UCLA School of Law

PUBLIC LAW & LEGAL THEORY RESEARCH PAPER NO. 22-12
LAW & ECONOMICS RESEARCH PAPER NO. 22-03

WHEN OLD HABITS DIE HARD:
A COMMENT ON SANDER AND STEINBUCH'S "MISMATCH AND BAR PASSAGE"

BY

SHEROD THAXTON
PROFESSOR OF LAW

In 72 JOURNAL OF LEGAL EDUCATION 382 (2024)

When Old Habits Die Hard: A Comment on Sander and Steinbuch's "Mismatch and Bar Passage"

Sherod Thaxton*

Contents

***Sherod Thaxton**, J.D., Ph.D., is Professor of Law and, by courtesy, of African American Studies, Public Policy, and Sociology at the University of California, Los Angeles. Professor Thaxton is one of the eight signatories of the Empirical Scholars Brief filed in *Students for Fair Admissions v. President and Fellows of Harvard College & University of North Carolina, et al.* (2023). He thanks Cheryl Harris, Richard Lempert, Intisar Rabb, the anonymous reviewers, and the editorial staff for helpful conversations and comments.

I. Introduction

Mismatch theory—which posits that race-conscious admissions policies harm racial minorities by admitting students into challenging schools where they cannot succeed—has figured prominently in the debate over affirmative action during the past half-century.[1] The theory was heavily emphasized by dissenting Supreme Court justices to the two notable cases of this century that narrowly upheld affirmative action in 2003 and 2016. And the theory fueled the renewed and ultimately successful bids in 2023 to eliminate affirmative action from university and law school admissions among both petitioners and an earlier dissenting-turned-majority voting justice.[2] The challenges to Harvard University's and the University of North Carolina's consideration of race as a factor in admissions decisions argued, *inter alia*, that universities "can eliminate this harmful mismatch and allow students to excel at schools for which they are most prepared by eliminating the use of racial preferences."[3]

The Court's most recent decision striking down affirmative action ironically bolstered the germaneness of mismatch theory to the mandate placed on universities to devise new admissions policies in the aftermath of that decision. The Court ruled that race-conscious admissions systems did not satisfy strict scrutiny, thereby violating the Equal Protection Clause of the Fourteenth Amendment. But it also was careful to note that the Constitution does not prohibit universities from considering how an applicant's race may have impacted her life, provided the impact is firmly connected to attributes of the applicant that a university deems valuable. In the aftermath of the Court's ruling, admission systems are permitted to look beyond numeric credentials—such as grade point averages and standardized test scores—when evaluating

1 *See* Sherod Thaxton, *How Not to Lie About Affirmative Action*, 67 U.C.L.A. L. Rev. 834, 856 (2020) ("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 recurring theme of arguments against affirmative action ever since.").

2 In *Fisher v. University of Texas (Fisher II)*, Justice Antonin Scalia opined that Black students admitted to a highly selective school, such as the University of Texas, would do better at "a slower track school . . . [because] they do not feel that they're—that they're being pushed ahead in—in classes that are too—too fast for them." Transcript of Record at 67, *Fisher v. University of Texas,* 579 U.S. 365 (2016). Over a decade earlier, in his dissenting opinion in *Grutter v. Bollinger*, Justice Clarence Thomas stated that University of Michigan Law School's affirmative action policy "tantalizes unprepared students with the promise of a University of Michigan degree and all of the opportunities that it offers. These overmatched students take the bait, only to find that they cannot succeed in the cauldron of competition." 539 U.S. 306, 372 (2003) (Thomas, J., dissenting).

