attending a second-choice school means going to a school that is perhaps a quarter to a third of a tier "lower" in the hierarchy than one's first-choice school. Going to a third-choice school means going to a school that is perhaps half a tier lower. For whites, there is no apparent difference between "second" and "third" choicers, but clearly a difference between those two groups and the first choicers.[344]

Ayres and Brooks conduct a different analysis (and potentially more useful one given the limitations of the BPS data) that explores why students enroll in schools other than their first choice. The outcome variable in the Ayres and Brooks analysis is not the tier location of the student, given they elected to attend their second (or third) choice school; rather it is a binary variable indicating whether the respondent attended their second (or third) school. Ayres and Brooks explain:

> We attempted to identify the factors that might influence an applicant's decision to forgo their first choice by running race-specific regressions using LSAT, UGPA, family income, status as male, law school tier, and the number of schools to which the student applied and was admitted. This analysis reveals that while almost all of these variables were insignificant for blacks, most were highly significant for whites. For example, white males and whites with more family income were significantly less likely to pass up their first choice. Compared to blacks, whites are also less likely to turn down their first choice when it is in a higher tier. . . . *We remain*

342. "So far as I know, Ayres and Brooks did not try to empirically validate these inferences, but I have attempted to do so, within the constraints of the BPS data." Sander, Whitepaper, *supra* note 4, at 7.

343. Sander, *Replication of Mismatch*, *supra* note 4, at 77 ("[F]irst-choice students with a given LSAT and UGPA were typically in a more elite tier . . . ."); *see also* Sander, Whitepaper, *supra* note 4, at 7 (providing more detailed discussions of his first choice and tier location analyses).

344. Sander, Whitepaper, *supra* note 4, at 7. *But see infra* Appendix C (describing important methodological shortcomings impacting Sander's analysis of the impact of second (or third) choice on tier location and first-time bar passage that call into question the inferences he draws from these models).

> *concerned about matriculation patterns within race, which could bias our results.*[345]

To paraphrase Ayres and Brooks, it appears that black and white students forgo attending their first choice schools for systematically different reasons. These differences may be attributable to, *inter alia*, the aforementioned limitations in the BPS data that may be correlated with race. In statistics jargon, there is a strong likelihood of correlated covariate measurement error. This means the values of first choice variable that we observe differ systematically, by race, from the values that we are interested in. I discuss the problem of covariate measurement error bias in greater detail in Subpart II.F.2. These problems, alone, render Sander's attempted analysis of the first choice effect on tier location unreliable, and the biasedness of Sander's results are further compounded for reasons that I have previously described, namely omitted variable bias (confoundedness),[346] interpolation bias (incorrect functional form assumptions within the range of observed data),[347] and extrapolation bias (incorrect functional form assumptions outside the range of observed data).[348] Stated differently, Sander's attempt to assess the core assumption of the first choice test is, itself, plagued by many of the same problems that bias his assessments of the effect of mismatch on bar passage rates.[349] And as I have already explained, even if a respondent's second (or third) choice results in them declining to attend a second tier school in in favor of a third tier school, the relevant mismatch characteristics of the two schools—UGPA, LSAT, and school selectivity—are likely to be indistinguishable.

Even putting these problems to the side, the analytical approach employed by Sander is not particularly illuminating because he does not explore whether the first choice effects on tier location are significantly different between black and white students.[350] These crossgroup comparisons would serve as tests of interaction/conditional effects—whether the first choice effects on tier location were moderated (accentuated or attenuated) by

345. Ayres & Brooks, *supra* note 12, at 1832 n.49 (emphasis added).
346. *See supra* Subpart II.C.
347. *See supra* Subpart II.D.
348. *See supra* Subpart II.E.
349. *See* King & Zeng, *supra* note 156, at 147 (describing requisite assumptions for establishing treatment effects); *accord* ANDREW GELMAN & JENNIFER HILL, DATA ANALYSIS USING REGRESSION AND MULTILEVEL/HIERARCHICAL MODELS 167 (2007).
350. *See* Ayres & Brooks, *supra* note 12, at 1832 n.49 (noting that the factors influencing applicants' decisions to forgo their first choice school differ between black and white students).

race.[351] Obviously, race is "formed prior to [the treatment variables] tak[ing] on their current values recorded in the data. [T]he fixed group [race] is not the cause of interest . . . ; rather, the data analyst's causal focus is on the [treatment] variables and how they may differ between groups."[352] The mismatch-relevant question is not whether black or white students who elect to attend their second (or third) choice school are more likely to attend a school in a lower tier than black or white students with the same UGPA and LSAT who elect to attend their first choice school. Rather, it would be informative to know whether there are meaningful differences in the effect of the first choice on tier location between black and white students.[353] The more

351. *See id.*; Dauber, *supra* note 238, at 1906 (noting that Sander's models in "Systemic Analysis" failed to test crossgroup comparisons and such comparison partly captures variation in causal effects).

352. Tim F. Liao, *Group Differences in Generalized Linear Models*, *in* HANDBOOK OF CAUSAL ANALYSIS FOR SOCIAL RESEARCH, *supra* note 48, at 153, 155; *see also* GELMAN & HILL, *supra* note 349, at 178 (discussing interactions of treatment effect with pretreatment variables).

353. It is important to emphasize that, under my suggested test, race is not the causal variable; first choice is. The analysis would simply capture variation in the treatment effect by race. *See* PEARL, *supra* note 50, at 127 (distinguishing between different types of direct effects and positing that racial discrimination turns on direct effects with interactions between race and a mediator). This is an important distinction because race is an immutable characteristic and, therefore, (1) it is not subject to change by intervention (and the goal of causal inference is to discover what would happened if we changed the world in some way), and (2) race is determined at a person's birth and all measured variables specific to a person are posttreatment.

Sander seems to misunderstand this point when responding to Ho's criticism of the use of white students as the control group and black students as the treatment group. Sander notes, "First, Ho suggests that my article is flawed because there is no 'control' group—a group that has not received racial preferences to whom blacks can be compared. Not so: The entire paper is organized around a comparison of 'treatment' blacks (who generally receive preferences) and 'control' whites (who generally do not)." Sander, *Mismeasuring*, *supra* note 24, at 2006. The Empirical Scholars Brief expressly identifies Sander's black-white comparisons as invalid, explaining:

> This comparison violates the principle of creating groups that are comparable in all pre-existing respects except for law-school tier. Usual tenets of research design require that a study hold constant pre-existing attributes such as race and gender. . . . By comparing *black* students at higher-tiered schools with *white* students at lower-tiered schools, Sander and [Doug] Williams violate these basic principles.

Empirical Scholars Brief in *Fisher I*, *supra* note 30, at 21–22 (citation omitted). Another influential causal inference scholar underscored this point when he noted:

> Causal variables are those that reflect such manipulations or varying experiences between units of study. . . . It inherently involves the use of counterfactual conditional statements (the result that would have obtained had the unit been exposed to another experience . . .). Properties or attributes of units are not the types of variables that lend themselves to plausible states of counterfactuality. For example, because I am a White person, it would be close to ridiculous to ask what would have happened to

important takeaway, however, is that Sander's first choice analysis on law school outcomes—which he posits provides strong support for mismatch theory—are doubly-contaminated by violations of the SUTVA: (1) the tier hierarchy does not accurately reflect whether an individual is mismatched and (2) the second or third choice analysis does not accurately reflect whether an individual actually enrolled in school that was less competitive than the individual's first or second choice school to which they were accepted.

c. Bias From Interference Between Units

The third requirement of the SUTVA is that there is no interference between treatment and nontreatment units that would influence potential outcomes. This means that there can be no contamination. This component naturally follows from the first two components because when there is interference between treatment and nontreatment units, one can neither determine which units did not receive the treatment (the control group) nor how much of the treatment was received by individual units in the treatment group (when the treatment dosage can vary). Consequently, the treatment effect cannot be properly assessed. A primary source of interference in studies with human subjects is their social interactions with one another, and it is often the case that such interference cannot be eliminated.[354] A textbook example of a violation of the SUTVA assumption in the education context is when students assigned to attend a tutoring program to improve their grades (the treatment group) interact with other students in their school who were

> me had I been Black. Yet, that is what is often meant when *RACE* is interpreted as a causal variable.
>
> PAUL W. HOLLAND, EDUC. TESTING SERV., RESEARCH REPORT NO. 03-03, CAUSATION AND RACE 8–9 (2003) (emphasis omitted and added) (citation omitted); *accord* Camilli & Welner, *supra* note 9, at 506 ("It is important to add that comparisons between members of different racial or ethnic groups . . . cannot provide an estimate of a causal effect of being a member of such a group. This is because a student cannot be randomly assigned to such categories.").
>
> *See* IMBENS & RUBIN, *supra* note 214, at 264 (explaining that it is implausible to compare treatment and control groups if the groups differed in terms of pretreatment characteristics such as gender). *But see* D. James Greiner & Donald B. Rubin, *Causal Effects of Perceived Immutable Characteristics*, 93 REV. ECON. & STAT. 775 (2011) (claiming that the examination of the effect of perceptions of race, and not race itself, can conform to the requirements of causal inference in experimental settings).

354. VANDERWEELE, *supra* note 64, at 397; GUANGLEI HONG, CAUSALITY IN A SOCIAL WORLD: MODERATION, MEDIATION AND SPILL-OVER 8 (2015) ("[D]ue to social interactions among individuals or groups of individuals, an intervention received by some may generate a spill-over impact on others affiliated with the same organization or connected through the same network.").

not assigned to the tutoring program (the control group) in ways that influence the grades of these nontreated students.

The SUTVA is unlikely to hold when assessing the impact of affirmative action on LGPA—especially first year LGPA—because the LGPA variable is standardized (set to curve) within law schools and captures class rank. If black students are mismatched because of affirmative action, then their performance influences the rankings of other students in their class. When mismatched black students occupy the lower distribution of the curve, they simultaneously elevate the class rank of nonmismatched students. The academic rankings of white students (the nontreated students) within a class are higher than they would normally be, holding UGPA and LSAT constant, because their LGPA is influenced by the academic performance of the students who benefit from affirmative action (the treated students). Sander claims that the class rank is a proxy for how much a student has learned, but this assumption is nonsensical because by simply removing mismatched students at the bottom of the grade distribution, the nontreated students' LGPA would fall by a corresponding amount; they would appear lower on the grade distribution although they did not actually learn less. This point has been underscored by others. "What [Sander] fails to tell us is that there was a lower tenth comprised of white students well before there was any affirmative action. There will also be a bottom tenth at the elite law schools and every other law school, under any regime we have."[355] Similarly, Ho notes the SUTVA may be violated because, in a world without affirmative action, grade curves and attendance patterns would have differed more systematically.[356]

The SUTVA would likely be violated even if one focused exclusively on black students because, according to Sander, one of the negative consequences of affirmative action is "social mismatch." Specifically, he claims that students are significantly more likely to make friends with other students who have similar levels of academic preparation and academic performance and law schools "create[e], through the use of admissions preferences, large gaps in

355. Cheryl I. Harris, *An Affirmative Act? Barack Obama and the Past, Present, and Future of Race-Conscious Remedies*, *in* THE OBAMA PHENOMENON: TOWARD A MULTIRACIAL DEMOCRACY 277, 304–05 n.13 (Charles P. Henry et al. eds., 2011) (quoting Charles E. Daye, *A Personal Perspective—Ten Reasons to Reject "A Systemic Analysis of Affirmative Action in American Law Schools"* 6 (UNC Legal Studies Rsch. Working Paper, Paper No. 927225, 2006)).

356. Ho, Evaluating Affirmative Action, *supra* note 23, at 3 ("Sander assumes [no interference among units] . . . I do not address this in our [sic] analysis, but to the degree that inference exists, estimates are both biased and falsely precise.").

academic preparation across distinct ethnic groups."[357] According to Sander, this social mismatch not only hinders academic achievement, but also undermines the diversity rationale of affirmative action.[358] Even if Sander is correct in that students choose friendship networks based, in part, on academic preparation and performance, there are clearly other reasons these networks form that are independent of performance, such as race/ethnicity.[359] So, for example, black students who are not mismatched are more likely to interact with greater priority, frequency, intensity, and duration with black students who are mismatched (and vice versa), and these interactions are likely to influence the academic performance of both groups of black students. Thus, the "spillover effect" remains relevant and mismatch effects cannot be properly identified at the individual level. In some situations, inference can only occur within a particular setting ("partial interference"), so causal inference can be carried out if other assumptions hold, but the effect can only be measured at the setting level.[360]

2. Correlated Measurement Error

Generally speaking, measurement error bias is of particular concern when the error in measuring a variable of interest is correlated with: (1) the true value of that variable; (2) the true values of other variables in the model; or (3) the errors in measuring those values.[361] Figure 17 illustrates the first condition. The latent constructs of interest (aptitude, selectivity, and knowledge) may be correlated with the measurement error of the manifest constructs. The dotted curved double-sided arrows represent potential correlations. So, for example, measurement error associated with school selectivity/competitiveness may be correlated with tier location. As a consequence, the (in)ability to properly capture meaningful mismatch-related

357. Sander, *Stylized Critique*, *supra* note 8, at 1643.
358. *Id.* ("Creating, through the use of admissions preferences, large gaps in academic preparation across distinct ethnic groups on campus can thus directly undermine the specific benefits campus diversity is supposed to achieve.").
359. *See, e.g.*, BEVERLY DANIEL TATUM, WHY ARE ALL THE BLACK KIDS SITTING TOGETHER IN THE CAFETERIA? (rev. ed. 2003) (explaining the dynamics leading to students' so-called self-segregation into racial/ethnic groups in educational institutions).
360. VANDERWEELE, *supra* note 64, at 426; HONG, *supra* note 354, at 371.
361. *See* John Bound, Charles Brown & Nancy Mathiowetz, *Measurement Error in Survey Data*, *in* 5 HANDBOOK OF ECONOMETRICS 3705, 3712 (James J. Heckman & Edward Leamer eds., 2001). Classical measurement error assumes that the error is random so, on average, its value is zero. Sander's claim concerning measurement error would only be correct if the measurement errors in the BPS data do not exhibit one or more of these biases.

distinctions based on differences in tier location may partly depend on whether one is comparing tiers near the top or bottom of the tier hierarchy.[362] Similar concerns may exist for measurement errors impacting aptitude (proxied by UGPA/LSAT) and legal knowledge (proxied by LGPA), which is to say that the accuracy of the measures of underlying phenomena may depend on their own values. So, for example, the higher (or lower) one's UGPA/LSAT, the better (or worse) UGPA/LSAT is able to capture aptitude.

The second condition occurs when the measurement error of a latent construct is related to a different latent construct (Figure 18) or manifest construct (Figure 19). So, for example, a student's aptitude for legal study may be correlated with measurement error for selectivity/competitiveness or LGPA. Similarly, a student's race/ethnicity or gender may correlate with measurement errors for aptitude or legal knowledge. For example, Kidder's research on racial disparities in LSAT performance for applicants to UC Berkeley's law school over a three-year span discovered strong racial/ethnic differences in LSAT performance even after taking into account applicants' age, undergraduate GPA, major, college/university attended, and year of college/university graduation. Specifically, relative to white applicants, black applicants scored 9.1 points lower, Latinx applicants scored 7 points lower, and Asian Americans scored 3.6 lower.[363]

Finally, Figure 20 illustrates the third condition. This condition differs from the first two measurement error conditions because it does not concern the relationship between the actual value of the theoretical construct and measurement errors; rather, it involves the potential correlation between the error terms of manifest constructs. An example of this would be between when measurement errors associated with UGPA/LSAT correlate with measurement errors for LGPA. In other words, the greater the mismeasurement of aptitude by UGPA/LSAT, the greater the mismeasurement of the legal knowledge by LGPA.[364]

362. Similarly, measurement error associated with first choice (as it relates to mismatch) may be correlated with its own value (for example, third choice options would be less likely to accurately capture mismatch effects than second choice options).

363. Kidder, *supra* note 299, at 1074–79 (noting that the racial gap in LSAT scores that favors white law applicants relative to all other racial groups remains after taking into account undergraduate institution, college major, UGPA, and grade inflation).

364. *See* discussion *infra* Subpart V.A.6.b, concerning the relationship between race/ethnicity, UGPA/LSAT, and assessment methods (in-class exam, take-home exam, and paper) in law school.

Figure 17: Sander's Causal Model (Condition 1 Measurement Errors)

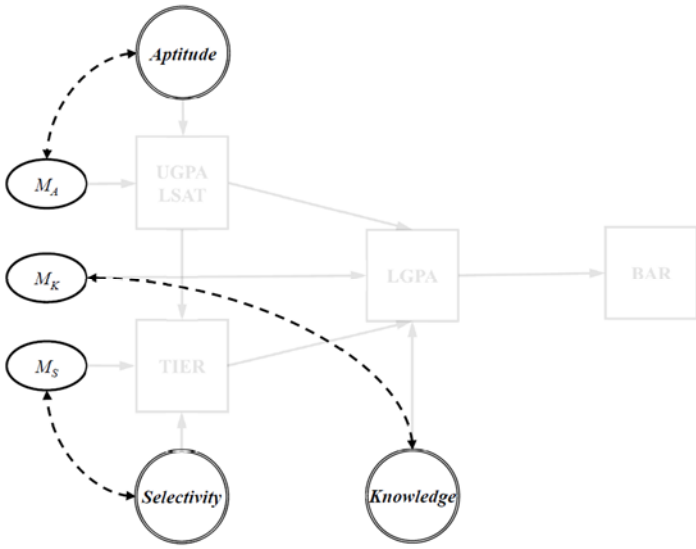


Figure 18: Sander's Causal Model (Condition 2 Measurement Errors)

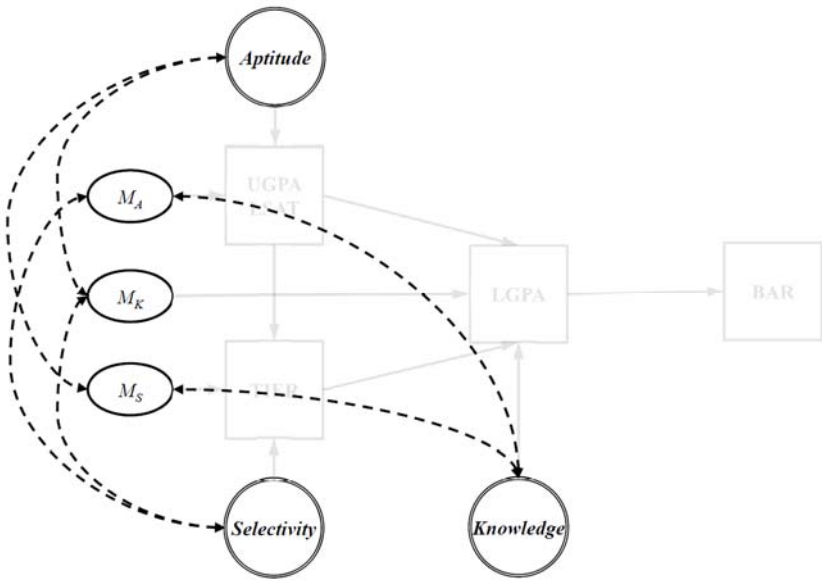


Figure 19: Sander's Causal Model (Condition 2 Measurement Errors)

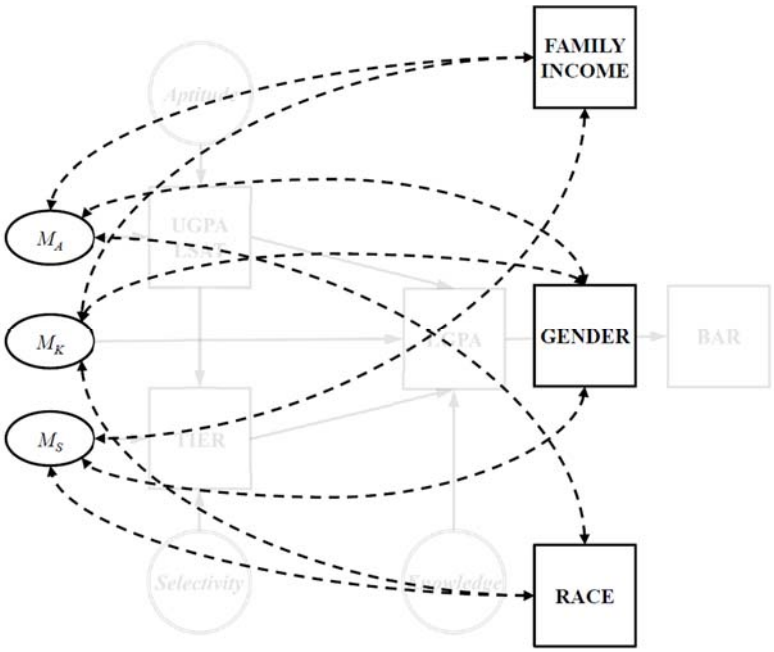
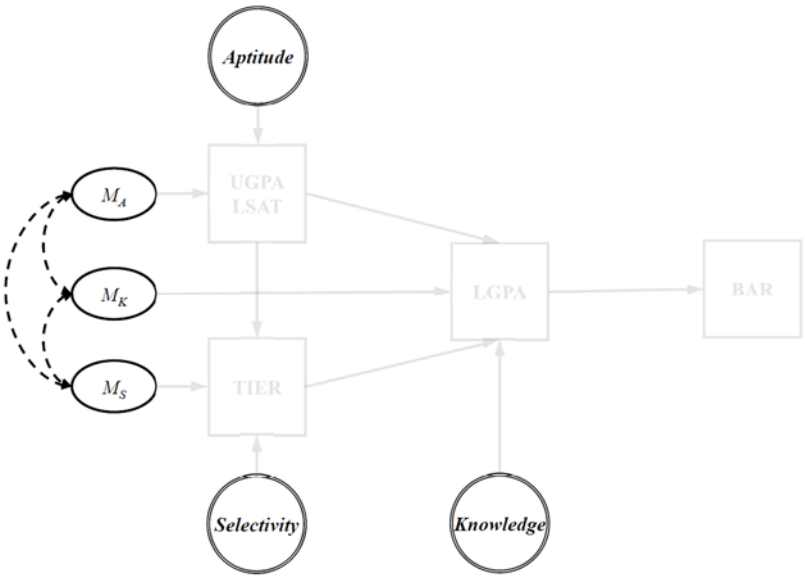


Figure 20: Sander's Causal Model (Condition 3 Measurement Errors)



The measurement problems associated with the LSAT merit further discussion for three reasons: (1) it is the more heavily weighted measure of the latent construct of aptitude by law schools during their admissions process; (2) it is, purportedly, a stronger predictor of FYGPA than UGPA; and (3) it figures prominently in research on mismatch effects in law school. The LSAC, which administers the LSAT, has repeatedly cautioned that the LSAT test suffers from measurement error, so an individual's true score lies within an interval (called a score band).[365] The underlying logic is that if an individual were to take the LSAT multiple times, then their score would fall within a certain range most of the time. LSAC reports a general measure of reliability for the LSAT in the form of a score band (plus/minus 2.6 points of the obtained score), which roughly represents that the particular testtaker's score would roughly fall into this interval 7 out of 10 times the exam was taken.[366] Of course, researchers have recognized for quite some time that the calculated reliability for tests like the LSAT are global (average) values, and the actual reliability for a specified score differs at various points along the score scale.[367] This is an example of the first measurement error condition in which the error in measuring a variable of interest is associated with the true value of that variable. And research has revealed that measurement error for test scores often reaches its peak in the middle of the score range, with the maximum error often more than twice the minimum. As Leonard S. Feldt would describe it, "the standard [average] error of measurement computed by the traditional formula for the test as a whole does not adequately summarize the error propensity of many—perhaps most—examinees."[368] Sander's analysis incorrectly treats a respondent's LSAT score as if it accurately captures their underlying ability without properly taking into the uncertainty in this measure. As previously explained, error-prone measurements may fail to balance the treatment and control groups with respect to the confounding variables, and Sander's analyses are susceptible to this problem. For quite some time, social scientists have advocated the use of techniques that incorporate the level of imprecision in the measurement of a covariate if the reliability of the measurement of that covariate is known in advance (as in the case of the LSAT) or can be estimated

365. *LSAT Score Bands*, LSAC, http://www.lsac.org/jd/lsat/your-score/score-band [https://perma.cc/8ATX-M6W8] (last visited Nov. 11, 2019).
366. *Id.*
367. Leo M. Harvill, *Standard Error of Measurement*, EDUC. MEASUREMENT: ISSUES & PRAC., Summer 1991, at 33, 35 (underscoring that there are different standard errors of measurement at different points on the score scale).
368. *Id.* at 35 (quoting Leonard S. Feldt et al., *A Comparison of Five Methods for Estimating the Standard Error of Measurement*, 9 APPLIED PSYCHOL. MEASUREMENT 351, 358 (1985)).

from the data (if there are multiple measurements of the same underlying phenomenon).[369] Sander neither mentions nor adopts any of these approaches, but there is good reason to believe measurement error biases Sander's results because of one or more of the aforementioned measurement error conditions. Although it is difficult to know, *a priori*, the magnitude and direction of the bias from measurement error, there are tests available that enable the analyst to make that assessment.[370]

Measurement error associated with the LSAT is also relevant to its low predictability of law school grades. Nearly every study of the relationship between UGPA, LSAT, and law school grades reveals that these variables account for less than one-quarter of the variation in LGPA.[371] Part of the low explanatory power of these models is likely attributable to the fact that the LSAT's predictive ability is highly variable across various types of assessment methods. A study examining the relationship between UGPA, LSAT, law school grades (in each year), and the type of assessment method used for the class (in-class exam, take-home exam, and paper) discovered that the LSAT is correlated with 1L in-class exam scores, but not scores on take-home exams and papers in the first-year year.[372] UGPA, on the other hand, is correlated with all three forms of assessment.[373] The effect of LSAT, controlling for UGPA, declines significantly after the first year for all assessment methods.[374]

369. *See, e.g.*, Jesse M. Rothstein, *College Performance Predictions and the SAT*, 121 J. ECONOMETRICS 297, 313–14 (2004) (noting the less than perfect reliability of the SAT and specifically accounting for this fact in the analyses through an errors-in-variables correction when analyzing the impact of SAT on college grades); James W. Hardin, Henrik Schmiediche & Raymond J. Carroll, *The Regression-Calibration Method for Fitting Generalized Linear Models with Additive Measurement Error*, 3 STATA J. 361, 363 (2003) ("Clearly, if we had an estimate of the unknown measurement error variance, we could adjust the results from an analysis using the error-prone covariates . . . by requiring a specification of the attenuation factor (reliability ratio)."); Sophia Rabe-Hesketh, Andrew Pickles & Anders Skrondal, *Correcting for Covariate Measurement Error in Logistic Regression Using Nonparametric Maximum Likelihood Estimation*, 3 STAT. MODELLING 215, 216–21 (2003) (describing a latent variable approach to account for measurement error for models with binary outcomes); Blackwell et al., *supra* note 296, at 311–13 (developing a multiple overimputation framework to address error-prone variables).