3 Complaint at 93, *Students for Fair Admissions, Inc. v. President & Fellows of Harvard College*, 397 F. Supp. 3d 126 (D. Mass. 2014) (No. 14-cv-14176) (claiming that race-conscious affirmative action produces a mismatch effect); Complaint at 49, *Students for Fair Admissions, Inc. v. University of North Carolina et al.*, 567 F. Supp. 3d 580 (M.D.N.C. 2014) (No. 14-cv-954) (same); *see also Students for Fair Admissions, Inc. v. President & Fellows of Harvard College (SFFA)*, 143 S. Ct. 2141, 2197 (2023) (Thomas, J., concurring) (endorsing the view that "large racial preferences for black and Hispanic applicants have led to a disproportionately large share of those students receiving mediocre or poor grades once they arrive in competitive collegiate environments").

applicants.[4] Thus, the Court's rulings in *SFFA* did not render the debate over mismatch irrelevant. In fact, given that "merit" must be understood capaciously, understanding student performance across various academic settings and how that performance relates to post-graduation outcomes may be more important than ever as universities are reevaluating their admissions systems.

The persistence of the idea of mismatch in both policy discussions and litigation is puzzling given that the empirical evidence in favor of mismatch not only is sparse[5] but has been deemed unreliable by the vast majority of analysts who have rigorously investigated the matter.[6] It might seem that, "with empirical resolution [seemingly] impossible, scientific inquiry is replaced by debate."[7] Richard Sander, perhaps the most well-known proponent of mismatch, referred to critics of mismatch as "ideological Zealots who cling to conventional affirmative action policies with almost religious fervor and see their attacks on mismatch as a sort of holy war."[8] Yet the consistent critique

4 *SFFA*, 143 S. Ct. 2175. Writing for the majority, Chief Justice John Roberts emphasized that "nothing in this opinion should be construed as prohibiting universities from considering an applicant's discussion of how race affected his or her life, be it through discrimination, inspiration, or otherwise." *Id.*

5 Brief of Empirical Scholars as *Amici Curiae* in Support of Respondents at 9, *Students for Fair Admissions, Inc. v. President and Fellows of Harvard College and University of North Carolina, et al.*, 143 S. Ct. 2141 (2023) (Nos. 20-119 & 21-707) [hereinafter Empirical Scholars Brief in *SFFA*] ("The mismatch hypothesis is premised on data that [Richard] Sander and a handful of others claim demonstrate that affirmative action policies cause negative outcomes for URG [underrepresented racial group] students."); *see also infra* note 74 and accompanying text.

6 Empirical Scholars Brief in *SFFA*, *supra* note 5 ("It would be a serious mischaracterization of the evidence to suggest that the mismatch hypothesis is a consensus view among social scientists. The data do not support key claims made by mismatch proponents, and their arguments are not based on reliable scientific principles."); *see also* Thaxton, *supra* note 1, at 968 (explaining that a large body of social science research has reached conclusions that directly contradict mismatch theory); *accord* Richard O. Lempert, *The Supreme Court Has Upheld Affirmative Action. So Let's Dump Mismatch Theory*, N.Y. Times (June 23, 2016) ("Despite its fragile empirical foundation, the [mismatch] idea has received outsize[d] attention . . . . A careful review of the data shows that affirmative action benefits minority students. It does not mismatch them . . . . The few empirical studies referenced by people who defend the mismatch theory do not hold up to critical scrutiny."); Anthony P. Carnevale et al., *The Concept of "Mismatch" at Play in the Supreme Court* Fisher *Decision is Empirically Unsound*, Georgetown University Center on Education and the Workforce 1 (2016) ("When average students are placed in the nation's best colleges and universities, they will graduate at a much higher rate. Rather than being intimidated by not being able to meet the standards of their peers . . . these students are instead challenged by the circumstances, and succeed at a rate comparable to their peers."); *see infra* note 74 and accompanying text.

7 Charles F. Manski, *Identification Problems in the Social Sciences*, 23 Soc. Methodology 1, 2 (1993).

8 Richard Sander, *The Stylized Critique of Mismatch,* 92 Tex. L. Rev. 1637, 1641 (2014) [hereinafter Sander, *Stylized Critique*]. In responding to critics of mismatch theory, Sander simultaneously targets his critics with *ad hominem* attacks and claims to be the target of *ad hominem* attacks. He characterizes those critical of his empirical research as "affirmative action Zealots," "anti-mismatch Zealots," "Zealots lack[ing] intellectual confidence," "deliberately misleading,"

is the dubious evidentiary support for mismatch. These debates are not mere ideological disagreements,[9] nor are they simply about discursive tone.[10]

The primary criticism of the various studies reporting support for mismatch is that they suffer a vast array of recurring theoretical and methodological shortcomings. And these critiques have been levied by a myriad of the most prominent statisticians and social scientists in academia.[11] In stark contrast to

"completely clueless," and "incapable of fairly summarizing what any piece of mismatch scholarship shows," and he describes their work as "a Case Study in Zealotry," "a polemic authored by Zealots," "fraudulent to its core," "silly, rather than merely erroneous," and "ideology wearing empirical makeup." Sander states that critiques of him are "a completely non-substantive, ad hominem, and unfair attack," and they avoid "honest debate on disputed points," and "a sort of ad hominem shell game;" he further accuses multiple critics of "imply[ing] that [he] was not-quite-all-there mentally." Sander, *id*. at 1638, 1646, 1650, 1658, 1662, 1664; Richard Sander, *Mismatch and the Empirical Scholars Brief*, 48 Val. U. L. Rev. 555, 558, 580, 583, 584 (2014) [hereinafter Sander, *Mismatch and the ESB*]; *see also infra* note 28.

9 Thaxton, *supra* note 1, at 840–45; *see infra* notes 69–71, 73–74 and accompanying text. *See generally* Manski, *supra* note 7, at 2 ("Disagreements presumably arise out of the conflicting self-interests and ideologies of the population, but normative differences are not the only contributing force. Many controversies reflect divergent beliefs about human behavior, specifically about the effects of government programs on behavior.").

10 There is an ongoing debate in academic circles over the appropriate level of humility and civility that should govern scholarly criticism (and responses to that criticism)—referred to as the "tone debate." Maarten Derksen & Sarahanne Field, *The Tone Debate: Knowledge, Self, and Social Order*, 26 Rev. Gen. Psych. 172 (2022). Much of the focus of this debate involves controversies over unsuccessful attempts to replicate findings that have been previously published in scholarly journals because "the inability to reproduce earlier results tends to raise questions about the ability and/or morality of the original researchers or the replicators." *Id*. at 180; *accord* Michael S. Weisbach, The Economist's Craft: An Introduction to Research, Publishing, and Professional Development 103 (2021) ("Replication is an important part of the scientific process, and there has recently been much controversy in the social sciences about papers that cannot be replicated."). Brown and colleagues explain, "In trying to counter the points of some authors or studies, some individuals resort to *ad hominem* arguments, often trying to undermine the credibility of arguments by attacking a person based on perceived expertise or presumed motives." Andrew W. Brow et al., *Issues with Data and Analyses: Errors, Underlying Themes, and Potential Solutions*, 115 Proc. Nat'l Acad. Sci. 2563, 2567 (2018). Other scholars are skeptical about the concern with tone, arguing that the tone debate is a "derailment tactic" used by purveyors of bad science to divert attention away from real issues in knowledge production in the sciences, such as replication, questionable research practices, and publication bias. Derksen & Field, *supra* note 6, at 180. There are numerous examples of individuals going beyond commenting on the work itself to criticizing the author(s), as well as countless instances of an author(s) characterizing criticism as personal attacks even when it is based solely on scientific grounds. Perceptions of discursive tone are complicated and can reside, in part, in the eye of the beholder. *Id*. at 175 ("The dividing line between legitimate scientific criticism and inappropriate personal attacks is always the main issue. . . . Critics should not make things personal, and those whose work is criticized should not take it personally.").