370. *See, e.g.*, VanderWeele & Hernán, *supra* note 298; Blackwell et al., *supra* note 296.

371. *See supra* notes 176, 341 (noting the LSAT's low predictive ability of law school grades across a range of studies).

372. Henderson, *supra* note 299, at 1043–44.

373. This finding rebuts the criticism that grades on take-home exams or papers are more subjective and less likely to produce a discernable grading curve. *Id.* at 981 n.29.

374. In the study, the proportion of in-class exams over three years of law school steadily declined from 74.9 percent (first year) to 60.3 percent (second year) to 52.1 percent (third year) (61.3 percent overall). The proportion of take-home exams varied from 13.6 percent (first year) to 16.0 percent (second year) to 15.7 percent (third year) (15.2 percent

Testtaking speed, which is captured by the LSAT, appears to drive the relationship between the LSAT and first-year grades. While test taking speed is relevant to the bar exam, strong reasoning ability is more important for a career in legal practice, and this ability is needed for strong performance across all assessment methods. The study suggests that UGPA is better at capturing motivation, persistence, and writing ability than the LSAT. This may be especially relevant because UGPA and LSAT are both proxies for aptitude, but they are differentially related not only to the various proxies for legal knowledge (based on assessment method), but also the particular class year (1L, 2L, or 3L) in which the assessment are made. When looking beyond just the proxies for aptitude and legal knowledge and incorporating race/ethnicity into the inquiry, the study found that that black students had the highest differentials, by far, between in-class and other methods at a national law school (but the difference was not statistically significant at a regional law school). For Latinx students, the difference was statistically significant at the regional school, but not the national school. But given that the number of black and Latinx students in the study was pretty small, the lack of statistical significance in some settings is likely attributable to the limited sample size for these groups.

In summary, there appears to be considerable cause for concern about various types of measurement error bias impacting assessments of mismatch theory using the BPS data. Instability bias impacts at least some of the tier location comparisons (ambiguous treatment assignment), the first choice analysis (ambiguous treatment assignment and ambiguous treatment dosage), and the assessment of mismatch on LGPA (interference between units). A direct assessment of the impact of the different types of correlated measurement error bias is not possible given the limitations of the BPS data, but prior studies strongly suggest that correlated measurement error creates genuine causes for concern when testing mismatch theory,[375] and approaches for handling some of these measurement errors have been well established.[376] Several of the shortcomings in the BPS data that contributed to these potential biases were identified by the LSAC several years before Sander's initial publication in the *Stanford Law Review* purporting to discover empirical

overall), papers steadily increased from 11.6 percent (first year) to 19.6 percent (second year) to 26.6 percent (third year) (20 percent overall), and clinical courses increased from 2.6 percent (second year) to 4.3 percent (third year) (2.5 percent overall). *Id.* at 1017 tbl.17.

375. *LSAT Score Bands*, *supra* note 365; WIGHTMAN, CLUSTERING, *supra* note 308; Kidder, *supra* note 299; Henderson, *supra* note 299.

376. *See supra* note 369.

support for mismatch theory.[377] Critics of Sander's work—especially those who have uncovered contradictory evidence with their own analyses of the BPS data—have been much more transparent about the inherent limitations of the BPS data, from the perspective of both highlighting the problems initially identified by the LSAC and specifically acknowledging how these problems may call into question the credibility of their conclusions.

III. A NOTE ON THE EMPIRICAL SCHOLARS BRIEF

The *Fisher I* Empirical Scholars Brief (ESB) was signed by eleven theoretical and applied statisticians.[378] Many of the signatories are considered giants in the field of causal inference and I can only think of a handful of prominent causal inference scholars who were not signatories.[379] The ESB

377. WIGHTMAN, CLUSTERING, *supra* note 308 (identifying a measurement error with law school clusters); LINDA F. WIGHTMAN, LAW SCH. ADMISSION COUNCIL, RESEARCH REPORT NO. 99-05, BEYOND FYA: ANALYSIS OF THE UTILITY OF LSAT SCORES AND UGPA FOR PREDICTING ACADEMIC SUCCESS IN LAW SCHOOL (2000) [hereinafter WIGHTMAN, BEYOND FYA] (identifying a potential correlated measurement error between race/ethnicity, UGPA, and LSAT); WIGHTMAN, BAR PASSAGE STUDY, *supra* note 36 (identifying measurement error associated with use of within-school standardized LGPA).

378. The signatories included: Guido Imbens (Stanford University—Economics), Donald B. Rubin (Harvard University—Statistics), Gary King (Harvard University—Political Science), Richard A. Berk (University of Pennsylvania—Criminology & Statistics), Daniel E. Ho (Stanford University—Law & Political Science), Kevin M. Quinn (UC Berkeley—Law & Political Science), D. James Greiner (Harvard University—Law & Statistics), Ian Ayres (Yale University—Law & Economics), Richard Brooks (Yale University—Law & Economics), Paul Oyer (Stanford University—Economics), and Richard Lempert (University of Michigan—Law & Sociology). *See* Empirical Scholars Brief in *Fisher I*, *supra* note 30, at 1–7. Since the filing of the ESB, Brooks moved to Columbia University and then to New York University, and Quinn moved to the University of Michigan.

379. Notable causal inference scholars who were not signatories on the ESB include Andrew Gelman (Columbia University—Political Science & Statistics), James Heckman (University of Chicago—Economics), Judea Pearl (UCLA—Computer Science), Paul Rosenbaum (University of Pennsylvania—Statistics & Wharton), and Christopher Winship (Harvard University—Sociology). I have not contacted these scholars to determine whether they were asked to be a signatory, so one cannot draw any inferences about the omission of their names from the ESB.

Rosenbaum's work is cited in most of the critiques of mismatch, including in the ESB, to directly challenge Sander's methodological approach—namely, the problems of posttreatment bias and omitted variable bias. *See supra* notes 76, 201 and accompanying text. Andrew Gelman's work on posttreatment bias is also cited in the ESB. Empirical Scholars Brief in *Fisher I*, *supra* note 30, at 22 (citing GELMAN & HILL, *supra* note 349, at 188–90). James Heckman has extensively written about sample selection bias, interpolation bias, and extrapolation bias—all issues that continue to impact Sander's work. *See* Heckman, *supra* note 167; Heckman et al., *supra* note 245. While Heckman's substantive research has not focused specifically on mismatch, he has examined the

began by stating, "Based on over 255 collective years of social-science research experience, *amici* have concluded that the research on which [the Sander-Taylor brief in *Fisher I*] relies, Professor Sander's 'mismatch' hypothesis, is unreliable, failing basic tenets of research design."[380] Kidder and Lempert underscore this point when they write:

> Far better methodologists than Sander have reached this conclusion [that research purporting to find mismatch effects suffers from fundamental flaws]. The definitive statement is found not in the usual scientific literature, but in an amicus brief submitted to the Supreme Court in response to a brief that Sander and Stuart Taylor filed in conjunction with *Fisher v. University of Texas at Austin* (2013). Signers evaluating Sander's data and methods included leading

economic effects of affirmative action policies on black students, and concluded that increased education attainment among black students, coupled with enforcing employment discrimination laws, did more to decrease the racial wage gap than other workplace-focused interventions. James J. Heckman & J. Hoult Verkerke, *Racial Disparity and Employment Discrimination Law: An Economic Perspective*, 8 YALE L. & POL'Y REV. 276, 288 (1990) ("[T]he passage of Title VII of the Civil Rights Act of 1964, and the adoption of affirmative action requirements for federal contractors all occurred at the same time, [and] many scholars believe government policy played an important role in improving black economic status.").

The ESB also cited an article coauthored by Christopher Winship that examined the impact of affirmative action on black students' graduation rates (at the undergraduate level) at twenty-seven elite institutions. Not only did Winship find that there is no empirical support for the "academic difficulty model" (mismatch), his study concluded that school selectivity (which, for many black students, is impacted by affirmative action) was the only variable that had any significant effect on graduation: "Contrary to common belief, selectivity improves black probabilities of graduation, and helps blacks more than it helps whites." Mario L. Small & Christopher Winship, *Black Students' Graduation From Elite Colleges: Institutional Characteristics and Between-Institution Differences*, 36 SOC. SCI. RSCH. 1257, 1257 (2007). *See also* WILLIAM G. BOWEN, MATTHEW M. CHINGOS & MICHAEL S. MCPHERSON, CROSSING THE FINISH LINE: COMPLETING COLLEGE AT AMERICA'S PUBLIC UNIVERSITIES 214 (2009) (examining twenty-one "flagship" public universities, as well as public university systems in four states, and finding "absolutely no evidence of any 'mismatch' or 'overmatch' problem" for black and Hispanic students); WILLIAM G. BOWEN & DEREK BOK, THE SHAPE OF THE RIVER: LONG-TERM CONSEQUENCES OF CONSIDERING RACE IN COLLEGE AND UNIVERSITY ADMISSIONS 59–65 (1998) (studying various outcomes for students at twenty-eight mostly private colleges and universities and finding no support for the mismatch hypothesis and some evidence of a reverse mismatch effect); CHARLES ET AL., *supra* note 8, at 199 (analyzing the impact of race-conscious admissions on undergraduate GPA and finding "no empirical support for the mismatch hypothesis" and explaining "[o]ther things equal, individual affirmative action beneficiaries earn the same grades as other students on campus"); Alon & Tienda, *supra* note 256, at 294 (examining multiple nationally representative datasets and concluding that the "findings do not support the 'mismatch' hypothesis for black and Hispanic (as well as white and Asian) students who attended college during the 1980s and early 1990s").

380. Empirical Scholars Brief in *Fisher I*, *supra* note 30, at 1.

> methodologists in economics, law, political science, sociology, and statistics . . . concluded: "[W]hether one finds Sander's conclusions highly unlikely or intuitively appealing, his 'mismatch' research fails to satisfy the basic standards of good empirical social-science research."[381]

Kidder and Onwuachi-Willig note that part of Sander's effort to rebut the collective wisdom of the foremost causal inference scholars has been "peculiar [when] speculat[ing] as to the reason that 'most of the distinguished signatories' agreed to sign the brief."[382] According to Sander:

> [T]he Empirical Scholars Brief, filed in August 2012, argued that the literature on law school mismatch theory has not 'proven' that these effects occur, because the research is not based on the sort of methods that scientists would usually require for convincing proof—such as, for example [sic], a randomized scientific experiment. This is an interesting point, and probably the point that persuaded most of the distinguished signatories to the brief to join it.[383]

But Sander mischaracterizes the critique presented in the ESB.[384] In fact, several of the flaws in Sander's work identified in the ESB (for example, posttreatment bias, extrapolation bias, improper interracial comparisons) would negatively impact his results even if Sander were able to conduct a randomized experiment.[385] At the outset of the ESB, its signatories explain that:

> Sander's research has major methodological flaws—misapplying basic principles of causal inference—that call into doubt his controversial conclusions about affirmative action. . . . [Sander's] research, which consists of weak empirical contentions that fail to

381. Kidder & Lempert, *Mismatch Myth*, *supra* note 22, at 108–09 (citations omitted) (quoting Empirical Scholars Brief in *Fisher I*, *supra* note 30, at 27).

382. William C. Kidder & Angela Onwuachi-Willig, *Still Hazy After All These Years: The Data and Theory Behind "Mismatch"*, 92 TEX. L. REV. 895, 920 n.112 (2014) (reviewing SANDER & TAYLOR, *supra* note 4).

383. Sander, Brief in *Schuette*, *supra* note 113, at 26.

384. Taken at face value, Sander's response to the ESB would render all nonexperimental research incapable of providing convincing proof; clearly, this is not the case and the credibility of statistical inference from nonexperimental data hinges on the plausibility of the underlying assumptions of the statistical model. As econometricians James Heckman and Richard Robb have explained, "Social scientists *never* have access to true experimental data of the type sometimes available to laboratory scientists. . . . In the place of laboratory experimental variation, social scientists use subjective thought experiments. Assumptions replace data." James J. Heckman & Richard Robb, *Alternative Methods for Solving the Problem of Selection Bias in Evaluating the Impact of Treatments on Outcomes*, *in* DRAWING INFERENCES FROM SELF-SELECTED SAMPLES 63, 63 (Howard Wainer ed., 1986).

385. *See supra* Subparts II.A, II.E.

> meet the basic tenets of rigorous social-science research, provides no basis for this Court to revisit longstanding precedent supporting the individualized consideration of race in admissions. . . . In light of the significant methodological flaws on which it rests, Sander's research does not constitute credible evidence that affirmative action practices are harmful to minorities, let alone that the diversity rationale at the heart of *Grutter* is at odds with social science.[386]

These "basic principles of causal inference" apply to both experimental and observational research, and as ESB signatory Gary King notes:

> [E]ven in the best experimental work, some information goes missing, randomly selected subjects sometimes refuse to participate, some subjects do not comply with treatment assignments, random numbers do not always become assigned as planned or must be assigned at a more aggregated level than desired and outcomes are not always measured correctly or recorded appropriately. To account for these problems, when they cannot be fixed through better data collection, more sophisticated methods become necessary.[387]

Similarly, Richard Berk, another signatory of the ESB, cautioned that "[a] careful look at randomized experiments will make clear that they are not the gold standard. But then, nothing is."[388] In addition to King and Berk, many of the other ESB signatories have emphasized that potential bias resulting from inconsistent treatment assignment/dosage (in violation of the SUTVA), sample selection, nonresponse (for example, attrition), and treatment and control group imbalance are equally applicable to experimental and observational studies.[389]

386. Empirical Scholars Brief in *Fisher I*, *supra* note 30, at 8–9 (citations omitted).

387. Kosuke Imai, Gary King & Elizabeth A. Stuart, *Misunderstandings Between Experimentalists and Observationalists About Causal Inference*, 171 J. ROYAL STAT. SOC'Y (SERIES A) 481, 499 (2008); *see also* King & Zeng, *supra* note 156, at 146.

388. Richard A. Berk, *Randomized Experiments as the Bronze Standard*, 1 J. EXPERIMENTAL CRIMINOLOGY 417, 417 (2005) (noting that randomized experiments have many of the same implementation problems as observational studies); *see also* Robert J. Sampson, *Gold Standard Myths: Observations on the Experimental Turn in Quantitative Criminology*, 26 J. QUANT. CRIMINOLOGY 489 (2010) ("[U]sers of experimental methods need to directly address fundamental assumptions and limitations that come into play when we are in a social world.").

389. Berk, *supra* note 388, at 429; *accord* Donald B. Rubin, *Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies*, 66 J. EDUC. PSYCHOL. 688, 695 (1974); Greiner, *supra* note 64, at 564; Ho, *supra* note 7, at 2000 n.16; IMBENS & RUBIN, *supra* note 214, at 31; *see also* Heckman & Robb, *supra* note 384, at 63 n.2 (noting there are "very real problems of designing or conducting true social experiments," including

Sander has continued to advance contradictory and unsupported claims about the ESB, including that: (1) "[the ESB] is fraudulent to its core";[390] (2) "most of the signatories of the ESB were [most likely] not involved in its actual drafting, signed on as a favor to friends who happened to be mismatch critics, and are now deeply embarrassed to have been associated with it";[391] (3) the ESB is a "Case Study of Zealotry" and "[t]he principal author was, in all likelihood, . . . the most zealous of the Zealots";[392] (4) three *different* scholars "signed []and apparently spearheaded[]" the ESB;[393] and (5) there "is no matter of interpretation or argument: the authors of the ESB simply lie."[394] The final point merits further discussion because it illustrates what has been the main obstacle to engaging in fruitful discussions about the impact of race-conscious admissions policies to help inform the judges, legislatures, and the general public: Sander's tendency to employ deceptive or evasive argumentation when fundamental flaws are identified in research that claims to provide support for the mismatch hypothesis.[395] As English philosopher and pioneer of the scientific method, Sir Francis Bacon, famously noted over four hundred years ago, "Truth emerges more readily from error than from confusion."[396]

In Sander's amicus brief in *Fisher II*, he accuses the ESB of "either [] not read[ing] the [mismatch] work they were critiquing, or deliberately misrepresent[ing] it,"[397] and concludes that "[i]n any case, the ESB is

nonrandom selection and nonrandom missing, which "make experimentation problematic if not infeasible [and a]lleged solutions to the[se] problems of self selection and attrition . . . *inject* into the analysis of experimental data subjective features which the experiments were proposed to avoid"); Imai et al., *supra* note 387, at 481 (explaining that "both [observationalists and experimentalists] regularly make related mistakes in understanding and evaluating covariate balance in their data").

390. Sander, *Stylized Critique*, *supra* note 8, at 1664.

391. *Id.*

392. *Id.* at 1662, 1664. The scholar to whom Sander refers, Richard Lempert, noted before the publication of Sander's statement that, "I [Lempert] was one of the signers [of the ESB], but the methodological analysis was not mine, and although I concur in the brief 's conclusions, I contributed little to its preparation." Lempert, *supra* note 67, at 20 n.44.

393. Sander, Whitepaper, *supra* note 4, at 2. The three scholars to whom Sander refers, by name, are Ian Ayers, Richard Brooks, and Daniel Ho.

394. Sander, *Mismatch & the ESB*, *supra* note 30, at 568.

395. Kidder & Lempert, *Mismatch Myth*, *supra* note 22, at 108 ("Sander . . . has acknowledged problems with his data and models, but his concessions have never extended to acknowledging fundamental flaws."); *see supra* notes 112–117 and accompanying text (explaining that Sander has attempted to sidestep a key criticism of his empirical analyses—posttreatment bias—by stating that he never claimed to present a causal test of mismatch theory; his current position, however, is directly at odds with his numerous efforts to defend his analytical framework that advanced a causal story).

396. FRANCIS BACON, THE WORKS OF FRANCIS BACON 210 (James Spedding et al. eds, 1875) (1610).

397. Sander, Brief in *Fisher II*, *supra* note 40, at 28.

laughably inept throughout."[398] Sander presents a nearly identical criticism of the ESB in his amicus brief in *Schuette v. Coalition to Defend Affirmative Action*.[399] Sander's statements are in reference to the ESB's assessment of Williams's research which, among other things, criticized some of Williams's statistical models as invalid because those models relied on inappropriate crossracial comparisons.[400] According to Sander:

> [T]he ESB contends that the law school mismatch research is invalid because it does not make "intra-racial comparisons" . . . which is, arguably, important because "interracial" comparisons (e.g., comparing blacks and whites) may fail to control for important interracial differences that affect empirical results. However, the Williams paper, cited above, and cited by the ESB as a prime example of law school mismatch work uses ***only*** intraracial comparisons throughout its entire analysis.[401]

Sander's assertion that the ESB either misunderstood or deliberately mischaracterized Williams's analyses is both incorrect and misleading. Sander cites to Williams's published article in the *Journal of Empirical Legal Studies* in 2013,[402] and based on that version of Williams's article, he claims that all of Williams's analyses involve within-race comparisons. Sander's statement is erroneous in two respects. First, Williams's 2013 article does contain crossracial comparisons. Although the comparisons do not directly test the mismatch hypothesis, they are presented to establish the foundation for mismatch effects in the law school context. Williams presents statistics demonstrating racial differences across several outcome measures

398. *Id.*

399. 572 U.S. 291 (2014); Sander, Brief in *Schuette*, *supra* note 113, at 25 ("[A]ll of Williams' analyses are within-race analyses. None of Williams' analyses are black-white comparisons in the sense criticized by the [ESB]. This and other errors of the [ESB] are so obvious, in fact, that amicus is hopeful that at least some of the authors will retract these claims in due course." (emphasis omitted)).

400. *See supra* note 353 and accompanying text (discussing the inappropriateness of crossracial comparisons for causal inference).

401. Sander, Brief in *Fisher II*, *supra* note 40, at 27–28.

402. Williams, *supra* note 22. Sander puts significant stock in Williams's article published in the *Journal of Empirical Legal Studies* (*JELS*), and, in his amicus brief filed in *Fisher II*, Sander claims that none of his critics have published rebuttals in peer-reviewed journals. Sander, Brief in *Fisher II*, *supra* note 40, at 21. This is also a selective presentation of the facts, as Lempert has publicly noted that he and a coauthor attempted to publish a rebuttal of Williams's article in *JELS*, but the editor of the *JELS* declined to publish Lempert's response. *See supra* note 321 and accompanying text. Lempert has explained that the editor's refusal to publish his response was not based on merits of Lempert's critique, but rather the journal's policy to not publish rebuttals. Lempert Brief, *supra* note 312, at 9–10 n.5.

(graduation rates, first-time bar passage rates, and ultimate bar passage rates).[403] The comparison group is white law students, and Williams reports that black, Native American, and Latinx law students perform worse across all of these outcome measures, with differences most pronounced for first-time bar passage rates.[404] Williams excludes Asian law students, claiming that the group is "more heterogeneous in terms of entering credentials and receive less preferences in law school admissions,"[405] but the BPS data reveal that Asians are also significantly less likely to pass the bar on the first attempt than white students—a difference of about 11 percentage points (91.9 percent for white students; 80.7 percent for Asian students).[406]

By doing this, Williams is explicitly using race/ethnicity as a proxy for affirmative action. He notes that these are unadjusted differences because they do not take into account entering credentials (UGPA and LSAT), and then examines how much of the race/ethnicity gap is explained by entering credentials, irrespective of law school selectivity. This calculation is performed by estimating the impact UGPA, LSAT, and a handful of other background variables on the various outcomes on the subsample of white students,[407] and then using the regression coefficients from this model to predict outcomes for each racial group. Williams reports that a sizable portion of the racial/ethnic gap in first-time bar passage remains after controlling for entering credentials,[408] and uses this finding to support his claim that something other than overall differences in entering credentials must be responsible for differences.[409] In Williams's view, "The mismatch hypothesis explains these racial gaps in performance as a product of too much 'distance' between the academic credentials of minority students and the

403. Williams, *supra* note 22, at 181.
404. *Id.*
405. *Id.* at 181 n.15. On the other hand, Williams treats "Mexican, Puerto Rican, or other Latino in the BPS" as a homogenous group, *id.*, although the BPS data disaggregates this group (but does not disaggregate Asians), and these groups differ in terms of the adjusted unexplained gap in first-time bar passage. *See infra* note 409.
406. WIGHTMAN, BAR PASSAGE STUDY, *supra* note 36, at 27 tbl.6. The gap in first-time bar passage for black, Native American, and Latinx students are, respectively, 31 percent, 26 percent, and 17 percent. Williams, *supra* note 22, at 181 tbl.1.
407. In addition to UGPA and LSAT, Williams's models include mother's education, father's education, family income, disability status, and an English as a second language. Williams, *supra* note 22, at 181 n.16.
408. *Id.* at 181 tbl.1.
409. The size of gap, after taking into account differences in UGPA and LGPA, in first-time bar passage rate significantly differs between Mexicans and "Other Latino," which partly undercuts Williams's claim that Latinx are a homogenous group. *See supra* note 405.

median [or white] student."[410] Williams comparisons are, inarguably, crossracial comparisons.

A simple illustration will underscore this point. Given that the vast majority of the BPS sample consists of white students, the same analysis could have been performed (and nearly identical results would have been obtained) by estimating his equations on the entire sample BPS students and including indicator variables for each racial/ethnic group. The coefficient for each group represents the gap in comparison to white students, and the differences between these coefficients in the unadjusted and adjusted models represent the percentage of the racial/ethnic gap unexplained by the entering credentials.[411] Sander's analyses in "Systemic Analysis" were criticized for similar reasons: Crossracial comparisons are inappropriate. An identical analysis is included in a 2009 prepublication version of Williams's 2013 article.[412] There are also additional analyses present in the 2009 draft that were ultimately excluded from the 2013 version, and Sander's critique of the ESB's assessment of Williams's analyses, which was based on the 2009 version, reveals the second significant flaw in Sander's characterization of the ESB.

The second problem with the manner in which Sander has responded to the ESB is that he neglects to mention that the ESB was filed in 2012, before the publication of Williams's article, and therefore was based on the 2009 prepublication version.[413] Williams makes two additional inappropriate black-white comparisons to directly test various aspects of the mismatch hypothesis in the prepublication version and, unsurprisingly, these erroneous comparisons provided some of the strongest support for mismatch effects across a range of outcomes.[414] In Williams's own words:

410. Williams, *supra* note 22, at 181.

411. The inadvisability of Williams's crossracial comparisons can be illustrated with a simple example. Williams assumes that the impact of entering credentials (UGPA and LSAT) is uniform across racial groups, but the vast literature on racial/ethnic discrimination across a variety of settings suggests otherwise. *See infra* note 478 and accompanying text (discussing research on the persistence of racial discrimination in the legal field). If one explicitly models the variability in the effects of credentials on first-time bar passage, estimates derived from the adjusted model (modeling controlling for UGPA and LSAT) focusing only on black law students to predict first-time bar passage for all other racial/ethnic groups, the overall predicted probability for white students decreases by nearly 7 percentage points (from 92 percent to 85 percent) and the unexplained gap between white and black students increases by approximately 5 percentage points (from 11 percent to 16 percent). Both sets of coefficients may be appropriate for within-group analyses (assuming unconfoundedness), but not between-group analyses.

412. Williams, Whitepaper, *supra* note 292.

413. *Id.*

414. *Id.* app. at 3 tbl.3, 6 tbl.6.

> Given the absence of a reliable estimate of distance for measuring mismatch in the BPS, an alternative strategy is to take advantage of the fact that almost all blacks receive preferences and that no whites receive preferences. Given this feature of affirmative action, one approach is to use "black" as a proxy for being negatively mismatched and "white" as a proxy for being matched.
>
> . . . .
>
> [Ayres and Brooks] employ a variant to the selectivity approach. They use the tier attended by a black relative to the median tier attended by a white with the same academic credentials [index] as a proxy for distance. The treatment variable *white relative tier* is defined to be the tier attended by an individual with a given academic index score relative to the tier attended by the median white with the same index score . . . . I adopt the Ayres and Brooks suggestion of computing the White Tier (Index) as the median tier for each 20 point intervals [sic] of the academic index. . . . As expected, relative tier has no effect on the bar passage rates of whites, since almost all whites are matched. For blacks and the minority subgroup, all of the signs in the bar passage outcomes have the negative sign predicted by the mismatch hypothesis.[415]

Williams's prepublication draft was the only version of his analyses available to the signatories of the ESB *and the court* when the brief was filed, and the ESB specifically cites to this version of the article.[416] Sander's deceptive criticism of the ESB's assessment of Williams's analysis becomes apparent when one reads the initial amicus briefs filed by Sander in *Fisher I* to which the ESB is filed in response. Sander filed two versions of his amicus brief, an initial version (Oct. 19, 2011)[417] and a revised version (May 29, 2012),[418] and both briefs cite to the 2009 version of the Williams article that was criticized by the ESB.[419] The ESB was filed Aug. 13, 2012, and therefore

415. *Id.* at 18, 27–28 (footnote omitted) (citation omitted); *compare id.* at 27–28, *with* Williams, *supra* note 22, at 173–74 (describing Ayres and Brooks's relative tier analysis, but making no reference to his own relative tier analysis present in the prepublication draft that purportedly provided support for the mismatch hypothesis).

416. Empirical Scholars Brief in *Fisher I*, *supra* note 30, at 15.

417. Brief Amicus Curiae for Richard Sander & Stuart Taylor, Jr. in Support of Petitioner, Fisher v. Univ. of Tex. at Austin (*Fisher I*), 570 U.S. 297 (2013) (No. 11-345), 2011 WL 5015112 [hereinafter Sander, Brief in *Fisher I* (Original)].

418. Brief Amici Curiae for Richard Sander & Stuart Taylor, Jr. in Support of Neither Party, *Fisher I*, 570 U.S. 297 (No. 11-345), 2012 WL 1950266 [hereinafter Sander, Brief in *Fisher I* (Revised)].

419. *See* Sander, Brief in *Fisher I* (Original), *supra* note 417, at 8; Sander, Brief in *Fisher I* (Revised), *supra* note 419, at 9.

Sander did not—indeed, he could not—submit a revised brief challenging the assertions of the ESB with respect to its criticism of Williams's inappropriate interracial comparisons. In Sander's amicus brief in *Fisher II*, filed on Sept. 10, 2015—more than three years after the ESB was filed—Sander solely cites to the 2013 published version of Williams's article that no longer contained most of the problematic crossracial comparisons identified by the ESB.[420] Sander repeated this fabrication about the ESB's critique of Williams's analysis in a law review article published in 2014, asserting that the ESB's criticism of Williams's work with respect to crossracial comparisons is "*simply untrue*" and "*In every case, Williams's analyses compared students within the same racial group or groups!*"[421] Clearly Sander, rather than the ESB, deliberately engaged in misrepresentation about Williams's work that was before the Court in *Fisher I*.[422]

Despite Sander's apparent grasping at straws in the aftermath of the ESB,[423] one thing remains abundantly clear: The signatories of the ESB are a veritable who's who of quantitative social scientists,[424] and several of these signatories have either published rebuttals of Sander's work,[425] commented on drafts of critiques of Sander's work (and acknowledged in the footnotes),[426] or have engaged in research (or assessments of empirical research) on either the effect of attending elite law schools on post–law school outcomes (such as employment and earnings)[427] or the effect of affirmative action on minorities

420. Sander, Brief in *Fisher II*, *supra* note 40.
421. Sander, *Mismatch & the ESB*, *supra* note 30, at 568.
422. Even with the crossracial comparisons removed from the analysis, Williams's conclusions are unreliable because his models suffer from several other forms of bias and rely on multiple implausible assumptions highlighted by the ESB and others. *See supra* note 323 and accompanying text.
423. *See also supra* note 40 (describing Sander's *ad hominem* arguments against the signatories of the ESB).
424. In addition to King and Rubin, *see supra* note 29, the group of signatories includes multiple elected fellows to the Academy of Arts and Sciences, American Association for the Advancement of Science, and the American Statistical Association.
425. *See, e.g.*, Ho, *supra* note 7 (direct rebuttal of *Systemic Analysis* by Ho); Ho, Evaluating Affirmative Action, *supra* note 23 (same); Ayres & Brookes, *supra* note 12 (direct rebuttal of *Systemic Analysis* by Ayres & Brooks); Chambers et al., *supra* note 68 (direct rebuttal of *Systemic Analysis* by Chambers, Lempert, and colleagues); Alice Xiang & Donald B. Rubin, *Assessing the Potential Impact of a Nationwide Class-Based Affirmative Action System*, 30 STAT. SCI. 297 (2015) (direct rebuttal of *Systemic Analysis* by Alice Xiang & Donald Rubin).
426. *See* Ho, Evaluating Affirmative Action, *supra* note 23, at 1 (thanking ESB signatories Greiner and King for their helpful comments).
427. Paul Oyer & Scott Schaefer, *The Returns to Elite Degrees: The Case of American Lawyers*, 72 INDUS. LAB. REL. REV. 446, 457, 477 (2019) (reporting that (1) minorities are more likely to attend a top 10 law school, controlling for parental education, elite reputation of

outside of the law school context.[428] Of greater significance, however, is the fact that most of the signatories have been instrumental in developing the core conceptual and methodological toolkit necessary for credible causal inference across a wide range of substantive domains.[429] Most of the scholars who have directly engaged Sander's work—several of them signatories on the ESB—attempt to closely adhere to the core principles of causal inference when evaluating mismatch theory and their conclusions uniformly reject Sander's findings.[430] By contrast, the few studies purporting to find support for mismatch effects in law school, conducted by "Richard Sander and others, predominantly people tied to him in some way,"[431] violate widely accepted norms of causal inference.[432]

undergraduate institution, and undergraduate major; and (2) graduates from top 10 law schools have a greater likelihood of working at an elite firm and earning more money, even after controlling for LSAT score).

428. Lempert, *supra* note 320, at 158–65 (identifying several errors, as well as questionable reporting practices about the sensitivity of results to model specification, in research conducted by Arcidiacono on alleged mismatch effects among science majors in the University of California system).

429. *See, e.g.*, IMBENS & RUBIN, *supra* note 214; LITTLE & RUBIN, *supra* note 177; DONALD B. RUBIN, MATCHED SAMPLING FOR CAUSAL EFFECTS (2006); Blackwell et al., *supra* note 296; Berk, *supra* note 388; Adam N. Glynn & Kevin M. Quinn, *Why Process Matters for Causal Inference*, 19 POL. ANALYSIS 273 (2011); Greiner & Rubin, *supra* note 353; Ho et al., *supra* note 218; Ho & Rubin, *supra* note 218; King & Zeng, *supra* note 156; Rubin, *supra* note 156; Rubin, *supra* note 389.

430. It is puzzling that Sander continues to reject the ESB's criticisms of his and Williams's research on mismatch when it was Sander, himself, who testified before a federal agency tasked with, *inter alia*, collecting and studying information on racial discrimination. During his testimony, Sander urged the agency to assemble a panel of social scientists to review and verify his research on the impact of affirmative action at American law schools. *See* U.S. COMM'N ON CIVIL RIGHTS, AFFIRMATIVE ACTION IN LAW SCHOOLS 9, 35 (2007). Sander remarked: "The Commission should appoint a panel of social scientists to review the research available on the impact of racial preferences in American law schools on academic performance, bar passage, and long-term career success of African-American law students." *Id.* at 9.

431. Kidder & Lempert, *Mismatch Myth*, *supra* note 22, at 106.

432. *Id.* at 114 (noting that the proponents of the mismatch hypothesis "have mustered little evidence to support this hypothesis and the little they have is often methodologically suspect").

Arcidiacono, another economist tied to Sander, has published several papers alleging deleterious effects of race-conscious admissions policies, including mismatch effects. Peter Arcidiacono, Esteban M. Aucejo & V. Joseph Hotz, *University Differences in the Graduation of Minorities in STEM Fields: Evidence From California*, 106 AM. ECON. REV. 525 (2016). Arcidiacono's research has also been criticized for (a) violating fundamental principles of causal inference, (b) employing inconsistent methodological choices and selectively reporting results that are supportive of his hypotheses, and (c) offering unfounded attacks of other scholars' work that identify core flaws in his work. *See* Lempert, *supra* note 320, at 158–65 (identifying several errors in Arcidiacono's research, as well as questionable reporting practices about the sensitivity of results to model specification).

Similar to the ESB that highlighted inexcusable flaws in Sander's and Williams's work on mismatch in law schools, sixteen leading economists, including Nobel laureate Robert

IV. MINING FOR MISMATCH

Rubin famously remarked that "it is critical to give adequate conceptual thought to any nonstandard statistical problem before attacking it with mathematical analysis or available computer programs"[433] because "the blind use of complicated statistical procedures . . . is doomed to lead to absurd conclusions."[434] King has also cautioned that many reoccurring "problems [in quantitative social science] are more than technical flaws; they often represent important theoretical and conceptual misunderstandings."[435] Guillermina Jasso has attributed this state of affairs to the fact that "student[s] typically learn[] an array of methods for empirical analysis but virtually no methods for theoretical analysis. Thus, the temptation to premature test is strong. As a result, propositions are rushed to empirical test before there is sufficient understanding either of a proposition taken alone or of a coherent system of propositions."[436] Two key concerns of theoretical analysis are clarity and logical coherence. The clarity of a theory is its capacity to be understood and requires addressing all critical questions necessary for the theory to be intelligible.[437] Logical coherence refers to both a logical consistency between parts of the theory (in other words, no internal contradictions) and the absence of

Solow, filed an amicus brief identifying significant methodological shortcomings in Arcidiacono's statistical analyses submitted in litigation over Harvard's undergraduate admissions process. Brief of Professors of Economics as *Amici Curiae* in Support of Defendant at 1, Students for Fair Admissions, Inc. v. Presidents & Fellows of Harvard Coll., 397 F. Supp. 3d 126 (D. Mass. 2019) ("*Amici* include, among others, a Nobel laureate, four former Chief Economists of federal agencies, current and former university administrators, editors of peer-reviewed journals, and multiple professors whose research focuses on higher education. *Amici* have a wide range of views about the appropriateness of using race as a factor in college admissions."). Of the sixteen signatories, only one of them (Guido Imbens of Stanford University) was also a signatory on the ESB. According to these economists, Arcidiacono's analysis "did not provide the principled justification that sound statistical methodology requires," *id.* at 1, and Arcidiacono's and plaintiff's amici's "criticisms of [defendant's statistical expert's] modeling approach . . . are not based on sound statistical principles or practices." *Id.* at 18 (noting that Arcidiacono and plaintiff's amici "flatly ignore" alternative analyses and findings that undermine their claims).

433. Rubin, *supra* note 91, at 300.

434. Paul W. Holland & Donald B. Rubin, *On Lord's Paradox*, *in* PRINCIPALS OF MODERN PSYCHOLOGICAL MEASUREMENT 3, 18 (Howard Wainer & Samuel Messick eds., 1983).

435. King, *supra* note 3, at 666.

436. Jasso, *supra* note 5, at 11 (emphasizing the importance of conceptual clarity and logical coherence before conducting empirical analysis).

437. *See, e.g.*, ARTHUR L. STINCHCOMBE, CONSTRUCTING SOCIAL THEORIES 6 (1968) ("Social theorists should prefer to be wrong rather than misunderstood. Being misunderstood shows sloppy theoretical work.").

spurious reasoning.[438] The primary conceptual flaws present in Sander's analyses pertain to the posited theoretical relationships among the key variables forming the core of the mismatch hypothesis—relationships that numerous scholars have identified as dubious, and sometimes nonsensical.[439]

Sander often claims that his research findings have been replicated, but that is quite different from stating that his findings have been validated by the scientific community.[440] By replication, Sander simply means that researchers can follow his recipe and reproduce the same results, not that the recipe has been deemed correct.[441] The ESB explains, "The hallmark of reliable empirical work is that it can be validated by other researchers. A wide array of social scientists have studied the impact of elite educational institutions on student outcomes, reaching conclusions directly contrary to those of mismatch."[442] Likewise, Kidder and Lempert note that "[validation] attempts, except for the work of Sander's friend and erstwhile coauthor Doug Williams . . . , have served not to bolster the case for mismatch theory, but rather to call it into question."[443]

The primary and ubiquitous criticism of Sander's work is his heavy reliance on implausible assumptions.[444] Rather than putting his theory to

438. *See* Jasso, *supra* note 5, at 3.
439. *See supra* Part II.
440. *See* Kidder & Lempert, *Mismatch Myth*, *supra* note 22, at 124 n.17 (noting that Sander's assertions that his results have been replicated "is itself revealing, for it reflects a crabbed view of what it means to 'replicate' research, a view that would not be asserted by those who believed in the robustness of his findings. . . . This tells us nothing about the trustworthiness of Sander's results.").
441. *Compare* SANDER & TAYLOR, *supra* note 4, at 69 ("[I]t is important to note at the outset that all the factual claims and the data presented in 'Systemic Analysis' withstood all scrutiny. All of its tables, models, and analyses were replicated."), *with* Kidder & Lempert, *Mismatch Myth*, *supra* note 22, at 108 ("With the exception of Sander's erstwhile coauthor Doug Williams, every social scientist we know of who has independently analyzed the data Sander used has reported results that dispute his conclusions."), *and* Camilli & Welner, *supra* note 9, at 521 ("Some [mismatch] studies suggest positive effects, some suggest negative effects, and some suggest no significant effects. If enough snark hunters return empty handed, there is not much reason to examine or explain the nature of the snarks.").
442. Empirical Scholars Brief in *Fisher I*, *supra* note 30, at 14.
443. Kidder & Lempert, *Mismatch Myth*, *supra* note 22, at 112.
444. For example, certain models in "Systemic Analysis" that concluded ending affirmative action would increase the number of black attorneys (although decreasing black law school enrollment) almost certainly falsely assumed "that (1) the matriculating minority law school applicants, without exception, would have been willing to attend considerably less selective schools than those they had aspired to attend, (2) that they would have been indifferent to the location of those law schools that might have admitted them, (3) that they would have been similarly indifferent to the costs of attending such schools, and (4) that there would have been space in such schools for them." *See* KIDDER & LEMPERT, WHITE PAPER, *supra* note 22, at 4.

more exacting tests (which is the correct way to conduct rigorous social science),[445] Sander bridges data gaps with assumptions that consistently favor his theory and fails to conduct (or report) appropriate sensitivity analyses to examine the consequences of different assumptions and the robustness of his results. Denouncing this approach to empirical research as "data grubbing" and "fishing," econometrician Michael Lovell noted these problems are most pronounced "when the investigator's choice of explanatory variables is not inhibited by well-defined a priori considerations" and the investigator is "quite modest in describing how industrious a search was undertaken in generating reported results; and the criterion by which variables have been selected is usually left unspecified."[446] Some model specifications are reported and designated as "correct," but this occurs only after the researcher has looked at the estimates from dozens (possibly hundreds or even thousands) of models that all vary with respect to how the variables are coded, what variables are included, the parametric form of the relationships, and what subset of the data were analyzed. Rather than providing a true test of an *ex ante* hypothesis, this practice often amounts to a demonstration that it is possible to find results consistent with a preferred hypothesis.[447] According to pioneering statistician, John Tukey: "The most important maxim for data analysis to heed, and one which many statisticians seem to have shunned, is this: 'Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise.'"[448]

445. *See, e.g.*, Dan L. Burk, *When Scientists Act Like Lawyers: The Problem of Adversary Science*, 33 JURIMETRICS 363, 366–67 (1993) (arguing that the professional norms of the legal community revolve around adversarialism and advocacy, whereas the professional norms of the scientific community involve the creation of concepts and their exploration through the tools of empirical research in an attempt to construct a coherent view of the universe and its workings); *see also, e.g.*, Lee Epstein & Gary King, *The Rules of Inference*, 69 U. CHI. L. REV. 1, 9 (2002) ("While a Ph.D. is taught to subject his or her favored hypothesis to every conceivable test and data source, seeking out all possible evidence against his or her theory, an attorney is taught to amass all the evidence for his or her hypothesis and distract attention from anything that might be seen as contradictory information." (emphasis omitted)).
446. Michael C. Lovell, *Data Mining*, 65 REV. ECON. & STAT. 1, 10–11 (1983).
447. *See* Merton, *supra* note 5, at 468 (describing *ex post* explanations as unscientific).
448. John W. Tukey, *The Future of Data Analysis*, 33 ANNALS MATHEMATICAL STAT. 1, 13–14 (1962) (emphasis omitted).

A. Baiting and Switching Bar Passage

Sander has expressly stated that several of his critics' research "showed evidence of having been carefully massaged,"[449] yet in replying to the initial wave of criticisms of "Systemic Analysis," Sander switched from focusing on eventual bar passage to first-time bar passage.[450] Sander argues that the first-time bar passage rate is a better reflection of how much a student has learned in law school than the ultimate bar passage rate, but he fails to offer any evidence supporting his claim. It is difficult to imagine any reason for Sander's change of focus other than the fact that it yields results that appear supportive of mismatch theory—assuming, of course, one ignores the myriad of methodological flaws with his analyses. If the claim of mismatch theory is that affirmative action reduces the number of lawyers, then the focus on ultimate bar passage is the only appropriate outcome variable. Kidder and Lempert explain:

> Although those who know nothing about the bar exam and bar exam preparation might think Sander's argument is self-evidently correct, it is not. Not only may a bar exam cover subjects, including local procedural rules, that a student has never studied in law school, but in all states a major portion of the bar exam will relate to courses that a student completed more than two years before he or she takes the bar. For these reasons bar test preparation has become a major industry, and bar prep courses may cost a thousand dollars or more. Thus first time bar passage may reflect not just what has been learned in law school, but what courses a student took in law school, whether to save money a student chooses to attempt the bar once without taking a prep course, the quality of the bar prep course(s) a student attends, whether by chance a bar prep lecturer chooses to highlight an issue that later appears on the bar exam, and the amount of time a student can devote to boning up

449. SANDER & TAYLOR, *supra* note 4, at 86: *id.* at 83–86 (claiming that critiques of mismatch research conducted by Ayres, Brooks, Jesse Rothstein, and Albert Yoon excluded both data and model specifications that were likely supportive of mismatch).

450. *Compare* Sander, *Systemic Analysis*, *supra* note 7, at 444 tbl.6.1 ("The dependent variable is whether a person passes the bar on one of her first two attempts."), *with* Sander, *Mismeasuring*, *supra* note 24, at 2007 ("Figure 1 shows the results of . . . the influences of several variables on first-time bar passage for all blacks in the Bar Passage Study . . . data set."); *also compare* Ho, *Reply to Sander*, *supra* note 23, at 2013 (noting that Sander "has also changed the outcome variable from eventual bar passage, as defined in his original article, to first-time bar passage"), *with* Sander, *Replication of Mismatch*, *supra* note 4, at 86 ("In reproducing Ho's result, we found an odd anomaly: the outcome variable that Ho used was not whether students fail a bar exam, but whether students ever pass.").

> for the bar. For [these] reason[s] one hears stories of high ranking graduates of top law schools who failed the bar on their first try.[451]

Beyond the conceptual limitations of focusing on first-time bar passage, there are important methodological ones. First, not only is Sander's explanation of the hypothesized relationship between mismatch and first-time bar passage (but not ultimate bar passage) unconvincing, it is not altogether clear that he analyzes this relationship in the most reliable manner. A more illuminating and statistically defensible analysis would involve employing methods that permit the examination of the effect of mismatch on eventual bar passage while taking into account unobserved factors that jointly influence the likelihood of passing the bar on both the first time or a subsequent time (without positing a direct association between first-time passage and ultimate bar passage).[452] Sander hypothesizes that mismatch effects influence first-time bar passage, but not subsequent passage, and his models implicitly assume that the two events do not share common causes that are unrelated to mismatch.[453] Second, the fact that Sander allegedly finds support for mismatch theory only if an individual has one opportunity to take the bar, but not a second (or third) opportunity, undermines support for the claim of a mismatch effect for first-time takers. This is because as more mismatch-related tests for bar passage are conducted on a set of observations, the likelihood of erroneous inferences based solely on chance also increases. In other words, these repeated tests on the same data inflate the likelihood of a "false positive." Ho underscored this point in his reply to Sander: "[O]f course if one conducts enough tests, with enough specifications, measures, models, and recoding, one can induce a statistically significant result even if the relationship is random (classic 'Type 1 error')."[454]

---

451. KIDDER & LEMPERT, WHITE PAPER, *supra* note 22, at 8–9 n.12; *see also* James Bandler & Nathan Koppel, *Raising the Bar: Even Top Lawyers Fail California Exam*, WALL ST. J. (Dec. 5, 2005, 12:01 AM), https://www.wsj.com/articles/ SB113374619258513723 [https://perma.cc/73K4-R6EG] (noting that former Stanford Law School Dean and Harvard Law School Professor, Kathleen Sullivan, failed the California Bar, although she was already a nationally acclaimed constitutional law scholar and previously licensed to practice law in New York and Massachusetts).

452. *See* JEFFREY M. WOOLDRIDGE, ECONOMETRIC ANALYSIS OF CROSS SECTION AND PANEL DATA 144 (2002) (noting that multivariate regression models with correlated disturbances across equations capture unobserved common causes).

453. *See supra* Subpart II.A.2 (discussing nonrandom sample selection bias).

454. Ho, *Reply to Sander*, *supra* note 23, at 2014. Sander's inattention to the problem of bias resulting from multiple comparisons—that is, repeated comparison of mismatched and nonmismatched students across a range of outcomes—is present throughout much of his work on affirmative action. Nearly all of his analyses make multiple pairwise comparisons of coefficient effects across the various models (for example, first choice

B. Avoiding Reliability Assessments

Even assuming, *arguendo*, that the switch from ultimate bar passage to first-time bar passage is defensible, Ho's critique of Sander's empirical analysis underscores Lovell's aforementioned admonition with respect to "data grubbing" because Sander does not provide reliable evidence that his results are insensitive (robust) to his specific (and highly questionable) modeling assumptions. Specifically, the reliability of the causal estimates of a mismatch effect, as well as the overall predictive capacity of the model, can be readily assessed with two widely adopted procedures in the social sciences—model averaging and crossvalidation—that Sander and Williams neglect to employ. I provide a discussion of these procedures here, rather than in Part II, because whereas Part II is concerned with explanatory modeling, this Part focuses on predictive modeling. Ho and colleagues explain, "model averaging . . . and cross-validation . . . are useful for predictive inference but not directly applicable in the context of causal inference."[455] Model averaging "offers no solutions to the problems of endogeneity or causal inference . . . [and] is best used as a subsequent robustness check to show that our inferences are not overly sensitive to plausible variations in model specification."[456] Similarly, crossvalidation is primarily concerned with examining the performance of a model, as a whole, rather than the proper identification of a specific causal effect. It is probably most helpful, though not entirely accurate, to think of model averaging and crossvalidation as tools that can be helpful in identifying potential problems with a proposed causal model; a particular model, however, may perform reasonably well according to model averaging and

versus second choice; first choice versus "consolidated" second choice; second choice versus third choice). *See* Sander, *Replication of Mismatch*, *supra* note 4, at 79–82. The confidence intervals and *p*-values for these pairwise comparisons should be adjusted to account for multiple comparisons. For a single test, if we choose a 5 percent significance level, we would expect a 5 percent chance of concluding that two regression coefficients are different when the population values are actually equal (Type I error). When making multiple comparisons at the 5 percent level, the likelihood would be higher than 5 percent. The basic idea is to modify the threshold or significance level for the individual tests so that the overall significance level or Type I error rate is still 5 percent—that is, so that there is only a 5 percent chance or less of concluding there is a statistically significant result when in fact there is none. Widely accepted approaches to deal with analyses that make several comparisons across groups in order to minimize false positives have been available dating back to the 1930s. *See, e.g.*, E. L. LEHMANN & JOSEPH P. ROMANO, TESTING STATISTICAL HYPOTHESES 348–84 (3d ed. 2005) (describing multiple comparison tests popularized by Carlo E. Bonferroni, Sture Holm, Henry Scheffé, and Zbyněk Šidák).

455. Ho et al., *supra* note 218, at 202 (citations omitted).

456. Jacob M. Montgomery & Brendan Nyhan, *Bayesian Model Averaging: Theoretical Developments and Practical Applications*, 18 POL. ANALYSIS 245, 266 (2010).

crossvalidation, yet still exhibit many of the biases impacting causal inference. Below, I briefly describe each procedure and underscore their importance when testing the mismatch hypothesis.

1. Model Averaging

A favored practice in the social sciences, called "model averaging," involves examining the stability of the magnitude and direction of the causal effect across various combinations of the explanatory variables in the model, as well as different measurements of those explanatory variables and different plausible functional form (or parametric) assumptions, and report both weighted and unweighted average causal effects, with the weights being determined by predictive accuracy of the model (in other words, the model fit).[457] Model averaging is the best choice when the research is interested in a single causal estimate because it removes modeling uncertainty by averaging over the distribution of the estimate, leaving only the uncertainty caused by analyzing a sample and not the entire population of cases.[458] If the results can be nullified by small, sensible changes in model specification and measurement, then one cannot have much confidence in a true causal relationship.[459]

Neither Sander nor Williams report conducting any type of model averaging to examine the robustness of their causal estimates. This is unfortunate given the fact that model averaging is especially appropriate for testing the mismatch hypothesis due to the lack of a clear theory identifying, *a priori*, the core explanatory variables and the specific manner in which they are related to bar passage. Ho's reanalysis of Sander's work underscores the sensitivity of Sander's work to the choice of explanatory variables included in the model, and as a result, Sander's analyses cannot provide credible evidence of any mismatch effect.[460] Specifically, Ho examined 180 pre–law school enrollment covariates included the BPS data, compared to the mere ten

457. The weighted estimates are a helpful single measure that averages over the entire modeling distribution. The unweighted estimates provide insight into the distribution of the estimates that can be obtained from the data. *See* Cristobal Young & Katherine Holsteen, *Model Uncertainty and Robustness: A Computational Framework for Multimodel Analysis*, 46 SOCIO. METHODS & RSCH. 3, 4–5 (2017) (developing a framework to assess the robustness of the estimates from a model across sets of possible controls, variable definitions, standard errors, and functional forms).

458. *See* King et al., *supra* note 211, at 350; *see also supra* note 156.

459. *See* Edward E. Leamer, *Sensitivity Analyses Would Help*, 75 AM. ECON. REV. 308, 308 (1985) ("A fragile inference is not worth taking seriously. All scientific disciplines routinely subject their inferences to studies of fragility. Why should economics be different?").

460. *See* Ho, Evaluating Affirmative Action, *supra* note 23, at 7–8.

variables examined by Sander, and he found no empirical support for the mismatch hypothesis.[461]

2. Crossvalidation

Whereas model averaging entails "sampling" variables (and their hypothesized measurements and functional forms) included in the model, crossvalidation involves repeatedly drawing samples from a subset of the data and refitting the statistical model of interest on each subset to examine the extent to which the predictive capacity of the model differs across the subsets.[462] "Model validation is done to ascertain whether predicted values from the model are likely to accurately predict responses on future subjects or subjects not used to develop our model."[463] There are numerous approaches to crossvalidation, but perhaps the most popular is "*k*-fold" crossvalidation. Specifically, (1) the data are randomly split into *k* roughly equal-sized parts (called "folds"), (2) a model is fit on *k*-1 folds ("training folds"), and (3) the model is used to predict the outcomes on the remaining fold ("validation fold"). The process is repeated for each subset, so *k* models are estimated.[464] Simply put, "[*k*]-fold cross-validation uses part of the available data to fit the model, and a different part to test it."[465]

461. *See id.* at 7–9. Another approach to assessing the credibility of causal estimates, often referred to as a "causal bounds test," differs from the model uncertainty tests in that the focus is on the magnitude of the effect of the unobserved factors rather than the sensitivity of the causal estimates to different combinations of observed factors. In a nutshell, the procedure matches pairs of students across the various tiers, but that otherwise have the same observed characteristics. Under the assumption of unconfoundedness (no omitted variable bias), the students will have equal odds of being in either tier and the odds ratio will equal unity. The causal bounds algorithm manipulates the odds that the matched students have the same probability of being selected into either tier—for example, students with similar observable characteristics could differ in their odds of attending the top tier school as opposed to the second tier school by a factor of two. In other words, the manipulation of the odds is a measure of the degree of departure from the unconfoundedness assumption. *See* ROSENBAUM, *supra* note 201, at 105–57 (discussing approaches to assess a statistical model's sensitivity to hidden bias); *see also* Ho, *Reply to Sander*, *supra* note 23, at 2015 n.26 (suggesting that Sander "conduct sensitivity analyses to examine to what degree inferences would change if some unobserved variable were correlated to the treatment and the outcome").

462. *See* TREVOR HASTIE, ROBERT TIBSHIRANI & JEROME FRIEDMAN, THE ELEMENTS OF STATISTICAL LEARNING: DATA MINING, INFERENCE, AND PREDICTION 241 (2d ed. 2009).

463. HARRELL, *supra* note 261, at 109.

464. *See* BERK, *supra* note 220, at 34 ("Common practice seems to favor either fivefold or tenfold cross-validation.").

465. HASTIE ET AL., *supra* note 462, at 241.

Crossvalidation allows the analyst to determine, *inter alia*, how well a given statistical model can be expected to perform on independent data, thereby providing a measure of the quality of the chosen model. When the prediction error of the model varies widely across the validation folds, there is strong evidence of "overfitting" (poor internal validity) and one cannot be confident that the underlying model appropriately captures the dynamics of interest.[466] Perhaps the most intuitive measure of the model fit for crossvalidation is the R^2 which approximately captures the fraction of the variance in the dependent variable explained by the model.[467] The R^2 can also be averaged across the *k*-folds to obtain a more reasonable estimate of the overall model fit (as well as the variability of the model fit using the standard error of the R^2).[468] Stuart Russell and Peter Norvig explain:

> Despite the best efforts of statistical methodologists, users frequently invalidate their results by inadvertently peeking at the [validation] data. . . .
>
> Peeking is a consequence of using [data] to both *choose* a hypothesis and *evaluate* it. The way to avoid this is to really hold the [validation] set out—lock it away until you are completely done with learning and simply wish to obtain an independent evaluation of the final hypothesis.[469]

Neither Sander nor Williams report any results from a crossvalidation analysis, which is troubling because (1) the BPS data are well-suited for this type of analysis given the large number of observations relative to the number of variables included in the model,[470] (2) crossvalidation is easy to implement in all of the popular statistical software packages,[471] and (3) other scholars have highlighted the dependency of Sander's and Williams's results on unreasonable model assumptions about the structure of the BPS data.[472] The sensitivity of Sander's and Williams's results to these assumptions could have been examined with crossvalidation. My reanalysis of Sander's and Williams's results using *k*-fold crossvalidation reveals that the predictive

466. *See* HARRELL, *supra* note 261, at 111–14.
467. *See id.* at 111; *see also supra* Subpart II.A.2 (discussing R^2 measure).
468. *See* HARRELL, *supra* note 261, at 113.
469. STUART J. RUSSELL & PETER NORVIG, ARTIFICIAL INTELLIGENCE: A MODERN APPROACH 708–09 (3d ed. 2010) (emphasis omitted).
470. *See* BERK, *supra* note 220, at 34 (explaining that "cross-validation is more likely to provide generalizable results from training data with a large number of observations").
471. For example, crossvalidation procedures are readily available in *R*, *Matlab*, *SAS*, *S-Plus*, *SPSS*, and *Stata*.
472. *See* Ho, *supra* note 7; *see also* Rothstein & Yoon, *supra* note 21; Camilli & Jackson, *supra* note 183.

capacity of their models is highly unstable across the validation subsamples.[473] This finding is unsurprising given the aforementioned concerns about the impact of interpolation bias and extrapolation bias on their conclusions. Recall from Subparts II.D and II.E that interpolation and extrapolation biases produce model-based counterfactuals that are insufficiently tethered to the data, and as a result, the model must "guess" data values based on the parametric assumptions of the model. Crossvalidation highlights the sensitivity of results to unwarranted (and unverified) model assumptions because small changes in the data (via random subsetting) cause large changes in the predictive power of the model. King and Zeng explain: "We merely need to recognize that some questions cannot be answered reliably from some data sets. Our linearity (or other functional form assumptions) are written globally—for any value of x—but in fact are relevant only locally—in or near our observed data."[474]

C. Accepting the Bad and Rejecting the Good

"Together, the twin flaws of accepting poor evidence and rejecting good evidence produce some odd results."[475] Sander has claimed that "it is now *undeniable* that this system [of race-based affirmative action] is producing grossly unequal results. . . . Criticism is vital, but critics who wish to reject the mismatch theory outright have a responsibility to offer their own explanation and cures for the disparate harm our current system inflicts on blacks."[476] He is correct about one thing: Scholars must carefully examine the causes and correlates of interracial differences in education and employment outcomes. And, fortunately, they have. In fact, in addition to identifying numerous vital errors in Sander's research (rather than outrightly rejecting it), his critics have

473. I conducted a five-fold crossvalidation analyses of Sander's and Williams's models. Sander has published many tests of mismatch with very different model specifications, so I focus on his most recent published work, Sander, *Replication of Mismatch*, *supra* note 4. For both Sander's and Williams's work, I uncovered widely ranging R^2 statistics for the subsamples. Because crossvalidation is based on random subsamples of the data, each crossvalidation produced different R^2 statistics, but the range was generally from 0.002 to .09 from Sander, and 0.04 to 0.30 for Williams.

474. King & Zeng, *supra* note 230, at 187–88.

475. Mark Cooney, *Still Paying the Price of Heterodoxy:* The Behavior of Law *a Quarter-Century On*, 31 CONTEMP. SOCIO. 658, 660 (2002) (book review).

476. Sander, *Mismeasuring*, *supra* note 24, at 2010 (emphasis added); *see also* SANDER & TAYLOR, *supra* note 4, at 86 ("Not one of the critics [of mismatch] has articulated—let alone tested—an alternative theory to explain the patterns that are plainly there."); Sander, *Replication of Mismatch*, *supra* note 4, at 77 ("Through the lengthy debate on mismatch, no critic has ever provided an alternate explanation of these facts.").

routinely identified other explanations that are on stronger conceptual and evidentiary footing than mismatch theory.[477] Indeed, "[c]ompeting theories are then still to be judged by the extent to which they account for events, those which have been least falsified but which are most falsifiable."[478] Ayres and Brooks, for example, emphasize that the plausibility of these alternative explanations—which identify racial discrimination and isolation—is underscored by the fact that race/ethnicity is statistically significant in predicting law school grades, even after adjusting for UGPA and LSAT.[479] My own analysis, which exactly matched black and white students in the BPS based on gender, UGPA, and LSAT in the top tier,[480] revealed similar results with respect to FYGPA and LGPA: Race/ethnicity remained statistically significant in the models *and* accounted for a substantial increase in the explained variation in grades—approximately 17 percentage points (a 136 percent increase).[481] The results from the next elite tier were nearly identical

477. Chambers et al., *supra* note 68, at 1885–86 (discussing stereotype threat, financial circumstances, and the scarcity of black faculty as important determinants of academic and early professional success among black graduates); *see also* Kidder & Lempert, *Mismatch Myth*, *supra* note 22, at 111–12 (same); Ayres & Brooks, *supra* note 12, at 1813, 1829, 1838–40 (same); Dauber, *supra* note 238, at 1903–04 n.34 (identifying a vast empirical literature documenting persistent racial discrimination in the labor market for black lawyers and other high-status black professionals); David B. Wilkins, *A Systematic Response to Systemic Disadvantage: A Response to Sander*, 57 STAN. L. REV. 1915, 1919 (2005) (summarizing research on the persistence of racial discrimination in the legal field and arguing that "[a]ffirmative action has played a crucial role in helping black lawyers to overcome the systematic and persistent obstacles" and its benefits far outweigh its potential risks); *infra* Subpart IV.C (highlighting the importance of racial and gender context for educational outcomes of minorities and women).

478. MARVIN HARRIS, CULTURAL MATERIALISM: THE STRUGGLE FOR A SCIENCE OF CULTURE 18 (1979).

479. *See* Ayres & Brooks, *supra* note 12, at 1834–38. This finding mirrors research on the relationship between race/ethnicity and college performance. Black students consistently underperform white students with the same admissions credentials at the same colleges. Rothstein, *supra* note 369, at 311–13 (discovering that racial minorities earned lower grade point averages in college when controlling for high school grade point average and SAT scores).

480. Recall that exact matching eliminates model dependency by limiting the analysis to counterfactuals that are actually observed in the data, and the matching procedure I employed was officially endorsed by the Office of Biostatistics & Epidemiology for the FDA. I reanalyzed the models and, in addition to exactly matching students on gender, UGPA, LSAT, and tier location, I also exactly matched students based on whether they attended their first choice school. Although this procedure significantly reduced the number of comparables, the results were very similar to the models that did not include exact matching on the first/second choice variable.

481. For FYGPA, including race in the model increased amount of explained variation (R^2) from 12.5 percent to 29.5 percent, and for LGPA, from 10.3 percent to 27.1 percent.

as well.[482] Equally illuminating and troubling is the magnitude of the race effect. Black students' grades were nearly an entire standard deviation lower than their white counterparts with the exact same entering credentials and attending a law school in the top tier.[483] This effect size was virtually identical when examining students in the next highest tier.[484] It is inadvisable to posit race/ethnicity as a causal variable;[485] nonetheless, my analysis, as well as Ayres and Brooks's results, potentially shed light on the question of the magnitude of the "black tax" paid by black law students that is unrelated to academic mismatch.[486] By focusing on the top two tiers, my analysis has the advantage of examining a small, and relatively homogenous, group of law schools.[487] The top tier is composed of sixteen schools and the second tier is composed of

482. The $R^{2}{}_{2}$ increased from 6.3 percent to 22.1 percent for FYGPA, and from 8.4 percent to 26.4 percent for LGPA.

483. *See* WIGHTMAN, BEYOND FYA, *supra* note 377, at 31–32 (analyzing the BPS data and reporting similar results).

484. The magnitude of the race effect diminishes to approximately 70 percent of a standard deviation when analyzing the next two tiers; nonetheless, it remains both substantively meaningful and statistically significant. The predictive power of race/ethnicity for the models analyzing the middle two tiers also substantially diminishes.

485. *See supra* note 353 (explaining that an immutable characteristic, such a race/ethnicity, is not a causal variable because it is determined at a person's birth and cannot be changed by intervention).

486. *See* JODY DAVID ARMOUR, NEGROPHOBIA AND REASONABLE RACISM: THE HIDDEN COSTS OF BEING BLACK IN AMERICA 13–14 (1997) ("The Black Tax is the price Black people pay in their encounters with Whites (and some Blacks) because of Black stereotypes. . . . [It] is not rooted in conscious [racial] animus. . . . It is unconscious discrimination on the one hand and ostensibly rational discrimination on the other that impose the lion's share of the Black Tax today [and it] comes in many varieties."); *see also* ELLIS COSE, THE RAGE OF A PRIVILEGED CLASS 5 (1993) (describing the cumulative impact of constant experiences of anti-black racism and the energy expended in dealing with them as "soul-destroying" and "at the heart of black middle-class discontent").

Chetty and colleagues note that wealth does not serve as a protective factor for black people in the way it does for white people. *See* Raj Chetty, Nathaniel Hendren, Maggie R. Jones & Sonya R. Porter, *Race and Economic Opportunity in the United States: An Intergenerational Perspective*, 135 Q. J. ECON. 711 (2020). For example, black people raised in households in the top 3 percent of the income distribution (97th percentile) have marriage rates similar to white people raised in households at the very bottom of the income distribution. *Id.* at 738. Black males raised in households in the top 1 percent of the income distribution (99th percentile) have identical incarceration rates to white males raised in households at the 34th percentile of the income distribution. *Id.* at 744–46.

487. Due to the small number of exact matches when analyzing the top two tiers separately, I reanalyzed the data after combining students from these two tiers, and exactly matching black and white students according to gender, UGPA, and LSAT. Unsurprisingly the results were nearly identical to the disaggregated analysis on each tier given the fact that the predictive power of race and the magnitude of the race effect were basically the same across the tiers.

fourteen schools,[488] so black and white students in these tiers with identical entering credentials will be similarly positioned relative to the median student at their school, academically speaking. Nevertheless, black students' academic performance lags substantially behind their academically similarly situated white classmates in both tiers. In the words of Ayres and Brooks, "However, there is more to 'race' than the race of the student; and . . . one must wonder if better race controls (including the racial composition of the school, the classroom and the faculty) would reveal an even bigger impact of race, as implied by the stereotype threat literature."[489]

Sander's own work provides indirect support for the proposition that law school racial context better explains the racial differential in law school grades than academic mismatch. Specifically, Sander examines the differences in FYGPA for black students across all of the law school tier pairings, exactly matching the students on gender, UGPA, and LSAT. The grade differentials are rather modest for the adjacent tier pairings (and also contradictory to the mismatch hypothesis), save for Tier 5/Tier 6.[490] Black students in Tier 5, on

488. *See supra* Table 2. Table 2 lists the top tier as comprising eighteen schools and the second highest tier comprises nineteen schools; two top tier schools and five second tier schools did not participate in the BPS study. WIGHTMAN, USER'S GUIDE, *supra* note 324, at 16 (describing the composition of the law school clusters in the BPS study).

489. Ayres & Brooks, *supra* note 12, at 1829. I also analyzed race effects for the Latinx and Asian American students, after exactly matching them, relative to white students, on gender, UGPA, LSAT, and tier location. Both groups performed worse, academically, relative to white students. The magnitudes of the race effects for Asian students and Latinx students were much smaller than those for black students (ranging from a quarter to a half standard deviation lower compared to white students for FYGPA and LGPA). *See also infra* Appendix B.3 (demonstrating that race/ethnicity exerts a statistically significant total effect on first-time bar passage, even after controlling for LGPA, but the strongest negative effect was found for black students).

Camille Charles and colleagues discover no support for mismatch effects at the student-level (and some support for a reverse mismatch effect), but do they uncover evidence that "affirmative action functions at the institutional level to undermine grades earned by black and Latino students." CHARLES ET AL., *supra* note 8, at 200. When there is a large test score gap at the aggregate level, black and Latinx achievement is directly lowered through a stigmatizing social context and indirectly lowered via subjective performance burdens (stereotype threat). *Id.*

490. Sander, *Replication of Mismatch*, *supra* note 4, at 87 tbl.14. Sander places the tiers in reverse order, so the top tier is labeled as Tier 6. I discuss the tiers in terms of their more intuitive ordering, so the top tier is Tier 1 and the last tier (the HBLS tier) is Tier 6.

Sander's analysis focusing solely on black students also reveals that, when making adjacent tier comparisons, students in the more competitive tier perform *better* in terms of FYGPA. Although the differences are modest, this finding directly contradicts Sander's assertion that these students are outmatched. Specifically, black students in Tier 2 performed better than black students in Tier 3, and black students in Tier 3 performed better than black students in Tier 4, even though these students were exactly matched in terms of UGPA, LSAT, and gender. In two of the other adjacent comparisons—Tier

average, have a FYGPA, that is nearly an entire standard deviation (*beta* = -0.97, $p < .001$) lower than otherwise similarly situated black students in Tier 6. Tier 6, as noted previously, consists entirely of HBLS. By comparison, the FYGPA differential for Tier 4/Tier 5 is one-fifth as a large (-0.19) and not statistically significant.[491] Although Tier 6 has the lowest average UGPA and LSAT, this cannot explain the magnitude of the difference in FYGPA—especially considering the fact that UGPA and LSAT account for only a small percentage of the variation in FYGPA (and LGPA). Furthermore, the difference in the average LSAT score for the Tier 4/Tier 5 (3.22) and Tier 5/Tier 6 (3.04) comparisons are nearly identical.[492] The differences in the average academic index scores for Tier 4/Tier 5 (56.2) and Tier 5/Tier 6 (73.33) are similar as well.[493]

These concerns about the impact of racial dynamics on legal education and outcomes parallel earlier concerns about gender dynamics in law schools.[494] Lani Guinier and colleagues' groundbreaking study explored the reasons female law students at the University of Pennsylvania underperformed their male counterparts in terms of LGPA despite having similar entering credentials.[495] The centerpiece of their study was performance data (UGPA, LSAT, and grades for each year of law school) for 981 law students. Their data were also supplemented by, *inter alia*, interviews with students and faculty, as well as classroom observations.[496] They note that:

> [The] performance differential between men and women [was] created in the first year of law school and maintained over the next

1/Tier 2 and Tier 4/Tier 5—there is no statistically significant difference in FYGPA. *Id.* I was unable to replicate Sander's finding for the Tier 3/Tier 4 comparison. Contrary to Sander, I found no statistically significant difference between students in these two adjacent tiers.

When the same analysis is conducted on the subsample of white students, FYGPA is consistently higher in the lower–adjacent tier (and statistically significant for all adjacent pairings). This finding also undermines the mismatch hypothesis because, according to Sander, black students receive racial preferences that place them at a disadvantage in law school, but white students do not. *See id.* at 85 (arguing that "there is no reason to think that whites, taken as a whole, will experience mismatch, since at any given school their credentials are generally slightly above the school medians").

491. *Id.* at 87 tbl.14.

492. *See supra* Table 2.

493. The average index scores for Tiers 4, 5, and 6 are, respectively, 681.46, 625.28, and 551.95. *See supra* Table 2.

494. *See, e.g.*, Dara E. Purvis, *Female Law Students, Gendered Self-Evaluation, and the Promise of Positive Psychology*, 2012 MICH. ST. L. REV. 1693, 1694–1703 (summarizing the research literature on women's experiences in law school).

495. LANI GUINIER, MICHELLE FINE & JANE BALIN, BECOMING GENTLEMEN: WOMEN, LAW SCHOOL, AND INSTITUTIONAL CHANGE (2d ed. 1997).

496. *Id.* at 30–31.

> three years. By the end of their first year in law school, men [were] three times more likely than women to be in the top 10% of their law school class.[497]

Guinier and colleagues underscore that their findings at the University of Pennsylvania are:

> [S]upported by those of the 1996 Wightman Report [based on the BPS data], whose findings 'constitute a consistent and cumulative set of data supporting the hypothesis that women tend to underperform academically in law school relative to their previous academic achievement.' . . . [The BPS data] in no way suggest that women are *unable* to perform adequately in law school or that the performance differential between women and men is a consequence of women selecting less rigorous undergraduate majors.[498]

More recent studies conducted at Harvard Law School, Stanford Law School, and the University of Texas Law School uncovered a similar dynamic with respect to gender and first-year grades.[499] The primary explanations for this performance gap—which preclude mismatch effects because the female and male students share similar entering credentials—were (1) alienation via the Socratic method, (2) perceptions of a hostile law school environment, and (3) the lack of gender diversity among the law faculty.[500]

---

497. *Id.* at 28 (emphasis omitted).

498. *Id.* at 105 n.13 (quoting LINDA F. WIGHTMAN, LAW SCH. ADMISSIONS COUNCIL, WOMEN IN LEGAL EDUCATION: A COMPARISON OF THE LAW SCHOOL PERFORMANCE AND LAW SCHOOL EXPERIENCES OF WOMEN AND MEN 26 (1996)).

499. *See* DAVID B. WILKINS, BRYON FONG & RONIT DINOVITZER, HARVARD CTR. ON THE LEGAL PROFESSION, THE WOMEN AND MEN OF HARVARD LAW SCHOOL: PRELIMINARY RESULTS FROM THE HLS CAREER STUDY (2015) (analyzing grades from cohorts from 1975, 1985, 1995, and 2000); *see also* HARVARD LAW SCH. WORKING GRP. ON STUDENT EXPERIENCES, STUDY ON WOMEN'S EXPERIENCES AT HARVARD LAW SCHOOL (2004) (analyzing seven years of data, 1997–2003); Daniel E. Ho & Mark G. Kelman, *Does Class Size Affect the Gender Gap? A Natural Experiment in Law*, 43 J. LEGAL STUD. 291, 299 (2014) (analyzing grades from 2001 to 2012); Allison L. Bowers, *Women at The University of Texas School of Law: A Call for Action*, 9 TEX. J. WOMEN & L. 117, 134 (2000) (examining thirteen years of data, 1984–1997).

500. *See generally* Purvis, *supra* note 494; Ho & Kelman, *supra* note 499, at 308–10 (discovering that small class sizes eliminate, and sometimes reverse the gender gap in law school grades, and that may be due to the fact that instruction is more interactive and less Socratic in smaller classes); *see also* Bostwick & Weinberg, *supra* note 320, at 9 (noting that increasing the number of women in Science, Technology, Engineering, and Math (STEM) doctoral programs had a positive impact on the overall graduation rates of women in those programs, and concluding that having more women in these STEM programs most directly impacted whether women dropped out of the doctoral programs within the first year).

With respect to racial/ethnic diversity among the professorate, underrepresented minorities, including African Americans, Latinx, Native Americans, and Alaskan

The fact that black law students who attend historically black law schools (HBLS) are more likely to graduate and pass the bar exam than their tier ranking would predict, meanwhile white students who attend HBLS are less likely to pass the bar than tier ranking would predict, strongly suggests that nonmismatch-related environmental factors are also operative.[501] But Sander is outright dismissive of these facts because in his mind black underperformance on the bar exam "turns out to be a fundamental question to which no plausible answer other than mismatch theory has ever been suggested."[502] Sander unequivocally states:

> Blacks were much less likely to pass the bar exam than were whites with the same academic index coming out of college.
>
> So clearly something was depressing black performance.
>
> . . . .
>
> . . . Blacks were doing badly on the bar not . . . because law schools were somehow unwelcoming. They were doing badly, it turns out, because the law schools were killing them with kindness by extending admissions preferences (and often scholarships to boot) that systematically catapulted blacks into schools where they were very likely not only to get bad grades but also actually have trouble learning. It was thus about race only to the extent that schools based large admissions preferences on that factor.503

Natives, composed 10.2 percent of fulltime tenured faculty positions in U.S. universities in 2013. MARTIN J. FINKELSTEIN, VALERIE MARTIN CONLEY & JACK H. SCHUSTER, TIAA INST., TAKING THE MEASURE OF FACULTY DIVERSITY 9 tbl.5 (2016). The proportion of faculty members who are African American and were hired between 2007 and 2016 fell from 7 percent to 6.6 percent. Krupnick, *supra* note 320.

501. Ho, *supra* note 7, at 2004 ("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 . . . . This result could be an indication that white students actually do better in homogenous environments . . . . "); *see also supra* note 320 (showing that a critical mass of minority students at an institution likely explains differential academic performance among science majors rather than school selectivity).

502. SANDER & TAYLOR, *supra* note 4, at 58–59.

503. *Id.* at 58, 60–61 (emphasis omitted). *But cf.* Daniel J. Morrissey, *Saving Legal Education*, 56 J. LEGAL EDUC. 254, 269 (2006) (noting that poor performing law schools target black law students and charge them full tuition, in part, to subsidize better credentialed (and mostly white) classmates); Elie Mystal, Opinion, *How to Con Black Law Students: A Case Study*, N.Y. TIMES (Mar. 20, 2017), https://www.nytimes.com/2017/03/20/opinion/how-to-con-black-law-students-a-case-study.html [https://perma.cc/6HGT-Q85P] (same); Genevieve Bonadies, Joshua Rovenger, Eileen Connor, Brenda Shum & Toby Merrill, *For-Profit Schools' Predatory Practices and Students of Color: A Mission to Enroll Rather Than Educate*, HARV. L. REV. BLOG (July 30, 2018), https://blog.harvardlawreview.org/for-profit-

But Sander's responses to his critics embody the low standards of statistical criticism in academia lamented by renowned biostatistician, Irwin Bross, nearly six decades ago: "The critic has the responsibility for showing that his [sic] counterhypothesis is tenable. In so doing, he [sic] operates under the same ground rules as a proponent."[504] Several of the widely used "put downs" of statistical work are operative in Sander's responses to his critics: (a) the hit-and-run, which points to a flaw without developing a counterhypothesis; (b) the speculative, which proposes a counterhypothesis but makes no attempt to reconcile it with extant evidence; and (c) the tubular (tunnel vision), which fails to see evidence contrary to the favored hypothesis.[505] Sander shoulders the responsibility of proving the tenability of the mismatch hypothesis in light of both his failure to find support for the mismatch hypothesis when adhering to widely accepted rules of causal inference *and* the burgeoning and methodologically sound research on the positive effects of race-conscious affirmative action on the educational and professional attainment of its intended beneficiaries across of a wide range of settings.[506] Sander would do well to heed the prudent advice appearing sixty-five years ago in Darrell Huff's seminal text, *How to Lie with Statistics*:

> [W]hen there are many reasonable explanations you are hardly entitled to pick one that suits your taste and insist on it. But many people do.
>
> To avoid falling for the *post hoc* fallacy and thus wind up believing many things that are not so, you need to put any statement of relationship through a sharp inspection.[507]

schools-predatory-practices-and-students-of-color-a-mission-to-enroll-rather-than-educate [https://perma.cc/E8WZ-S2QE] (same).

504. Irwin D. J. Bross, *Statistical Criticism*, 13 CANCER 394, 394 (1960) (emphasis omitted).

505. *See id.* at 395–98.

506. *See, e.g.*, Empirical Scholars Brief in *Fisher I*, *supra* note 30 (noting that Sander's work violates many of the fundamental tenants of causal inference); Richard O. Lempert, David L. Chambers & Terry K. Adams, *Michigan's Minority Graduates in Practice: The River Runs Through Law School*, 25 LAW & SOC. INQUIRY 395, 395 (2000) (failing to find any support for the mismatch hypothesis at the University of Michigan Law School); *see also supra* note 379 (listing multiple national studies finding that race-conscious affirmative action improves the likelihood of graduation among African American students); *supra* note 196 (citing a wide variety of studies that find race-conscious affirmative action does not adversely impact employee performance and compensation); Consolidated Brief of Lt. Gen. Julius W. Becton, Jr. et al. as Amici Curiae in Support of Respondents at 5, Grutter v. Bollinger, 539 U.S. 306 (2003) (Nos. 02-241 & 02-516) (explaining that "the military cannot achieve an officer corps that is *both* highly qualified *and* racially diverse unless the service academies and the ROTC use limited race-conscious recruiting and admissions policies").

507. DARRELL HUFF, HOW TO LIE WITH STATISTICS 89 (1954).

V. PROPERLY MEASURING MISMATCH

What would a scientifically defensible examination of the mismatch hypothesis with the BPS data entail? It is important to acknowledge from the outset that no approach is infallible because of the "fundamental problem of causal inference": We can only observe one of the potential outcomes (according to treatment status) for each unit, so causal inference is justified only to the extent that one can plausibly claim that (1) the only important factor that distinguishes the treatment and control groups in the analysis is, in fact, exposure to the treatment and (2) the data contain sufficient information on both treated and nontreated groups to draw empirically-driven, as opposed to model-driven, inferences. Nevertheless, the careful analyst can attempt to ameliorate the various types of biases I describe in this Article that threaten internal validity.[508] The ensuing advice is suggestive, nonexhaustive, and merely provides a rough template for investigating the role of mismatch in law schools using the BPS data. I conclude this Part by highlighting generalizability concerns with the BPS data that may undermine external validity. These data were initially collected in 1991, so to the extent that present-day law school matriculants—especially minorities—and law school environments significantly differ from those examined by the BPS, the implications from any study based on the BPS may have limited relevance for today's current and aspiring law students.

A. Improving Internal Validity

1. Posttreatment Bias

Recall that posttreatment bias can exist in two forms: controlling for a mediating variable (such as LGPA) and analyzing a nonrandom subset of students (such as law graduates) when analyzing bar passage. Sander's analyses have suffered from both types of problems and his attempts to address these biases are statistically indefensible and fail to remove (or even significantly ameliorate) the bias.[509] A proper analysis of mismatch in the BPS data would employ widely available methods that appropriately address both types of bias.[510] Analysts should also be explicit about the limitations of the

508. *See, e.g.*, Leamer, *supra* note 459, at 308 ("We must insist that all empirical studies offer convincing evidence of inferential sturdiness.").
509. *See* Sander, *Replication of Mismatch*, *supra* note 4.
510. *See supra* Subparts II.A.1–II.A.2.

mediation analysis (through LGPA) that I have previously discussed. Sander does not take these steps.

2. Nonresponse Bias

At best, discarding information from respondents because of "nonresponse" in the BPS, as Sander does, will merely make the estimates less precise; at worst, these estimates will be biased and undermine analysts' ability to draw reliable inferences from the data. I have shown that missing values for one of Sander's key explanatory variables, LGPA, can be systemically predicted by several of the other variables in his models, thereby rejecting the assumption of Sander's models that the remaining cases in his analyses are a random subset of all of the cases in the overall sample.[511] Quantitative social scientists have employed methods to properly address nonresponse bias since the mid-1970s.[512] Such approaches are commonplace in empirical analyses of survey data that populate the pages of the most reputable peer-reviewed social science journals.[513] Sander employs no such methods.

3. Omitted Variable Bias

Several of the biases plaguing Sander's mismatch work that have been described in this Article can be reconceptualized as omitted variable bias.[514] Although one can seldom be certain that there are no relevant variables omitted from the analysis, it is advisable to take into account as many pretreatment variables as possible to achieve balance between the treatment and control groups. Ho identified nearly two hundred additional variables in the BPS data. Further analyses of the BPS data must incorporate this information if the "no confounder bias" assumption is to be believed. Any analysis using significantly less than the full complement of plausible pretreatment variables in the BPS should test and report the sensitivity of the results to changes in the variables included in the model. Further, analysts

511. *See supra* Subpart II.B.
512. *See, e.g.*, Donald B. Rubin, *Inference and Missing Data*, 63 BIOMETRIKA 581 (1976); *see also supra* Subpart II.B.
513. *See* Donald B. Rubin, *Multiple Imputation After 18+ Years*, 91 J. AM. STAT. ASS'N 473, 486 (1996) ("Multiple imputation is doing well, perhaps even flourishing, as documented by recent sessions at the annual meetings of the American Statistical Association and other professional associations . . . and by the variety of recent publications documenting its applicability and extending its theory.").
514. *See supra* Subpart II.C ("Posttreatment bias and nonresponse bias are special cases of omitted variable bias . . . . ").

should pay close attention to the predictive ability of the model (model fit) and not merely the statistical significance of parameter estimates. With a dataset as large as the BPS, statistically significant effects tell us very little about the statistical model's ability to accurately describe the process driving LGPA, graduation, and bar passage.[515] It is often the case that when the model does a poor job of explaining the phenomenon under investigation, the model is improperly specified, which often indicates an exclusion of important explanatory variables. Even if the parameter estimates of key variables were to remain statistically significant after including other factors that increase the predictive ability of the model, the relative magnitudes of the variables are likely more important to understand what factor(s) are primarily responsible for a particular phenomenon. Statistical significance relates to our level of confidence that a hypothesized effect exists in the larger population, but practical significance concerns the magnitude of an effect, and therefore its relative importance in explaining the outcome under investigation.

4. Interpolation Bias

Various parametric assumptions are necessary to estimate and interpret a wide class of statistical models, but such assumptions should be warranted and, to the extent possible, specifically examined. Some of these parametric assumptions can be avoided altogether via semiparametric and nonparametric methods.[516] Generally speaking, a fully parametric model is preferable to a semiparametric or nonparametric model only when the parametric assumptions are justified. At minimum, the analyst must ensure that the functional form of the relationships between variables in the model and the outcome is consistent across treatment and control groups, and the conditional probability distribution for the

515. *See supra* Subpart II.C (noting that statistical significance does not measure the strength of an association between variables, and may simply be the result of a small correlation and a large dataset). According to statistician Adrian Rafferty:

> In the past 15 years, however, some quantitative sociologists have been attaching less importance to *P*-values because of practical difficulties and counterintuitive results.
>
> These difficulties are most apparent with large samples, where *P*-values tend to indicate rejection of the null hypothesis even when the null model seems reasonable theoretically and inspection of the data fails to reveal any striking discrepancies with it.

Adrian E. Raftery, *Bayesian Model Selection in Social Research*, 25 SOCIO. METHODOLOGY 111, 112 (1995).

516. *See supra* note 226 and accompanying text.

outcome variable is properly specified.[517] A class of statistical models that enable the analyst to simultaneously relax functional form (parametric) assumptions and properly account for posttreatment bias due to censoring may be particularly well-suited for analyzing the BPS data.[518] Sander does not evaluate the sensitivity of his conclusions to the parametric assumptions he adopts about the relationships between variables with values in the range of observed data.

5. Extrapolation Bias

The examination of mismatch effects must be based on information that actually exists (or could possibly exist) in the range of observed data. "What if" questions based on hypothetical comparisons that cannot be located within the observed range are not empirically supported counterfactuals—they are "extreme counterfactuals"—and cannot be answered with the available data. Analysts should not offer conjectures masquerading as science. Most parametric statistical methods permit analysts to assume data exist when they do not, and fail to provide any indication that the counterfactual lacks any evidentiary anchoring. But the results are extremely model-dependent because there is no true counterfactual case in the data, so analysts must resist the temptation to pose fanciful "what if" questions without first ensuring that these questions can be reasonably answered with the available data. Data-driven, as opposed to model-driven, counterfactuals often require careful preprocessing before fitting a statistical model to the data in order to fully exploit the available information.[519] For the past forty years, researchers have developed an impressive suite of techniques to facilitate this preprocessing, but Sander continues to avoid incorporating these approaches into his work.[520] My reanalysis of one of Sander's primary models that is correctly

517. *See supra* Subparts II.D, II.F; *see also* Liao, *supra* note 352, at 163 (emphasizing the importance of examining the robustness of parametric assumptions across groups).

518. *See, e.g.*, Wojtyś et al., *supra* note 159, at 1 (describing the generalized joint regression modeling framework).

519. *See supra* Subpart II.E.

520. Williams purports to employ matching techniques to check the sensitivity of his results to extrapolation bias, but he fails to present any of the necessary diagnostic tests to demonstrate that the preprocessing was successful. My own reexamination of Williams's matching model reveals that, when done correctly, the results contradict Williams's conclusions. *See supra* Subpart II.E.

limited to law students with overlapping values across the six tiers reveals that there are no statistically significant mismatch effects.[521]

6. Measurement Error Bias

a. Instability Bias

Combining tiers that are indistinguishable on mismatch-relevant characteristics, or by modeling the effects of the actual values of the variables used to form the tiers, would minimize this type of bias using BPS data. Admittedly, both approaches suffer from measurement error, but the latter approach more accurately reflects average differences in tier-level student quality.[522] The first choice analysis should be abandoned because the BPS data do not contain information permitting the determination of whether the second (third) choice law school is more, equally, or less selective than the first (second) choice law school (or, in some cases, whether respondents were admitted to their first or second choice).[523] As noted earlier, Sander and Taylor claim that "the first choice/second choice model provides the soundest direct empirical test of mismatch with available data, and it strongly confirms the [mismatch] hypothesis."[524] Their confidence in this approach is misplaced because, for reasons I describe above, it is implausible that first choice framework substantially reduces potential omitted variable bias such that it outweighs the bias from violating the SUTVA. Neither Ayres and Brooks nor Sander's analysis of first choice effects attempt to test the plausibility of the assumption by controlling for the nearly two hundred pretreatment variables available in the BPS data. Whereas Ayres and Brooks are much more circumspect about the interferences they derive from the first/second choice framework, Sander and Taylor are sanguine about the merits of the approach. Finally, the analyses of the mismatch effect on LGPA—especially FYGPA—must

521. *See infra* Appendix C ("[I]nferences drawn from Sander's analysis are implausible because of the lack of comparable students in the BPS data" and "[w]hen limiting [Sander's] analysis to students with acceptable covariate balance, the [mismatch] effect . . . is indistinguishable from zero . . . . As I have shown [in *supra* Parts II.D, II.E], the only plausible . . . comparisons are those involving students in adjacent law school tiers.").

522. *See supra* Subpart II.F.1.a. The degree of unreliability of the actual values of the variables used to form the tiers can be incorporated in the model, thereby resulting in more plausible parameter estimates and standard errors. *See, e.g.*, Rothstein, *supra* note 369, at 312; Hardin et al., *supra* note 369, at 363.

523. *See* text accompanying *supra* notes 336–339 (noting that respondents in the BPS view their first, second, and third choice school in drastically different ways that contribute to instability bias under that analytical framework).

524. SANDER & TAYLOR, *supra* note 4, at 86.

be reformulated to account for interference between units (or spillover effects).[525] This is of particular concern in Sander's analysis because LGPA is standardized within school, so it represents rank within a class and is not directly comparable to students in other schools.[526] These concerns notwithstanding, it is important to note the analysis of mismatch on LGPA is not particularly illuminating because holding UGPA and LSAT constant, attending a law school in an elite law school negatively impacts LGPA regardless of race.[527]

b. Correlated Measurement Error Bias

Improperly measured variables are ubiquitous in social science research—especially survey research. Nearly all standard statistical models assume that the explanatory variables are measured without error. Violation of this assumption may bias causal inference when the measurement errors between the variables in the model are correlated with the true variable of interest, other variables in the model, or the measurement errors of other variables in the model. By Sander's own admission, the key variables in his mismatch analysis suffer from substantial measurement error. Whenever possible, analysts should attempt to minimize potential bias resulting from measurement error by utilizing information either already available in the data or preexisting knowledge of the reliability of particular measurements.[528] I describe two ways potential bias from certain types of measurement error in the BPS data can be assessed, and sometimes minimized, when examining mismatch effects: (1) combining the second and third tiers because these tiers (and their centroid law schools) are statistically indistinguishable along

525. *See supra* Subpart II.F.1.c.

526. *See* Sander, *Mismeasuring*, *supra* note 24, at 2007 n.12 ("But all of the data sets I use standardize grades within each individual school. The 'grade' variable is thus more of an indicator of class rank.").

As noted earlier, under certain assumptions, a group-level rather than individual-level causal effect can be measured. One should also employ various sensitivity tests in the presence of spillover effects to determine whether those assumptions are justified and group-level causal effects can be properly estimated. *See* VANDERWEELE, *supra* note 64, at 416–17.

527. *See, e.g.*, Arcidiacono & Lovenheim, *supra* note 7, at 17 (explaining that Sander's results strongly suggest that everyone is harmed by attending a more elite school with respect to grades and bar passage because the negative effect of grades swamps the positive effect of school quality); Harris, *supra* note 355, at 304 n.13 (noting that, before affirmative action, there was a bottom tenth of the law school class).

528. *See supra* Subpart II.F.2.

dimensions relevant to mismatch (UGPA, LSAT, and selectivity)[529] and (2) specifically incorporating the degree of uncertainty in a respondent's actual LSAT score.[530]

B. Improving External Validity

Another type of bias, heretofore unidentified in this Article, is bias resulting from external invalidity. It cannot be overlooked that the BPS data are getting increasingly old, and with each passing year, it becomes less and less likely the data can tell us much that is useful about the effects of affirmative action in law schools today even if the work by Sander and Williams was scientifically defensible (which it is not) because the entering credentials of law students have significantly changed overtime, including black law students, as well as the availability and utilization of LSAT and bar exam preparation courses and materials. Black students with academic index scores in 1991 that predicted a less than 50 percent chance of graduating and passing the bar have essentially disappeared from the ranks of law school matriculants. For example, in 1991, 22.1 percent of black students had entering credentials that predicted less than a 50 percent chance of passing the bar.[531] In more recent years, those percentages are significantly lower: 4.9 percent (2004), 5.8 percent (2005), 7.5 percent (2008), and 5.3 percent (2011).[532] In other words, in the twenty-year period since the BPS cohort began law school, the percentage of black students least likely to pass the bar has dropped by 16.8 percentage points (or by 76 percent). Analysts of the BPS data that seek to generalize to current law school dynamics must be mindful of this fact and qualify their analyses and conclusions accordingly.[533]

529. *See supra* Subpart II.F.1.a.
530. *See supra* Subpart II.F.2.
531. Kidder & Lempert, *Mismatch Myth*, *supra* note 22, at 113 tbl.6.1.
532. *Id.*
533. Sander also assumes that affirmative action is responsible for concentrating black students with lowest entering credentials in lowest ranked law schools because these students would not have been admitted to any law school absent affirmation action. Sander, *Systemic Analysis*, *supra* note 7, at 478 ("[T]he system [of affirmative action] as a whole leads to the admission of an additional five or six hundred black students—about one-seventh of the annual total—who would not otherwise be admitted to any accredited school."); *accord* U.S. COMM'N ON CIVIL RIGHTS, *supra* note 430, at 26 (testifying to a federal agency that race-based affirmative action results in a "net addition of blacks . . . in the lowest-tier schools, . . . [with] such marginal academic credentials that they face long odds against ever becoming attorneys"). His assumption is implausible for at least two reasons. *First*, 18 percent of the black students in the BPS study graduated from a HBLS, and these schools do not need to practice race-based admission policies for black students because the overwhelming majority of applicants are black. Only one HBLS, Howard University (ranked 108th by U.S. News in 2019) receives a nontrivial number of nonblack

Despite this recent trend, Sander has opined:

> All of the findings about bar exam results in *Systemic Analysis*, and in much of the ensuing debate, are based on the 1994 numbers.
>
> The current situation is much more severe . . . [for blacks] [because] 1994 was the historical high-water mark of national bar passage rates . . . [and] the rate has steadily fallen since then . . . .[534]

---

applicants, and its student population was 88 percent black in 2019. Indeed, if HBLS's did practice race-based admissions, it would be for the small number of nonblack applicants.

Second, many of the lowest ranked law schools admit a very large percentage applicants of their applicants in an effort to cover operating costs with revenue from tuition. In fact, in recent years, several law schools have made national headlines because their student demographics have shifted dramatically. These schools now enroll a greater percentage of black and Latinx students, in large part, because of the availability of federal and private educational loans and the exodus of many white students from these lower ranked schools to higher ranked schools after the first year of law school. These black and Latinx students are much more likely to pay full tuition (or close to full tuition) than white students, and their tuition dollars are used, in part, to subsidize white students with higher entering credentials. *See* Erin Thompson, *Law Schools Are Failing Students of Color*, NATION (June 5, 2018), https://www.thenation.com/article/archive/law-schools-failing-students-color [https://perma.cc/N55T-3NW6] (noting that the lowest ranked law schools charge higher tuition than the most elite schools, but offer dismal employment prospects for graduates—many of whom are black and Latinx); Mystal, *supra* note 503 (describing the various ways that for-profit law schools prey on black students with high aspirations but little knowledge of how the postgraduate system works, and charge as much as triple the tuition as similarly ranked not-for-profit law schools); Wesley Whistle, *Is Your Law School Worth It?*, FORBES (Nov. 21, 2019, 4:36 PM), https://www.forbes.com/sites/wesleywhistle/2019/ 11/21/is-your-law-school-worth-it/ #2b9a798038c7 [https://perma.cc/Y4WV-JXHH] (analyzing Department of Education data and discovering that graduates from the five law schools with the highest median debt ($195,000) also reported the lowest median annual salary ($38,000), and these law schools were all ranked among the bottom fifty); Morrissey, *supra* note 503 (explaining that students from the least selective schools are more likely to have borrowed the full cost of their education, more likely to subsidize the tuition of their better credentialed (and mostly white) classmates, and face the poorest employment prospects). Consequently, it is unlikely that these schools would admit significantly fewer black and Latinx students absent race-based affirmative action because they must satisfy somewhat static enrollment thresholds in an attempt to remain solvent. This, in turn, makes the quality of the overall applicant pool much less relevant. For example, for-profit Arizona Summit Law School, with abysmal bar passage and employment rates, formed an affiliation with Bethune-Cookman, an HBCU, in an effort to gain legitimacy and attract black college graduates. *See* Mystal, *supra* note 503. Black students who attended Florida A&M College of Law, an HBLS, with the same median LSAT score as Arizona Summit, paid one-third of the tuition ($14,000 versus $45,000) and passed the bar on the first attempt at twice the rate (50 percent versus 25 percent). *Id.*; *accord* Bonadies et al., *supra* note 503; FED. RESERVE BD., REPORT ON THE ECONOMIC WELL-BEING OF U.S. HOUSEHOLDS IN 2018, at 40 (2019) (reporting that black students are five times more likely to attend private for-profit colleges than white students).

534. Sander, *Reply to Critics*, *supra* note 10, at 2004 (citations omitted).

A very useful illustration of the problems with Sander's unsupported assertion was provided by Lempert and colleagues' analysis of the impact of affirmative action at the University of Michigan Law School (UMLS).[535] They note that close to 95 percent of the UMLS minority graduates now pass the bar,[536] but the situation was quite different in the earlier days of affirmative action:

> An uncomfortably high proportion of the [UMLS's] initially small African American cohorts were either in serious academic difficulty or after graduation failed at least one bar exam. But seeing this the school quickly adjusted its affirmative action admissions criteria, and research shows that soon thereafter nearly all of Michigan's students, both minority and white, graduated, passed the bar and went on to have successful careers by almost any measure [including current income, self-reported satisfaction, and professional service contributions]."[537]

Sander challenged Lempert and colleagues' assertion that black students at UMLS passed the bar at rates similar to white students, claiming that he "was always highly skeptical of Lempert's claim because it did not line up with [the BPS data]" that tracked the 1991 law school matriculants and revealed that 30 percent of the black students in the top two tiers failed the bar on the first attempt and 15 percent never passed the bar.[538] Sander obtained bar passage data from UMLS and compared those records to photos of UMLS graduates (to assess their race) and concluded that black students' first-time bar passage rate was 62 percent and the overall bar passage rate was 76 percent.[539] But Sander's analysis is suspect for at least two reasons. First, by nearly every measure, UMLS is a top tier (national) school. In the BPS data, 81.8 percent of black students who graduated from a law school in the top tier passed the bar on the first attempt, and 94.2 percent of black students eventually passed the bar.[540] Even assuming, for the sake of argument, that UMLS was grouped in the second tier in the BPS, according to Sander's own statements, UMLS would be indistinguishable in terms of UGPA, LSAT, and selectivity from many schools in the first tier.[541] Second, nearly 250 student

535. *See* Lempert et al., *supra* note 506.
536. *See id.* at 422.
537. KIDDER & LEMPERT, WHITE PAPER, *supra* note 22, at 2 n.2; *see also* Lempert et al., *supra* note 506, at 395.
538. Sander, *Listening to the Debate*, *supra* note 113, at 941.
539. *See id.* at 943.
540. Lempert, *supra* note 67, at 21; WIGHTMAN, BAR PASSAGE STUDY, *supra* note 36, at 28 tbl.7, 33 tbl.11.
541. Sander, *Mismeasuring*, *supra* note 24, at 2009 (emphasizing that the BPS tiers do not provide a cardinal ranking of law schools).

photos were missing, seriously undermining Sander's ability to appropriately determine the race of the UMLS graduates—and this is particularly important considering the small number of black students in the sample.[542] Lempert's reanalysis of Sander's data, which identified the race of students in two-thirds of the 250 missing photos in Sander's analysis, revealed that 91 percent of UMLS black graduates who took the bar for the first time in 2004 or 2005 passed the bar by the end of 2006.[543] The percentage of white, Asian, and Latinx students who graduated from UMLS and took the bar in 2004 or 2005 and passed the exam by 2006 were, respectively, 99 percent, 95 percent, and 100 percent.[544] Lempert and colleagues explain:

> [T]here is little evidence that affirmative action deflates Michigan's bar passage results. Hispanics, who benefit from affirmative action, did about as well on the bar as whites, who do not, and African Americans, who benefit from affirmative action, did about as well on the bar as Asians, who do not.[545]

A similar pattern held for first-time bar passage rates.[546]

CONCLUSION

After painstakingly describing the core errors of statistical inference frequently reproduced by political scientists (and explaining how to avoid them) in "How Not to Lie with Statistics," King optimistically observed, "Fortunately, in each case, there are plausible reasons for the initial mistake[] . . . or conceptual problem and a relatively painless solution to the problem."[547] Nearly twenty years later, Ho—a protégé of King—explained,

> The empirical investigation of affirmative action is important and should be subjected to scientific scrutiny. The fortunate fact is that we are not alone in this venture. Tools for analysis have been developed across academic fields, providing some easy fixes to reassess the mismatch hypothesis with more credible and theoretically consistent assumptions.[548]

542. *See* Lempert, *supra* note 67, at 21.
543. *Id.*
544. *Id.* at 22 tbl.1.
545. *Id.* at 21.
546. Data for UMLS revealed that 78.3 percent of black students passed the bar on the first time, compared to 95.9 percent of white students, 83.1 percent of Asian students, and 97.4 percent of Latinx students. *Id.* at 22 tbl.1.
547. King, *supra* note 3, at 684; *see also supra* text accompanying note 3.
548. Ho, *Reply to Sander*, *supra* note 23, at 2016.

In the roughly fifteen years that have followed Ho's admonition, Sander has been provided ample opportunities to un-"muddy" the waters,[549] yet his responses to his numerous critics, as well as his attempted interventions in the courts, demonstrate his inability (or unwillingness) to adhere to well-established norms of social scientific inquiry and causal inference.[550] As previously noted, equally troubling has been Sander's tendency to selectively identify findings from the research literature that merely appear supportive of mismatch, ignore or mischaracterize unsupportive findings, rely on a number of contradictory assumptions, overstate implications, and understate caveats.[551] Sander's most recent work continues in this vein, nibbling at the margins of the scholarship of those who have revealed persistent and seemingly fatal laws in his research, while avoiding tackling the fundamental shortcomings that have plagued the entire enterprise.[552] Theoretical physicist Richard Feynman famously remarked, "It does not make any difference how beautiful your guess is. It does not make any difference how smart you are, who made the guess, or what his [sic] name is—if it disagrees with experiment it is wrong. That is all there is to it."[553]

549. *See* Dauber, *supra* note 238, at 1902, 1910 n.58 ("Unfortunately, Sander['s "Systemic Analysis"] has muddied rather than clarified the waters with a flawed and ultimately misleading contribution. . . . This is apparently not the first time that Sander has muddied the waters of an important public policy debate by making claims that were declared by experts in the field to be without empirical basis and contrary to the relevant substantive and methodological literature. . . . The unimpressive record racked up by Sander's involvement in the living wage and affirmative action debates suggests that it might be wise to subject any future policy interventions issuing from his research shop to a healthy degree of skepticism."); *see also* Stephen Menendian & Richard Rothstein, *Putting Integration on the Agenda*, 28 J. AFFORDABLE HOUS. 147, 176 (2019) (reviewing Sander's work on housing segregation and concluding: "In an effort to develop contrarian narratives, at times [Sander] overreach[es]. Moreover, the nearly exclusive reliance on dissimilarity scores [for housing segregation] at a time when the quality and quantity of alternative measures of segregation has blossomed is especially puzzling.").

550. *See, e.g.*, SANDER & TAYLOR, *supra* note 4, at 85–86 ("None of these [critiques of mismatch] could have survived peer review by any expert aware of the facts discussed here. . . . At the end of the day the evidence points overwhelmingly toward a large law school mismatch problem . . . . "); Sander, *Replication of Mismatch*, *supra* note 4, at 88 ("[My] results add significantly to the body of research finding support for the law school mismatch hypothesis and subtracts from the thinning body of substantive critiques.").

551. In social scientific parlance, Sander's research appears to be plagued by confirmation bias—a cognitive bias that leads individuals to misinterpret new information as supporting previously held hypotheses, as well as induces a degree of overconfidence such that the individual may come to believe with near certainty in a false hypothesis despite receiving an infinite amount of information. *See* Matthew Rabin & Joel L. Schrag, *First Impressions Matter: A Model of Confirmatory Bias*, 114 Q. J. ECON. 37, 59–62 (1999).

552. *See* Sander, *Replication of Mismatch*, *supra* note 4 (failing to address various sources of bias identified by critics of mismatch theory in Sander's response to those same critics).

553. RICHARD FEYNMAN, THE CHARACTER OF PHYSICAL LAW 156 (M.I.T. Press 1967) (1965).

In an article reflecting on the growth (and growing pains) of the empirical legal studies movement, law and society scholar Richard Lempert lamented, "The work [in empirical legal studies] being produced is not always of the highest quality. Some of it fails to understand the logic of social science inquiry or gives insufficient attention to model specification, data flaws, and the operationalization of crucial concepts."[554] This deeply flawed research can influence judicial decisions, legislation, or systemic reform, especially if it captures the population's imagination in the form of a pithy takeaway message.[555] Lempert lists several (in)famous examples of methodologically unsound research that was coupled with strong publicity efforts before the work being properly vetted by the scientific community, including: James Q. Wilson and George L. Kelling's work on neighborhood disorder and crime ("broken windows");[556] Isaac Ehrlich's work on the deterrent effect of capital punishment;[557] John R. Lott and David B. Mustard's work on the impact of right-to-carry laws on violent crime;[558] and Sander's work on the effects of affirmative action in law schools.[559] These studies "resounded in the halls of policy when at best they raised issues for further examination and at worst seemed aimed at promoting particular policies

554. Lempert, *supra* note 67, at 13.
555. *See id.* at 15.
556. George L. Kelling & James Q. Wilson, *Broken Windows: The Police and Neighborhood Safety*, ATLANTIC MONTHLY (Mar. 1982), https://www.theatlantic.com/magazine/ archive/1982/03/ broken-windows/304465 [https://perma.cc/SPD4-8GNE]; *accord* James Q. Wilson & George L. Kelling, *Broken Windows: The Police and Neighborhood Safety*, *in* THE CRIMINAL JUSTICE SYSTEM: POLITICS AND POLICIES 103 (George F. Cole & Marc G. Gertz eds., 7th ed. 1998). To be clear, James Wilson and George Kelling did not attempt to provide a direct or comprehensive test of order maintenance policing when introducing their theory; rather, the authors report observational evidence of the impact of the order maintenance policing and provided indirect evidence supporting their hypothesis based on some experiments conducted by Stanford University psychologist, Philip Zimbardo. Several years after the Wilson and Kelling article was published, Northwestern University political scientist, Wesley Skogan, claimed to have found convincing support for the maintenance order policing. *See* WESLEY G. SKOGAN, DISORDER AND DECLINE: CRIME AND THE SPIRAL OF DECAY IN AMERICAN NEIGHBORHOODS (1990). Skogan's work received significant praise from members of the academy, *see, e.g.*, Tracey L. Meares & Dan M. Kahan, *Law and (Norms of) Order in the Inner City*, 32 LAW & SOC'Y REV. 805, 822–23 (1998), before serious methodological errors and questionable reporting practices were revealed in Skogan's work. *See, e.g.*, BERNARD E. HARCOURT, ILLUSION OF ORDER: THE FALSE PROMISE OF BROKEN WINDOWS POLICING (2001) (providing a critical review of Skogan's work, as well as several scholars' insufficiently skeptical acceptance of Skogan's conclusions).
557. Isaac Ehrlich, *The Deterrent Effect of Capital Punishment: A Question of Life and Death*, 65 AM. ECON. REV. 397 (1975).
558. John R. Lott, Jr. & David B. Mustard, *Crime, Deterrence, and Right-to-Carry Concealed Handguns*, 26 J. LEGAL STUD. 1 (1997).
559. Sander, *Systemic Analysis*, *supra* note 7.

regardless of weaknesses in the science."[560] Fortunately, the larger scientific community identified and responded to these scholars' numerous errors of inference and misrepresentations of fact.[561] But such corrective efforts continually divert scholars who are most mindful of the requirements of credible social science from engaging in new areas of inquiry with strong policy relevance.[562] The twin sins of this flawed research, then, are the propagation of unreliable information *and* hindrance of the growth of the field.[563]

APPENDIX

This appendix includes additional material on several shortcomings in Sander's work that undermine the validity of his results. In the ensuing discussion, I offer reanalyses of Sander's models and demonstrate that his results are highly sensitive to minor plausible changes to his statistical models.[564] The

560. Lempert, *supra* note 67, at 15; *accord* Kidder & Lempert, *Mismatch Myth*, *supra* note 22, at 106 ("[Mismatch] research [in the law school context] has been effectively marketed through both conservative and mainstream media. As a result, it has had a major role in shaping public perceptions and debate.").

561. *See, e.g.*, Edward E. Leamer, *Let's Take the Con Out of Econometrics*, 73 AM. ECON. REV. 31 (1983) (identifying numerous methodological flaws in Isaac Ehrlich's work on the deterrent effect of capital punishment and concluding that there is no deterrent effect); Empirical Scholars Brief in *Fisher I*, *supra* note 30, at 1 (concluding that Sander's research on mismatch effects in law schools is "unreliable, [and] fail[s the] basic tenets of research design"); HARCOURT, *supra* note 556 (concluding that the broken windows thesis lacks an evidentiary foundation); Tomislav V. Kovandzic & Thomas B. Marvell, *Right-to-Carry Concealed Handguns and Violent Crime: Crime Control Through Gun Decontrol?*, 2 CRIMINOLOGY & PUB. POL'Y 363 (2003) (reporting that right-to-carry laws do not impact violent crime rates); *see also generally* Leamer, *supra* note 459, at 308 ("Decentralized studies of fragility [of statistical inferences] are common whenever an inference matters enough to attract careful scrutiny.").

562. In the preface to the second edition of his popular text on regression analysis, statistician Richard Berk, a signatory of the ESB, remarked:

> [UC Berkeley Statistics Professor David Freedman] was my bridge from routine calculations within standard statistical packages to a far better appreciation of the underlying foundations of modern statistics. He also reinforced my skepticism about many statistical applications in the social and biomedical sciences. Shortly before he died, David asked his friends to "keep after the rascals." I certainly have tried.

BERK, *supra* note 220, at xii.

563. *See, e.g.*, Dauber, *supra* note 238, at 1913–14 (noting that "[r]esponding to the errors in Sander's article has occupied an enormous amount of time and attention from social scientists that might have been more profitably spent, a calculation certainly not lost on those scientists themselves").

564. "We [economists] need to be shown that minor changes in the list of variables do not alter fundamentally the conclusions, nor does a slight reweighting of observations, nor correction for dependence among observations, *etcetera*, *etcetera*." Leamer, *supra* note 459, at 308.

modifications I adopt are in accordance with widely accepted practices in the field statistical inference. Appendix Subpart A addresses the first four problems that were mentioned, but not specifically addressed, in Subpart II.A.2: (1) the focus on fully standardized regression coefficients which give the misimpression that the total tier effect on the probability of bar passage is nearly three times as large as it actually is; (2) the miscalculation of standard errors for the total tier effect that ignores the nonindependence of students enrolled in the same law schools and biases tests of statistical significance in favor of the mismatch hypothesis; (3) the sensitivity of Sander's results to removal of the tier containing all HBLS schools; and (4) the extremely poor ability of Sander's model to accurately predict both LGPA and bar passage (both first-time and eventual). Appendix Subpart B closely examines Sander's claim that there is "zero evidence of [posttreatment] bias" based on his conducting more than a dozen different tests.[565] I show that his tests, which I label his "Sequential Analysis," not only fail to properly test for posttreatment bias, but also do not support the conclusions he derives from them because of the: (1) incorrect interpretation of marginal effects for tier location; (2) miscalculation of standard errors and associated significant tests; and (3) miscalculation of the effect of race/ethnicity. Appendix Subpart C provides a more detailed discussion of the impact of functional form misspecification that was presented in Subpart II.D. In particular, this Part describes how Sander's conclusions concerning the impact of first choice on law school tier location and first-time bar passage are not only incorrect, but also overstate the impact of first choice on these outcomes of interest.

A. Mediation Analysis

1. Calculation of Tier Total Effect

The preferred approach in the social sciences is to "comput[e] quantities of interest and their uncertainties."[566] Scholars have emphasized that "substantively informative [interpretations] . . . convey[] a key quantity of interest in terms the reader wants to know. . . . In contrast, bad interpretations are substantively ambiguous and filled with methodological jargon . . . ."[567] Sander's discussion of the total effect of location on first-time bar passage not only lacks substantive meaning, but it implies a much greater practical effect than his model

565. *See supra* note 80 and accompanying text.
566. King et al., *supra* note 211, at 349.
567. *Id.* at 347–48.

describes.[568] Specifically, he claims that his mediation analysis reveals that tier location has a statistically significant negative impact on LGPA and a statistically significant positive impact on first-time bar passage; meanwhile, LGPA has a statistically significant positive impact on first-time bar passage.[569] Putting aside, for the moment, several methodological errors with Sander's analysis,[570] I was able reproduce his findings based on the problematic assumption that there were no common causes of the tier-LGPA (A_3) and LGPA–bar passage (A_2) relationships unaccounted for in the analysis (see Figure 4). Based on Sander's model, the total effect of tier location results in a 2.4 percentage point reduction in the probability of passing the bar on the first attempt. This effect is statistically significant at the 95 percent confidence level ($p < .05$).[571] Sander buries discussion of the total effect of tier location, which is the most important estimate from his mediation analysis, in a footnote without any reference in the text;[572] yet, in the main body of his text, he reports the standardized parameter estimates from the various direct effects of all of the variables in the model. One must wonder whether Sander excluded this because the magnitude of the total tier effect estimate is quite trivial.

Not only does Sander neglect to provide a transparent discussion of the total tier effect, he reports a fully standardized total tier effect in the footnote. This statistic is both unhelpful in understanding the impact of tier location on first-time bar passage and misleading with respect to the magnitude of the total tier effect. A fully standardized effect transforms both the explanatory variables and dependent variable into deviations from their average scores. The fully standardized total effect Sander reports (-0.073) reflects a change in standard deviation units in the probability of passing the bar on the first attempt for black students, for (essentially) a one–standard deviation change in tier location. This is the fully standardized effect because both variables—tier location and probability of first-time bar passage—are standardized, and a total tier effect because it captures both the direct effect on tier location on

568. Sander repeats this misleading reporting practice in his first choice analysis. *See infra* note 658.

569. *See* Sander, *Mismeasuring*, *supra* note 24, at 2008. Sander limits his analysis to the subsample of black students in the BPS. *Id.* at 2007.

570. *See supra* note 122 and accompanying text.

571. *See infra* note 577 and accompanying text.

572. *See* Sander, *Mismeasuring*, *supra* note 24, at 2008 n.16. In economics, the total effect of an explanatory variable with direct and indirect paths to an outcome variable is often called a reduced-form parameter. A reduced-form equation is produced by solving for each dependent variable such that the resulting equations express the endogenous variables as functions of the exogenous variables. CAMERON & TRIVEDI, *supra* note 130, at 25.

first-time bar passage and the indirect effect of tier location on bar passage that operates through the LGPA. To obtain an estimate of the total tier effect on the probability of first-time passage, one would need to take the product of the total standardized effect and the standard deviation of the first-time bar passage variable (-0.073 × 0.487). This is the semistandardized effect (-0.036) because the probability of first-time bar passage is now measured on its original scale, but the tier location variable is still measured on the standardized scale. Consequently, a one–standard deviation increase in tier location results in a 3.6 percentage point decrease in the probability of passing the bar on the first attempt for black students. Notice that the fully standardized effect gives the impression that the magnitude of the total tier effect on the probability of bar passage is twice as large (-0.073 vs. -0.036) as its actual value.

The problem with Sander's reporting of the total tier effect does not stop there. Because he standardizes the tier location variable, the 3.6 percentage point decrease is for a one–standard deviation change in tier location, and not a one-unit change on the original scale of the tier location variable. This difference is particularly meaningful because a one–standard deviation change on the standardized tier variable equals a 1.483 change in the unstandardized tier location variable—nearly 50 percent larger than a one-unit change on the original scale of the tier location variable. This would be roughly equivalent from moving from Tier 3 to Tier 1.5, rather than just to Tier 2. To obtain the total unstandardized tier location effect, one must divide the semistandardized effect reported earlier by the standard deviation of the unstandardized tier location variable (-0.036 ÷ 1.483). This is the causal estimate of the total tier effect of -0.024 (or a 2.4 percent decrease) that I reported at the outset of this discussion. Not only does this final measure of a 2.4 percent decrease in the probability of first-time bar passage for a one-level increase in tier location make the most intuitive sense, but that is also the most policy relevant metric.[573] This measure is also one-third of the magnitude (-0.024 ÷ -0.073 = 0.329) of the fully standardized total tier effect that Sander reports.

Even if one ignores the host of methodological problems present in Sander's mediation analysis and accepts the finding of a -0.024 total tier effect, Lempert has argued that a "2.5 percent decrement vis-à-vis white students is a small price to pay for the bar passage advantage associated with attending

573. The fact that the fully unstandardized tier effect (2.4 percent decrease) is the most policy relevant metric is underscored by the fact that, when referencing Sander's results, other analysts have focused on the unstandardized parameter rather than the standardized one. *See* Arcidiacono & Lovenheim, *supra* note 7, at 19; Ho, *Reply to Sander*, *supra* note 21, at 2014; Lempert, *supra* note 320, at 149.

an elite school."[574] In his *Fisher II* amicus brief, Sander favorably cites to a recently published review of the mismatch literature by economists Arcidiacono and Lovenheim. These authors conclude that, even under the rather strong assumption of no omitted variable bias, the magnitude of the mismatch effect in the BPS data is extremely modest: "African American students at selective law schools are about 2.5 percentage points less likely to pass the bar than white students at selective law schools."[575]

For these reasons, scholars have cautioned against the use of standardized coefficients because "[t]hey do not measure what they appear to; they substitute statistical jargon for [substantive] meaning; they can be highly misleading; and in nearly all situations, there are better ways to proceed."[576] The purpose of Sander's mediation analysis was to demonstrate how tier location, through its negative effect on LGPA, offsets its positive effect on first-time bar passage. Consequently, the calculation of a sensible the total effect was integral to his analysis.[577] By highlighting the separate direct effects, hiding the total tier effect in a footnote, and using transformed estimates that obscure the impact of the variables in the model on the natural metric of the dependent variable, Sander gives the misleading impression that law school

574. Lempert, *supra* note 320, at 149.

575. Arcidiacono & Lovenheim, *supra* note 7, at 19. Of Sander's first-time passage model, Ho notes that:

> [T]he model does lead to a borderline significant effect of tier, but not at the significance level of .01 that Sander reports for all other coefficients. The marginal effect of going to a higher-tier law school in this model is roughly a 2.4% increase [sic] in the probability of first-time bar passage, plus or minus 2.5% at a 99% confidence level. This effect is neither statistically distinguishable from zero [for eventual bar passage] nor close to explaining the substantial black/white bar passage gaps as Sander originally claimed.

Ho, *Reply to Sander*, *supra* note 21, at 2014 (footnotes omitted). Ho's sentence contains a typographic error. Going to a higher tier law school leads to a 2.4 percentage point *decrease* in the probability of passing the bar on the first attempt for black students. It is clear from Ho's discussion of Sander's results that this was merely an editing error rather than a substantive mistake in Ho's analysis. I reexamined Sander's model and obtained the same results as Ho: going to a higher-law school leads to a 2.4 percentage point decrease in the first-time bar passage for black students (controlling for LGPA). The effect is statistically significant at the 95 percent confidence level (*beta* = -0.024, std err. = 0.01).

576. King, *supra* note 3, at 669.

577. *See* Ho, *Reply to Sander*, *supra* note 21, at 2014 n.17 (noting that the "substantive effect [of tier location] is glossed over by [Sander's] standardization, which does not accord with a substantive quantity of interest and does not make the estimates comparable"); *cf.* King et al., *supra* note 211, at 360 (noting that "statistical analysts have a responsibility to present their results in ways that are transparent to everyone").

selectivity substantially hurts beneficiaries of affirmative action when, in fact, even granting all of Sander's implausible assumptions, the impact is minimal.

2. Calculation of Standard Errors

Standard errors measure the accuracy with which sample statistics, such as the effect of tier location on first-time bar passage, represent the actual population. Sander's mediation analysis fails to take into account the nonindependence of students in the various tiers when calculating the standard errors of his causal estimate.[578] Sander's independence assumption is peculiar because his entire analysis is constructed around identifying differences across tiers in first-time bar passage rates. His inclusion of the tier location variable in his models captures between-tier differences in the likelihood of passing the bar, but his standard errors (and corresponding tests of statistical significance) do not account for the grouping of respondents into those tiers. Standard errors that ignore clustering can either underestimate or overestimate the true standard errors because omitted variables and explanatory variables are often correlated within a cluster.[579] Proper calculations of standard errors for nonindependent observations in clusters—often referred to as "clustered" or "robust-clustered" standard errors—were introduced in the 1960s and have been common practice in economics (and social science, more generally) since the 1980s,[580] so it would be quite odd if Sander were unaware of the importance of employing them.[581] Moreover, Williams's analysis of mismatch

578. *See infra* Appendix B.2 (discussing Sander's miscalculation of standard errors in his sequential analysis).

579. The direction of the bias will depend on whether the within-cluster variance is larger/smaller than the between-cluster variance.

580. Halbert White, *A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity*, 48 ECONOMETRICA 817, 821 (1980) (noting that methods for the calculation of standard errors under conditions of nonindependence were first introduced by Eicker in 1963). Whereas taking into account tier location attempts to remove unobserved heterogeneity between the different tiers, clustered standard errors account for situations in which observations within the tiers are not independent. When there is heterogeneity in the tier location effects, clustered standard errors are necessary even when including tier effects. Alberto Abadie, Susan Athey, Guido W. Imbens & Jeffrey Wooldridge, *When Should You Adjust Standard Errors for Clustering?* 17 (Nat'l Bureau of Econ. Rsch., Working Paper No. 24003, 2017).

581. Two scholars surveyed the top three political science journals—*American Journal of Political Science*, *American Political Science Review*, and *International Organization*—between 2009 and 2012 and discovered that robust standard errors were used in, respectively, 45 percent, 66 percent, and 73 percent of all articles that use regression analysis. The authors also found that, across all academic fields in 2014, more than 75,000 articles reported using robust standard errors. Gary King & Margaret E. Roberts, *How Robust*

theory, which Sander has favorably cited on multiple occasions,[582] employs robust standard errors to account for the clustering of respondents in tiers.[583] When the correct standard errors are used, the results directly contradict Sander's core conclusions. Specifically, the total effect of tier is statistically insignificant ($p = 0.215$).[584]

3. Sensitivity to Removal of the Sixth (HBLS) Tier

I reanalyzed a mediation analysis adopting Sander's methods and assumptions, but removed the sixth tier, which is entirely composed of HBLS for reasons I have previously discussed.[585] Sander models the effect of tier location linearly,[586] which may bias his estimate of tier effect on LGPA and first-time bar passage because, as I mentioned earlier, black law students who attend HBLS may perform better in law school and on the bar exam than similarly situated black students attending schools in more selective tiers for reasons unrelated to mismatch.[587] Removing the HBLS tier excludes 18

Standard Errors Expose Methodological Problems They Do Not Fix, and What to Do About It, 23 POL. ANALYSIS 159, 159 (2015).

582. *See, e.g.*, Sander, *Stylized Critique*, *supra* note 8; Sander, *Mismatch & the ESB*, *supra* note 30; Sander, Brief in *Fisher II*, *supra* note 40.

583. Williams, *supra* note 22, at 187–93 (reporting robust standard errors for all of his analyses using the BPS data).

584. The analytical formula for calculating standard errors when observations are clustered into tiers presumes that the number of clusters is large, and often leads to underrejecting (or downward bias) with few clusters. An alternative approach involves repeated resampling the data (with replacement) to estimate standard errors based on the empirical distribution of the parameters across the samples. These resampling methods are much more likely to produce unbiased estimates of the standard errors. A. Colin Cameron, Jonah B. Gelbach & Douglas L. Miller, *Bootstrap-Based Improvements for Inference With Clustered Errors*, 90 REV. ECON. & STAT. 414, 414–15 (2008). To test the sensitivity of my results, I adopted two approaches to calculating standard errors: (1) the analytical formula ("cluster-robust"), and (2) resampling ("cluster-bootstrap"). Both approaches provide strong indication that the effect of tier location on first-time bar passage rates is statistically insignificant after properly accounting for nonindependence.

585. *See supra* note 320 and accompanying text (explaining that the HBLS tier is inappropriate for the comparative tier analysis because black students overperform in these schools for nonmismatch-related reasons).

586. *See* Ho, *supra* note 7, at 2001.

587. *See id.*; *see also* Arcidiacono & Lovenheim, *supra* note 7, at 19 ("One caveat is that historically black colleges and universities are in this bottom tier and may operate differently from traditional law school environments."); Lempert, *supra* note 320, at 161–62 (reporting findings that strongly suggest that campus environment and a critical mass of students of color are more likely to explain differences in graduation rates for science majors than mismatch when comparing UC Berkeley and UCLA, which have similar campus profiles with respect to students' entering credentials and acceptance rates); Bostwick & Weinberg, *supra* note 320, at 9 (discovering that the six-year graduation rate

percent of the black students in the sample and renders the total effect of tier location on the probability of first-time bar passage statistically insignificant (*beta* = -0.001, std. err. = 0.023, $p = 0.967$). This underscores how sensitive Sander's results are to the inclusion of the HBLS tier.

I also reanalyzed the models predicting the probability of first-time bar passage for Asian, Latinx, "Other Race," and white students with and without the HBLS tier to examine the sensitivity of the tier location effect. For models including the HBLS tier, there was no statistically significant tier total effect for any of these groups. Models excluding the HBLS tier produced similar results: There were no statistically significant tier effects for any group.[588] It is important to emphasize that all of these models employ Sander's mediation approach, which controls for LGPA and assumes tier-LGPA and LGPA–bar passage unconfoundedness. As I explained in Subpart II.A.2, when this assumption is relaxed, the miniscule mismatch effect that Sander reports (and that is statistically insignificant when correct standard errors are used) completely disappears.

4. Inadequacy of Model Fit

Sander's analyses exhibit very poor agreement between the data and his hypothesized models, yet he routinely fails to acknowledge these stark discrepancies when describing and defending his results. A model's "fit" refers to its ability to adequately describe the underlying data generation process, and an adequately fitting model is one that is consistent with the observed data without being unnecessarily complex.[589] Sander reports the value of a single model fit statistic: the Bentler-Bonett Normed Fit Index (NFI) statistic (0.998).[590] Sander provides no discussion of the NFI, nor reports any of the other standard measures of model fit, although researchers strongly recommend that multiple measures of model fit should be provided when utilizing structural equation modeling (mediation analysis) because the individual measures are subject to various

for women in STEM doctoral programs was significantly impacted by the percentage of women in the cohort).

588. Eliminating the sixth tier resulted in dropping approximately 2.6 percent of students from the BPS sample: 18 percent of black students, 1.4 percent of white students, 1.6 percent of Asians, 6.2 percent of Latinx students, and 3.8 percent of "Other Race" students.

589. JOHN K. KRUSCHKE, DOING BAYESIAN DATA ANALYSIS: A TUTORIAL WITH R, JAGS, AND STAN 290 (2d ed. 2015) ("[D]ata are contaminated by random noise, and we do not want to always choose the more complex model merely because it can better fit noise. Without some way of accounting for model complexity, the presence of noise in data will tend to favor the complex model.").

590. Sander, *Mismeasuring*, *supra* note 24, at 2008 fig.1.

limitations.[591] Generally speaking, a model is considered to have adequate fit when NFI > .95, but there are two major shortcomings of the NFI that are directly applicable to Sander's model: (1) the NFI does not penalize models for estimating unnecessary parameters that increase the likelihood of capitalizing on noise in the data and (2) the NFI is sensitive to sample size, so models estimated on data with over two hundred observations will almost invariably indicate good model fit according to this measure.[592] Sander's model assumes every variable in the model is either correlated with, is a cause of, or is caused by every other variable in the model, with the exception of the relationships between gender and bar passage, and between UGPA and LSAT,[593] so his formulation almost exactly reproduces the observed associations (variances and covariances) and it is unsurprising that his NFI statistic suggests adequate model fit.[594] Sander's mediation analysis is also performed on sample of 1225 black law students—well above the two hundred-observation threshold that biases the NFI measure. Again, the statistical literature is replete with warnings about the NFI measures and the general recommendation is that it should not be used, or if used, is only one of several model fit statistics reported.[595] I reanalyze Sander's mediation model and, below, I report two widely accepted model fit statistics that impose a penalty for unnecessary model complexity and that are not unduly sensitive to sample size. Both measures reveal that Sander's mediation analysis fits the data very poorly.

A common approach to assessing model fit is to compare a proposed (theoretically-consistent) model to an alternative (theoretically-inconsistent)

591. Jeff S. Tanaka, *Multifaceted Conceptions of Fit in Structural Equation Models, in* TESTING STRUCTURAL EQUATION MODELS 10, 15–16 (Kenneth A. Bollen & J. Scott Long eds., 1993) (explaining that measures of model fit, "fit indices," can be grouped into several classes and each class is subject to certain limitations, so it is best to use indices from different classes to cover the limitations of each index).

592. *See id.* at 14–15.

593. Sander, *Mismeasuring*, *supra* note 24, at 2008 fig.1.

594. Standard practice in the literature is to estimate an "overidentified model," which is a model with fewer constrained parameters (nonestimated covariances/correlations between variables in the model) than unconstrained parameters (regression coefficients that must be estimated from the data). PAXTON ET AL., *supra* note 145, at 27. Overidentified models possess more information than what is needed to estimate the parameters in the proposed model, so there is more than one way to determine the value of a parameter and all parameters have at least one unique solution. Analysts should aim to constrain several parameters, thereby increasing one's ability to adequately examine how changes in the assumptions of the model (for example, which variables are included/excluded from the model) impact the results. *See id*. Given the hypothesized causal structure of Sander's model, which posits that nearly all of the variables are related to one another, except for the two aforementioned constraints, it is difficult to assess the sensitivity of his results to different modeling assumptions. *See id.*

595. *See* Tanaka, *supra* note 591.

model and determine whether the proposed model fits the data better than the alternative model. Model selection based on comparative fit is typically favored over selecting a single model and reporting model fit statistics because there is always uncertainty about model structure, "yet simpler and more interpretable models can be identified without compromising the richness of the structural equation [mediation analysis] framework for representing substantive research hypotheses."[596]

Two complimentary measures for assessing relative model fit are the Consistent Akaike Information Criterion (CAIC) and the Bayesian Information Criterion (BIC).[597] These model fit measures enjoy advantages over the NFI, as well as other model fit statistics, because they adequately take into account model complexity and sample size.[598] They quantify the degree of discrepancy between the model and the data, so the smaller the statistic, the smaller the discrepancy and the better the fit. These measures reveal that a model that does not include tier location as an explanatory variable for first-time bar passage fits the data better than a model that includes tier location when taking model complexity into account.[599] The CAIC and BIC for the model with tier location included are, respectively, 20141.727 and 20141.717; whereas the CAIC and BIC for the model excluding tier location variables are, respectively, 16217.122 and 16217.114. Some analysts prefer the Adjusted BIC (ABIC) to the traditional BIC because the BIC has been criticized for being biased towards simpler models, so the ABIC penalizes model complexity slightly less than the BIC.[600] The ABIC statistic provides similar

596. Adrian E. Raftery, *Bayesian Model Selection in Structural Equation Models*, *in* TESTING STRUCTURAL EQUATION MODELS, *supra* note 591, at 163, 175.

597. *Id.* The Bayesian Information Criterion (BIC) is also called the Schwartz Bayesian Information Criterion or the Schwartz Information Criterion. The Consistent Akaike Information Criterion (CAIC) differs from the traditional (nonconsistent) AIC in that the CAIC incorporates sample size, and therefore is more consistent than the AIC as the sample size increases. Hamparsum Bozdogan, *Model Selection and Akaike's Information Criterion (AIC): The General Theory and Its Analytical Extensions*, 52 PSYCHOMETRIKA 345, 358 (1987).

598. Dominique M. A. Haughton, Johan H. L. Oud & Robert A. R. G. Jansen, *Information and Other Criteria in Structural Equation Model Selection*, 26 COMMC'N STAT.—SIMULATION & COMPUTATION 1477, 1500–09 (1997) (showing the AIC and BIC significantly outperform traditional fit indices, including the Normed Fit Index (NFI), when selecting the most appropriate structural equation model).

599. The full model includes a system of equations where LGPA is modeled as a function of UGPA, LSAT, tier location and gender; first-time bar passage is modeled as a function of UGPA, LSAT, tier location, gender, and LGPA.

600. Both the BIC and Adjusted BIC (ABIC) account for sample size and model complexity, but the ABIC penalty is not as high as the BIC. *See* Stanley L. Sclove, *Application of Model-Selection Criteria to Some Problems in Multivariate Analysis*, 52 PSYCHOMETRIKA 333, 335 (1987).

results: 20103.6 for the model with tier location included and 16185.349 for the model excluding tier location. Similar results were obtained when comparing models excluding both LGPA and tier location: The simpler models fit the data better.[601]

My reanalysis of Sander's model underscores the implausibility of a mismatch effect in law schools. Sander's model has very poor predictive power of both LGPA and first-time bar passage, as revealed by another popular measure of model fit that assesses the predictive accuracy of models: the coefficient of determination (R^2).[602] The R^2 statistics are 26.3 percent and 21.9 percent for the model predicting LGPA and first-time bar passage, respectively.[603] In other words, the model explains approximately one-quarter of the variability in LGPA and one-fifth of the variability in the first-time bar passage rate.[604] The *overall* R^2, which accounts for the fact that LGPA is both a dependent variable and an explanatory variable in the first-time bar passage model, is 34.6 percent. The paltry explanatory power of LGPA presents a major challenge for Sander's analytical model. Recall that, according to Sander's theory, "learning mismatch" is the primary mechanism through which outmatched students are harmed by affirmative action, yet UGPA and LSAT only explains a quarter of the variance in LGPA in this model. This problem is not unique to the BPS data—studies conducted at the law schools at UC Berkeley, Columbia University, UMLS, and the University of Pennsylvania reveal similar results: UGPA and LSAT explain less than a quarter of the variation in LGPA.[605] According to the LSAC, UGPA and LSAT explain, collectively, about a quarter of the variance in first year grades at all

601. CAIC = 16336.348; BIC = 16336.641; ABIC = 16308.053.

602. *See* LESLIE A. HAYDUK, LISREL: ISSUES, DEBATES, AND STRATEGIES 190–93 (1996) (explaining coefficients of determination in mediation models).

603. The publicly available BPS data do not include a measure of undergraduate institution selectivity, but as noted earlier, *see supra* note 181, additional analyses conducted by the LSAC on the full data reveal that college selectivity, as measured by the Astin Index, was only weakly correlated with LSAT, LGPA, and bar passage, and as a result, did not appreciably contribute to the predictive ability of the model. *Supra* note 181.

604. These models include UGPA, LSAT, LGPA, and tier location in the model predicting first-time bar passage, and UGPA, LSAT, tier location, and gender in the model predicting LGPA. Technically speaking, the R^2 ("coefficient of determination") is an improper measure of model fit for a binary variable, such as first-time bar passage. I report this measure because Sander employs a linear probability model (LPM), so R^2 measure is the standard fit statistics for this model.

605. Grutter v. Bollinger, 137 F. Supp. 2d 821, 861 (E.D. Mich. 2001) (University of Michigan), *rev'd*, 288 F.3d 732 (6th Cir. 2002); Kidder, *supra* note 299, at 1101 (UC Berkeley); James C. Hathaway, *The Mythical Meritocracy of Law School Admissions*, 34 J. LEGAL EDUC. 86 (1984) (reporting on Columbia University); GUINIER ET AL., *supra* note 495 (reporting on University of Pennsylvania).

ABA-accredited law schools.[606] Moreover, the predictive power of LSAT for grades decreases in both the 2L and 3L years, and deprecation of predictive power of the LSAT is more pronounced for black and minority students.[607]

Sander's use of inappropriate model fit statistics that suggest a stronger fit between this hypothesized model and the observed data is not limited to his mediation analysis. As I discuss in Subpart II.C, Sander also reports a misleading model fit statistic, Somers's *D*, when describing results from his nonmediation binary regression model of bar passage, and does not present additional measures of model fit despite strong warnings in social science literature about the limitation of Somers's *D* for the type of data that Sander analyzes—a problem that did not go unnoticed by his critics.[608] Chambers and colleagues highlight that Sander fails to "present a range of diagnostics that would have suggested the shakiness of its statistical foundations,"[609] and he reports "only the Somers's *D* statistic [for his] logistic [binary] regression results"[610] but "given that about 95% of those who took the bar passed, Somers's *D* presents a misleading portrait of how the model does."[611]

As the above discussion highlights, the statistical significance of the tier location effect (as well as other explanatory variables) is secondary to the larger question of practical significance. Statistical significance indicates that a particular effect observed in sample data drawn from a larger population is unlikely due to chance and, therefore, would be observed in the larger population (with some degree of imprecision); it does not tell us, however, whether the size of the effect is meaningful or whether the variable does an adequate job of explaining variation in the dependent variable(s). Model fit statistics are helpful in this regard, but only theory or policy considerations can perform that ultimate function. As such, the extremely modest explanatory power of Sander's models for the observed variability in LGPA and bar passage rates brings into serious question the relevance of mismatch theory in understanding academic performance and post–law school outcomes. And this critique holds true even when granting Sander's

606. Wightman, *supra* note 161, at 32.
607. Hathaway, *supra* note 605, at 91; Henderson, *supra* note 299, at 1029, 1051; Kidder, *supra* note 299, at 1101 (citing to a study conducted by LSAC at the request of Berkeley Law School showing that the correlation between academic index scores and law school grades dropped each year for black students (from 0.50 to 0.26 to 0.11)).
608. *See supra* text accompanying notes 204–207 (describing the shortcomings of the Somers's *D* statistic and presenting an appropriate alternative).
609. Chambers et al., *supra* note 68, at 1871.
610. *Id.*
611. *Id.* at 1872.

implausible assumptions about unconfoundedness for his hypothesized causal impact of tier location.[612]

B. Sequential Analysis

Sander attempted to augment his mediation model analyses with alternative model specifications in an effort to demonstrate that his results were not impacted by the various problems that Ho identified in his critique of *Systemic Analysis*. Sander explains that his:

> [R]eply [to Ho] showed one method—structural equation modeling [or mediation analysis]. Another very good (but less intuitive) method . . . [consists of] a series of [regressions], all of which attempt to predict who does, and does not, pass the bar on the first attempt. . . . Each model adds or subtracts variables to assess the marginal importance of different factors on bar passage. [The] approach also provides insight into how the variables interact with one another.[613]

Sander claims that "[t]hese models provide an even better demonstration of the story in *Systemic Analysis* than does the original Table 6.1 [in *Systemic Analysis*]."[614] Specifically, he reports that tier location, its interaction with UGPA (TIER × UGPA), and its interaction with LSAT (TIER × LSAT) are all statistically insignificant in his models before controlling for LGPA.[615] According to Sander, these results support the mismatch thesis because the "inclusion of [LGPA] in a model with Tier does not obscure some positive, indirect effect of Tier on bar passage [and] a model that omits GPA altogether but controls for the background credentials of students essentially eliminates any impact from Tier."[616]

I reanalyze Sander's supplemental analysis and my results contradict both of those conclusions.[617] With respect to his latter claim—the models

612. *See supra* Subpart II.A.2 (testing the unconfoundedness assumptions with the BPS data).
613. Sander & Doherty, *supra* note 80, at 1. *But see* Jonah B. Gelbach, *When Do Covariates Matter? And Which Ones, and How Much?*, 34 J. LAB. ECON. 509 (2016) (discussing several shortcomings of the sequential analysis framework).
614. Sander & Doherty, *supra* note 80, at 2.
615. In the statistics literature, an interaction between two or more variables is typically depicted as a product (multiplication) between the variables, such as A × B or A · B.
616. Sander & Doherty, *supra* note 80, at 2–3.
617. I was unable to exactly reproduce Sander's results for the supplementary analysis. The effects of race/ethnicity, UGPA, LGPA, and LSAT are essentially the same, but the effect of tier location and its interaction with UGPA and LSAT substantially differ from those reported by Sander. I also label my tier location variable differently from his analysis.

omitting LGPA, but controlling for UGPA and LSAT, eliminate any effect of tier location—his conclusion is unreliable for two reasons. First, his interpretation of the marginal effects of tier location is nonsensical because the interaction term in his model changes their meaning from an overall marginal effect to the conditional marginal effect when UGPA and LSAT have values of zero. Second, his clearly erroneous assumption that the respondents in the BPS are independent within tiers when calculating tests of statistical significance.

Sander also claims that race/ethnicity exerts no independent direct effect on first-time bar passage after taking into account LGPA, but this assertion is refuted by the data when the analysis is conducted properly. Race/ethnicity exerts a direct effect on LGPA (controlling for UGPA, LSAT, and tier location) and an indirect effect on first-time bar passage, operating through LGPA. These two processes result in a total effect of race/ethnicity on first-time bar passage that cannot be attributed to mismatch. I discuss each problem, in turn.

1. Interpretation of Marginal Effects

Sander interprets the marginal effects[618] of tier location inappropriately for his logit models (binary regression models) including interaction terms between tier location. A statistical interaction occurs when the effect (regression coefficient) of one explanatory variable on the dependent variable changes depending on the level of another explanatory variable. When an analyst examines an interaction effect between two (or more) variables in a statistical model, one typically includes both the constituent variables (such as tier and UGPA) and the product of the two variables (TIER × UGPA). The conditional relationship is then assessed by combining the "simple effect" of one of the constituent variables with the interaction effect. A simple effect is the effect of the constituent variable at a particular level of the other component variable in the interaction.[619] When there is no interaction term included in the model, the effect of a constituent variables represents a "main effect," which is the effect of the particular component variable on the

Sander lists the elite tier as "Tier 6," the next elite tier as "Tier 5," and so on. I, in contrast, list the elite tier as "Tier 1, the next elite tier as "Tier 2," and so forth. This does not change the substantive results, but it is important to keep in mind when analyzing Sander and Doherty's tables.

618. A marginal effect provides an approximation of the amount of change in the outcome variable that will be produced by a one-unit change in the explanatory variable. CAMERON & TRIVEDI, *supra* note 130, at 122.

619. *See* Robert J. Friedrich, *In Defense of Multiplicative Terms in Multiple Regression Equations*, 26 AM. J. POL. SCI. 797, 810 (1982).

dependent variable, ignoring the potential conditioning effect of the other component variable. It is customary to mean-center component variables at "zero" before creating multiplicative interaction terms to assist in interpreting the simple effects.[620] Mean-centering entails subtracting the mean value of a variable from each instance of the variable. The resulting values represent a deviation from the average value. Mean-centering is similar to standardizing, but the mean-centered variable retains its natural scale, and therefore avoids the interpretative problems that standardization presents. When variables are centered at zero, then the simple effect is the effect of the component variable when the other variable is at its average value.[621] If component variables have natural zeros, then mean-centering may not be necessary; but if the value of zero is nonsensical for a variable, then the simple effect will be nonsensical as well. Usually this can be easily detected in the statistical results because the parameter estimates for one (or both) of the component variables drastically changes (sometimes changing direction) and the standard error becomes extremely large.[622] Both of these issues are present in Sander's analysis.

The coefficients that Sander reports for the tier location contrasts represent the differences in the likelihood of first-time bar passage (on the logit scale) for each tier compared to the top tier for students in those tiers for whom both UGPA and LSAT equal "zero." Obviously, a value of zero is meaningless for these two variables because the tier location effect Sander is measuring applies to individuals who do not exist in the data (or at any law school). The minimum UGPA value for students in the model is 1.5, and 99.2 percent of the students have a UGPA over 2.1. Similarly, the minimum LSAT for students included in Sander's model was 11, and 99 percent of the students have a LSAT score over 22. Sander's methodological error is akin to extrapolation bias—specifying a model that, essentially, pretends the existence of data values that are not available, so the parametric assumptions are less constrained by the data.[623] When Sander's interpretive error is corrected, two findings emerge that were incorrectly suppressed. First, in the model excluding LGPA and including entering credentials (UGPA and LSAT), there is a statistically significant difference in first-time bar passage

620. *Id.* at 804.
621. *Id.*
622. *See id.* at 803.
623. *See supra* Subpart II.E.

rate when comparing students between the Tier 1/Tier 5, Tier 2/Tier 3, Tier 2/Tier 4, Tier 2/Tier 5, Tier 3/Tier 5, Tier 4/Tier 5, and Tier 5/Tier 6.[624]

Sander's model including LGPA suffers from the same problem. Statistically significant differences exist (again, when UGPA and LSAT are set to average values) between the Tier 1/Tier 2, Tier 1/Tier 4, Tier 1/Tier 5, Tier 1/Tier 6, Tier 2/Tier 3, Tier 2/Tier 5, Tier 2/Tier 6, Tier 3/Tier 4, Tier 3/Tier 6, Tier 4/Tier 5, and Tier 4/Tier 6 comparisons. These results indicate that tier location exerts a statistically significant impact on the probability of first-time bar passage even after controlling for UGPA, LSAT, and LGPA. They also reveal that the tier location effects become more pronounced in the model adjusting for LGPA—a finding that directly contradicts the central claim of mismatch theory. These results underscore how a simple and widely recommended rescaling of the constituent variables (mean-centering)—which does not change the association between the variables in the model, but only the interpretation of the simple effects—provides a very different picture of the mismatch effects. It should be emphasized that these statistically significant differences in tier location are not simply an artifact of exploring these differences for students with typical values for UGPA and LSAT. The models both excluding and including LGPA were reevaluated for students with entering credentials at the 25th and 75th percentiles and many of the tier location comparisons remained statistically significant.

Sander's interpretive error is easily identifiable based on large changes in the magnitude or direction of parameter estimates, as well as a large increase in the standard error of the estimate—these are typical indications of model instability do to extrapolation.[625] The difference in the tier location parameter estimates in Sander's models excluding and including the interactions between tier location and UGPA/LSAT are dramatic, changing direction (from negative to positive) for Tier 5 (from -0.846 to 0.772) and Tier 6 (from -1.516 to 0.981), and decreasing by 73 percent for Tier 4 (from -0.328 to -0.088).[626]

Sander's decision to not rescale the UGPA and LSAT variables to facilitate substantively meaningful and empirically supported interpretations of the tier location effects in his sequential analysis is particularly odd because he reports

624. Sander & Doherty, *supra* note 80, at 4. There is also a marginally significant (p = .07) effect between the first (top) and fourth tiers. Sander and Doherty only report tier contrast with the top tier, but pairwise comparisons between all of the tiers are equally as important because, according to the authors, their "model that omits GPA altogether but controls for the background credentials of students essentially eliminates any impact from Tier." *Id.* at 2–3.

625. Friedrich, *supra* note 619, at 803.

626. The top tier is the comparison category. Sander & Doherty, *supra* note 80, at 4 tbl.1 (Models II & IV).

parameters estimates for mean-centered variables in his aforementioned mediation analysis, so he was aware that rescaling variables changes their interpretation.[627] Furthermore, Sander conducted the mediation and sequential analyses around the same time,[628] and Sander expressly references the sequential analysis in the mediation article.[629] The standardized effects reported in Sander's mediation analysis are the effects of standardized variables in the model, and the standardization process entails both mean-centering variables at "zero" and representing their values as standard deviation units.[630] Had Sander reported the standardized effects in his sequential analysis, as he did in his mediation analysis, he would have obtained results identical to what I have identified for the tier location contrasts.[631]

The standard errors for the parameter estimates of the tier comparisons also substantially increase in absolute size—from 0.134 to 1.266 (Tier 6), from 0.057 to 0.606 (Tier 5), 0.034 to 0.352 (Tier 4), 0.026 to 0.269 (Tier 3), and 0.022 to 0.225 (Tier 2).[632] But the issues with Sander's standard error do not solely stem from the incorrect scaling of the constituent variables of his interaction terms. The standard errors he reports are inappropriate due to his unwarranted and indefensible assumption that students within tiers are independent. I address this issue below.

2. Calculation of Standard Errors

Sander also assumes the independence of observations when calculating standard errors (and significance tests) for his sequential analysis. When the correct standard errors are used, the results directly contradict Sander's core conclusions. Specifically, Sander's model excluding LGPA reveals statistically significant effects for all of the tier pairwise comparisons (except for Tier 3/Tier 4 and Tier 4/6).[633] Including LGPA in the model reveals that all tier location

627. *See supra* Subpart V.A.1.
628. Both analyses were published in 2005.
629. Sander, *Mismeasuring*, *supra* note 24, at 2007 n.13.
630. *See supra* Subpart V.A.1 (discussing the interpretation of standardized effects).
631. A standardized interaction term consists of the product of the two standardized constituent variables, and not the standardization of the product of the two constituent variables. LEONA S. AIKEN & STEPHEN G. WEST, MULTIPLE REGRESSION: TESTING AND INTERPRETING INTERACTIONS 40–42 (1991).
632. Sander & Doherty, *supra* note 80, at 4 tbl.1 (Models II & IV).
633. I conducted post hoc adjustments to minimize the false positive rate resulting from multiple comparisons for the models excluding and including LGPA. *See supra* note 454 and accompanying text (discussing the importance of post hoc adjustment procedures in the presence of multiple comparisons).

pairwise comparisons are statistically significant, except Tier 1/3 and Tier 5/Tier 6. The Tier 1/3 comparison was marginally significant ($p = 0.075$). The statistically significant pairwise comparisons refute Sander's claim that UGPA, LSAT, and LGPA eliminate the impact of tier location on first-time bar passage. Therefore, Sander's models fail even when examining them on their own terms (controlling for LGPA) if the analyst uses statistically defensible standard errors.[634]

3. Race/Ethnicity Effects

Sander claims that his sequential analysis "implies that neither race, nor other unobserved factors that might be correlated with race (e.g., attendance at bar preparation courses) is strongly correlated with bar passage or failure"[635] after accounting for LGPA. The analytical framework he employs, however, is incapable of supporting the assertion.[636] In fact, the National Research Council (of the National Academies) unequivocally concluded that "measures of [racial] discrimination that focus on episodic discrimination at a particular place and point in time may provide very limited information on the effect of dynamic, cumulative discrimination . . . [and] the effects of cumulative discrimination can be transmitted through the organizational and social structures of a society."[637]

I examined the overall effect of race/ethnicity with Sander's sequential analysis framework by calculating the direct effect of race/ethnicity on first-time bar passage and the indirect effect of race/ethnicity, operating through LGPA. The hypothesized direct and indirect relationships can be represented with the specific segments of the causal chain: a direct effect (RACE→BAR) and an indirect effect (RACE→LGPA × LGPA→BAR). This yields an overall race/ethnicity effect that is a sum of the direct effect and indirect effects. The purpose of this analysis was not to posit any causal relationship between race/ethnicity and LGPA/bar passage;[638] rather, the inquiry was designed to determine the extent to which the cumulative effects of racial discrimination—roughly proxied by race/ethnicity—impact LGPA and first-

634. *See* King & Roberts, *supra* note 581, at 160.
635. Sander & Doherty, *supra* note 80, at 2.
636. Sander has also frequently reported that race/ethnicity has no statistically significant direct effect on first-time bar passage after controlling for LGPA in his nonmediation analysis models. *See, e.g.*, Sander, *Systemic Analysis*, *supra* note 7, at 444–46.
637. NAT'L RESEARCH COUNCIL, MEASURING RACIAL DISCRIMINATION 226 (Rebecca M. Blank et al. eds., 2004) (emphasis omitted).
638. *See supra* note 58 (describing the necessary assumption for the identification of causal effects with the framework of mediation); *supra* note 353 (explaining the problems associated with positing an immutable characteristic, such as race/ethnicity, as a causal variable).

time bar passage rates.[639] If Sander is correct that neither race nor unmodeled factors correlated with race impact success on the bar exam, then race/ethnicity should not exert an overall effect. Similar to Sander's sequential analysis, my analysis focuses on the entire sample of law students, and includes gender, UGPA, LSAT, race/ethnicity indicator variables for each group (with white students serving as the comparison group), tier location indicator variables for each tier (with Tier 1 serving as the comparison group), (TIER × UGPA) interaction, and (TIER × LSAT) interaction. In contrast to Sander, I use an appropriate nonlinear probability model for binary outcomes (probit) when examining the impact of the explanatory variables on first-time bar passage.

Four findings emerged. First, all nonwhite racial groups, compared to white students, have lower LGPAs. The strongest negative effect is found for black students, while the effects for Asian and Latinx students are about the same, and the smallest negative effect is for students designated "Other Race."[640] Second, when controlling for LGPA, Asian and "Other Race" students performed worse than white students on their first attempt at the bar, but there was no statistically significant difference for black or Latinx (when compared to white) students. Third, the indirect effect of race/ethnicity on first-time bar passage, operating through LGPA, is negative for all racial/ethnic groups, compared with white students, with the strongest effect for black students, the weakest effect for "Other Race" students, and Asian and Latinx students being very similar. Fourth, the total effect of race/ethnicity on first-time bar passage, which combines the direct and indirect effects, reveals the true magnitude of racial/ethnic differences in first-time bar passage—nonwhite students perform worse on the bar compared to white students. Interestingly, there were no significant differences between any of the nonwhite pairwise comparisons (such as Asian-black, Latinx-Other Race). The results are nearly identical when removing the HBLS tier from the analysis—which is to be expected because the estimates are based primarily on the white students who compose approximately 85 percent of the estimation sample—and when adjusting for students' undergraduate major

639. *See supra* note 477 and accompanying text (discussing research on the persistence of racial discrimination in the legal field); *see also* Ayres & Brooks, *supra* note 12, at 1829 ("However, there is more to 'race' than the race of the student; . . . one must wonder if better race controls (including the racial composition of the school, the classroom and the faculty) would reveal an even bigger impact of race . . . . ").

640. *See also supra* Subpart IV.C (discussing the persistence of racial/ethnic differences in LGPA between white and nonwhite students, even after adjusting for UGPA, LSAT, undergraduate major, and law school tier).

(Humanities; Social Sciences; Business; Science, Technology, Engineering, and Math (STEM); and "Other").[641]

It is unfortunate that Sander does not examine whether racial/ethnic differences exist with respect to an overall race/ethnicity effect using his mediation analysis framework and instead focuses solely on the overall tier effect because the analytical frameworks are nearly identical.[642] Indeed, inquiry into an overall tier effect naturally lends itself to investigate other types of overall effects when the model posits both direct and indirect effects operating through an intervening variable. To be sure, Sander's analysis of the overall tier effect is not on strong statistical footing because he models the effect of tier location linearly and employs a linear probability model, rather than a binary choice model, when investigating bar passage. Both of these modeling choices for his mediation analysis are inadvisable and my reanalysis of the overall race/ethnicity effect on bar passage avoids them.[643]

The purpose of this supplemental analysis is not to provide definitive answers about the effects of race/ethnicity on LGPA and bar passage; rather

641. I also reexamined the models specifying that part of the effect of race/ethnicity operated through tier location. *See, e.g.*, Williams, *supra* note 22, at 182–83 (noting law school admission preferences for all nonwhite groups); Rothstein & Yoon, *supra* note 21, at 16 (noting that black students were more likely to be admitted to highly selective law schools than white students with similar credentials). This model has the advantage of capturing the two types of racial/ethnic indirect effects—the first operating through tier location and the second operating through LGPA. The key limitation of the model is that one cannot simultaneously measure the (TIER × UGPA) and (TIER × LSAT) interaction effects. The results of this model were nearly identical to the models that did not specify a direct relationship between race/ethnicity and tier location. The removal of the HBLS tier did not appreciably alter the results.

642. *See supra* Subpart II.A.1.a (describing Sander's assessment of direct, indirect, and total tier effects in his mediation model); *supra* Appendix A.1 (same).

643. The results from a linear probability models—including and excluding the HBLS tier—were similar in terms of the direction of the effects and their statistical significance, but the magnitude of the effects for the racial/ethnic contrasts were much larger. For example, whereas the total effect for "black" on the probability of first-time bar passage from the nonlinear probability model was -0.059, the effect was -0.163 in the linear probability model—276 percent larger. For Latinx students, the total race/ethnicity effect from the nonlinear and linear probability models were, respectively, -0.054 and -0.101—a 187 percent difference. The linear probability model makes the implausible assumption that the impact of race/ethnicity on the probability of first-time bar passage is constant across other values of the explanatory variables in the model (UGPA, LSAT and tier location); the nonlinear probability model does not. Another way to understand it is that the linear probability model makes an incorrect functional form assumption, *see infra* Appendix C, and this problem becomes magnified when assessing direct and indirect effects. Despite differences in the relative magnitudes of the race/ethnicity effects, statistical significance of the racial/ethnic pairwise contrasts (such as Asian-black, Latinx-other, black-white) were the same across both nonlinear and linear probability models.

my aim is to show that, adopting a similar framework employed by Sander's, I find results that are inconsistent with his core conclusions about the importance of race/ethnicity in explaining racial differences in students' performance in law school and on the bar exam. My findings suggest that something other than mismatch is suppressing nonwhite students' achievement in law school (proxied by LGPA) and on the first attempt at the bar examination. This interpretation is buttressed by the fact that Asian and "Other Race" students perform worse than white students with respect to LGPA and first-time bar passage, yet according to Sander and Williams, black and Latinx students are most harmed by affirmative action because Asian and "Other Race" students benefit the least from these policies in law.[644]

C. Functional Form Misspecification

Recall that both interpolation bias and extrapolation bias arise when the model improperly specifies the interrelationships (functional form) between variables—in other words, the model incorrectly posits how variables covary.[645] In the context of causal analysis, these biases impact inferences about how a treatment variable affects another variable(s). A peculiar example of Sander's problematic modeling assumptions that is relevant to functional form misspecification appears in his analysis of the effect of attending a first choice law school on law school tier location (for black and white students, separately).[646] Sander states that the primary purpose of these models is to rectify problems apparent in Ayres and Brooks's second choice analysis. Sander employs an ordinary least squares (OLS) linear regression model; the OLS framework he employs, however, is inappropriate because his dependent variable is ordinal, rather than continuous. Modeling tier location as a continuous variable is inappropriate because it results in the violation of several of the core assumptions of OLS regression and leads to invalid statistical inferences.[647] Of particular concern for Sander's analysis is his

644. *See* SANDER & TAYLOR, *supra* note 4, at 86 (remarking that there is a large mismatch problem in law school that affects black and Latinx students); Williams, *supra* note 22, at 181 n.15 (claiming that Asians and "Other Race" students are more heterogenous in terms of entering credentials and receive less preferences compared to black, Latinx, and Native Americans students).

645. *See supra* Subparts II.D–II.E.

646. Sander, Whitepaper, *supra* note 4, at 7 tbl.2; Sander, *Replication of Mismatch*, *supra* note 4, at 77.

647. The ordinary least squares (OLS) model for an ordinal variable is not suitable for several reasons: (1) standard errors are inconsistent and hypothesis tests for effects of explanatory variables are invalid; (2) the model produces predicted values of the

implausible assumption that the effect of attending a first choice law school on tier location is uniform across the different tiers. Sander posits that a first choice selection has the same effect on the probability of attending a second tier school relative to a first tier school as it would on the probability of attending a third tier school relative to a second tier school, a fourth tier school relative to a third their school, and so forth. Ho made a similar observation in the context of Sander modeling the effect of tier location as an explanatory variable on bar passage rates[648] and Sander, himself, has acknowledged on multiple occasions that law school tier only roughly approximates the selectivity hierarchy.[649] When Sander uses OLS regression to estimate the differences in the tier of first choice school (relative to the first choice school), holding UGPA and LSAT constant, his interpretation of the results is nonsensical because it is impossible to tell how far down the respondent is in the tier hierarchy because the selectivity distance between the adjacent

dependent variable that are impossible; (3) the functional form specified by the OLS model will be incorrect because it is unlikely that extreme values of an explanatory variable will have the same effect on the response variable as more moderate values; and (4) misleading model fit tests. For an accessible discussion of these issues, see J. SCOTT LONG, REGRESSION MODELS FOR CATEGORICAL AND LIMITED DEPENDENT VARIABLES 114–47 (Richard Berk ed., 1997).

It is widely accepted that the OLS framework is inappropriate for most types of data that social scientists investigate. This point was recently underscored by two prominent political methodologists:

> Data in the social sciences tend to be lumpier, often categorical. Nominal, truncated, and bounded variables emerge not just from observational datasets but in researcher-controlled experiments as well (e.g., treatment selection and survival times). Indeed, the vast majority of social science data comes in forms that are profitably analyzed without resort to the special case of OLS. . . . *[Y]ou will have to look hard for recent, state-of-the-art empirical articles that analyze observational data based on this approach.*

MICHAEL D. WARD & JOHN S. AHLQUIST, MAXIMUM LIKELIHOOD FOR SOCIAL SCIENCE: STRATEGIES FOR ANALYSIS, at xviii (2018) (emphasis added).

648. Ho, *supra* note 7, at 2001 (noting that Sander's assumption that "going from a sixth- to a fifth-tier school is roughly the same as going from a second- to a first-tier school" is both "unjustified and unnecessary").

649. *See, e.g.*, Sander, *Systemic Analysis*, *supra* note 7, at 415; Sander, *Mismeasuring*, *supra* note 24, at 2006, 2009; Sander, *Reply to Critics*, *supra* note 10, at 1987. Similar to the problem of analyzing tier location as a continuous dependent variable, when tier location is modeled as a continuous independent variable, these models incorrectly specify the functional form of the relationship between tier location and bar passage and undermine the ability of the models to properly adjust for the relationships between tier location and all of the other variables in the model that are correlated with tier location. *See* Arcidiacono & Lovenheim, *supra* note 7, at 19 n.27 (noting that Sander improperly includes the tier location variable as a linear predictor, but the proper approach to modeling the effect of law school tier is to include an indicator variable for each tier).

tiers is not uniform across tiers.[650] In statistics parlance, Sander uses a model for interval data when, in fact, his data are ordinal.[651] An interval-level regression model (for example, an OLS model) assumes the size of the differences between the categories are substantively meaningful, whereas an ordinal-level regression model does not. An ordinal regression model allows the analyst to answer a different type of question: Given the respondent's UGPA and LSAT and first choice selection, what is the probability that the respondent is enrolled in a specific tier relative to all of the lower tiers (for example students in Tier 1 versus Tiers 2–6, followed by Tier 2 versus Tiers 3–6, and so on)?

Alice Xiang and Rubin examined various aspects of the mismatch hypothesis using the BPS data and employed the appropriate ordinal regression framework to analyze the effect of race (black versus white) on the probability of enrolling in the various law school tiers, controlling for UGPA and LSAT.[652] In contrast to the OLS model, the ordinal regression model specifies a different (and more plausible) functional form between the independent variables and the probabilities of attending a law school in each of the six tiers.[653] This last point merits further discussion because Sander's implausible functional form assumptions appear throughout his scholarship on mismatch effects. Based on the incorrect assumption of the OLS model, the first choice effect on tier location is constant across all levels of UGPA and LSAT; under the more realistic assumptions of the ordinal regression model, however, the true marginal effect of first choice will differ depending on the particular UGPA and LSAT of the students. Consequently, it is likely that the marginal effect of first choice will be close to zero for a nontrivial number of the respondents in the data because the marginal effect will be much smaller near the lower or upper segments of the UGPA and LSAT distributions. This nonconstant marginal effect results from the ordinal nature of the tier location variable and the need for a nonlinear probability model, which

650. Sander, Whitepaper, *supra* note 4, at 7 ("The estimates suggest that for blacks, attending a second-choice school means going to a school that is perhaps a quarter to a third of a tier 'lower' in the hierarchy than one's first-choice school. Going to a third-choice school means going to a school that is perhaps half a tier lower."). The problem with Sander's interpretation is compounded by the fact that, as explained above, *see supra* Subpart II.F.1.a, Tier 2 and Tier 3 are indistinguishable along mismatch-relevant characteristics.

651. *See supra* note 293 and accompanying text (describing the four different levels of measurement).

652. Xiang & Rubin, *supra* note 425, at 303–08 (reporting that class-based affirmative action in law school would reduce racial diversity, but not improve academic outcomes of black students).

653. GELMAN & HILL, *supra* note 349, at 119.

properly accounts for the fact that the effect of first choice depends on the values of the other variables in the model.[654]

A similar concern is present when examining Sander's analysis of the impact of first choice on first-time bar passage.[655] Sander reports:

> By our calculation, choosing not to attend one's first-choice school raises the typical black student's chance of passing the bar from 65% to 78%.
>
> . . . .
>
> . . . [A] typical first-choice African-American has a 61.3% chance of graduating and passing the bar on the first attempt; for otherwise comparable third-choice students, the success rate is 85.9%.[656]

This alleged 13 percentage point difference (78 percent versus 65 percent) in first-time bar passage rates for black law students electing to enroll in their second or third choice school formed the centerpiece of the empirical discussion in the "Debate on Law School Mismatch" chapter in *Mismatch*, and Sander and Taylor identify Williams's article on mismatch effects in law school as the source of these findings.[657] It is important to emphasize that Sander analyzes the binary variable of "pass/fail" with a logistic regression model and states that the mismatch effect calculations are based on the "typical" black student, which is the average black student with respect to their UGPA, LSAT, tier location, and number of submitted law school applications. But Sander's results are misleading. As previously discussed, he neglects to mention that the effect of first choice on the probability of bar passage varies across the range of the values of the variables in his model.

So, for example, when focusing on students with values for UGPA, LSAT, and number of submitted law school applications at the 75th percentile of the distribution of those variables (rather than to their "typical" values), the overall impact of a black student attending a law school other than their first choice is a 7.5 percentage point differential in the probability of passing the bar exam on the first attempt—nearly half of the effect that Sander reports. For students with values at the 90th percentile of UGPA, LSAT, and number

654. CAMERON & TRIVEDI, *supra* note 130, at 122 ("[Unlike linear regression models,] [f]or nonlinear regression models . . . [the marginal effect] is a function of both parameters and regressors, and the size of the marginal effect depends on [both].").

655. Sander, *Replication of Mismatch*, *supra* note 4, at 80 tbl.6.

656. *Id.* at 80.

657. SANDER & TAYLOR, *supra* note 4, at 80 fig.5.1. These findings reported in *Mismatch* are susceptible to the same critiques of Williams's article that are described throughout in this Article. *See supra* text accompanying notes 131–135, 279–285, 321–323, 399–403.

of submitted applications, the impact of first choice is a 3.1 percentage point difference. In the opposite direction, for students at the 25th and 10th percentiles of the distribution of the aforementioned variables, the impact of enrolling in one's nonfirst-choice increases the bar passage rate by, respectively, 18 and 16 percentage points. This dynamic holds true when separating second choice from third choice students and examining those who elected to attend their third choice law school: The percentage point differentials drop to 8.3 and 4.2 at, respectively, the 75th and 90th percentiles of the UGPA, LSAT, and number of submitted applications distribution. At the 25th and 10th percentiles of UGPA, LSAT, and number of submitted applications, the differentials are 33.5 and 36.2 percentage points, respectively. It is obvious from these results that the impact of attending a lower choice law school on first-time bar passage depends, heavily, on where the student is situated with respect to their entering credentials.[658]

One might reason that the marginal effect of attending a first choice law school for the typical black student is still informative, but this ignores another shortcoming of Sander's first choice analysis: Students in Tier 2 and Tier 3 are indistinguishable with respect to UGPA, LSAT, and law school selectivity.[659] This is extremely important because the typical black student in Sander's analysis has a UGPA, of 3.02, a LSAT of 31.4, and is located in Tier 3. The BPS data do not contain information about the tier location of the student's first (or second) choice and it is questionable whether a student's first (or second) choice is likely to be a school in a higher tier than the student's second (or third) choice;[660] nonetheless, even if we were to assume, *arguendo*, that the typical similarly situated first (or second) choice black student was located in the adjacent higher tier, which would be Tier 2, then the increased bar passage rate for the second (or third) choice student in Tier 3 would not be attributable to mismatch-related

658. Similar to his mediation analysis, *see supra* Appendix A.1, Sander engages in misleading reporting of the marginal effects of first choice in his tables. Specifically, the coefficients he reports in the primary tables for the first choice analysis are the marginal effects on the natural logarithm (*ln* or $\log_e$) of the odds ("log odds" or "logit") of passing the bar exam on the first attempt. Sander, *Replication of Mismatch*, *supra* note 4, at 80 tbl.6. It is standard practice in the social sciences to report the odds ratio, which represents the odds of passing the bar over the odds of failing the bar for students electing to enroll in their second or third choice law school, rather than the log odds. Not only are the effects on the log odds difficult to interpret because they are not on the natural metric of the variable of interest, but in the context of Sander's work, the log odds imply a much stronger effect size for the second or third choice variable than the odds ratio (0.76 versus 0.47). *See, e.g.*, King et al., *supra* note 211, at 347–48 (imploring social scientists to report "substantively informative" interpretations of key quantities of interest).

659. *See supra* Subpart II.F.1.a.

660. *See supra* Subpart II.F.1.b.

dynamics. Moreover, the statistically significant 13 percentage point difference in first-time bar passage for the typical student is also a function of Sander incorrectly modeling the tier location effect linearly.[661] The fact that Sander, himself, fails to find significant differences in first-time bar passage rates for black students when he relaxes the linearity assumption and compares students in adjacent tiers reinforces this point.[662]

It must be emphasized, however, that inferences drawn from Sander's analysis are implausible because of the lack of comparable students in the BPS data. As explained earlier, there are very few students who are identical on the variables included in Sander's models, save for the second (or third) choice variable. When limiting the analysis to students with acceptable covariate balance (students who actually have comparables in the data), the effect of the second (or third) choice variable is indistinguishable from zero, irrespective of whether the students are observed at the 10th, 25th, 50th, 75th, or 90th percentile of the distribution on the relevant variables.[663] As I have shown, the only plausible (and data-driven, as opposed to completely model-driven) comparisons are those involving students in adjacent law school tiers.[664] When investigating potentially positive and deleterious impact of race-conscious admissions practices, one need look no further than to the sage advice offered by Shpitser and Pearl: "[O]nly by taking counterfactual analysis seriously is one able to distinguish testable from untestable counterfactuals, then posit the more advanced question: what additional assumptions are needed to make the latter testable."[665]

661. *See supra* note 648 and accompanying text (describing Sander's error in modeling the tier location effect linearly).
662. Sander, *Replication of Mismatch*, *supra* note 4, at 87 tbl.15.
663. After covariate balancing, the largest number of plausible comparable students was forty-nine across the six law school tiers (twenty-three second/third choice and twenty-six first choice) out of 469 students. This underscores the strong model dependence on which Sander's 13 percentage point second choice bar passage differential rests. It is also worth noting that I employed a covariate balancing approach that maximized the number of comparable students at the expense of achieving the closest possible comparisons. I examined the robustness of my results by using balancing approaches that achieved even closer comparisons, but effective sample sizes varied between eight and thirty-three students. My analyses of these small samples produced similar results to the ones I reported with the larger sample of forty-nine. This problem is magnified when focusing on Sander's disaggregated second and third choice models. There are only twenty-five third choice black law students who took the bar examination that are included in Sander's models, and only ten comparable students (five third choice and five first choice) out of 371 students in the sample. *See also supra* Subpart II.E (discussing the almost complete lack of plausible counterfactuals in Williams's first/second choice analysis after employing balancing methods on the less than ten variables he uses in his analysis).
664. *See supra* Subparts II.D–II.E.
665. Shpitser & Pearl, *supra* note 216, at 359.