11 Thaxton, *supra* note 1, at 956 n.378 & 965–66 (noting that many of the critics of mismatch "are a veritable who's who of quantitative social scientists" who "have been instrumental in developing the core conceptual and methodological toolkit necessary for credible causal inference across a wide range of substantive domains"); *see, e.g., SFFA,* 143 U.S. 2256 (Sotomayor, J., dissenting) (noting the "[U.S. Supreme] Court previously declined to adopt

engaging in zealotry,, several of Sander's earliest (and most persistent) critics expressly stated that they would not support affirmative action if there was reliable evidence of that policy harming its intended beneficiaries.[12] In an effort to preserve the legal and political relevance of mismatch in light of these mounting negative assessments, proponents of mismatch now question the probative value of existing data. They suggest that the limitations of available data hamper the precise measurement of the degree of mismatch and that better data would reveal stronger support for their theory.[13] This claim prompted two analysts to remark that "in the scholarly and public discourse mismatch claims seem to function largely as a way to cloak value-based objections to affirmative action with a veneer of concern for the success of affirmative action minorities."[14]

The extent to which this "better" data support or challenge mismatch theory remains an important area of inquiry. It turns out that the evidence does not support the theory, and claims to the contrary rely on inappropriate and

---

this so-called 'mismatch' hypothesis for good reason: It was debunked long ago"); *see also infra* note 28.

12 Thaxton, *supra* note 1, at 856.

13 Brief for Richard Sander as *Amicus Curiae* in Support of Petitioner at 28, *Students for Fair Admissions, Inc. v. President and Fellows of Harvard College and University of North Carolina et al.*, 143 S. Ct. 2141 (2023) (Nos. 20-119 & 21-707) [hereinafter Sander, Brief in *SFFA*] ("[T]he low quality of the available data would tend to bias estimates away from finding mismatch. Thus, data allowing direct measurement of the gap between individual student credentials, and the median credentials at their school, might well show larger mismatch effects."); Sander, *Stylized Critique*, *supra* note 8, at 1647 (arguing that critics of mismatch "opposed the release of data . . . [and] are afraid of what better data will show . . . [and] oppose transparency in higher education"); *see also* Carol J. Williams, *Does Affirmative Action Help or Hurt Lawyers? Professor Seeks State Bar Exam Data to Study Racial Bias. Bar Says No,* L.A. Times B1 (Sept. 8, 2008) (noting that Sander requested what he described as "the perfect database" to test mismatch theory—i.e., thirty years of data from the State Bar of California for exam-takers, including age, race, gender, academic records, and bar scores); *SFFA,* 143 U.S. 2198 n.8 (Thomas, J., concurring) ("And, of course, if universities wish to refute the mismatch theory, they need only release the data necessary to test its accuracy."). *But see infra* notes 137–138 and accompanying text (challenging the claim that poor data quality has necessarily biased results against mismatch).

14 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*, University of Houston Institute for Higher Education Law & Governance Working Paper No. 13-12 17 (2014) [hereinafter Kidder & Lempert, *White Paper*] (shorter version published as 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*]). *Compare* Dan L. Burk, *When Scientists Act Like Lawyers: The Problem of Adversary Science*, 33 Jurimetrics 363, 366–67 (1993) ("Whereas the professional norms of the legal profession revolve around adversarialism and advocacy, the professional norms of science . . . involve[] the creation of concepts and their exploration through the tools of empirical research to construct a coherent view of the universe and its workings.").

indefensible methods of parsing and interpreting these data. This response to the latest article making a case for mismatch theory demonstrates how.

In *Mismatch and Bar Passage: A School-Specific Analysis*,[15] Richard Sander and Robert Steinbuch (hereinafter S&S) offer a statistical analysis of the "law school mismatch hypothesis" in an effort to explain racial differences in the likelihood of passing the bar examination on the first attempt.[16] S&S claim that their analysis is an improvement on prior studies of mismatch because they have better data. The prior studies relied on data from the Law School Admission Council's (LSAC) Bar Passage Study (BPS),[17] but their recent study uses school-specific data, which permits a more granular measure of mismatch than was possible with the BPS. Whereas the BPS combines data on nearly 28,000 students from 163 (of 172) law schools into six clusters of schools,[18] S&S's school-specific data on approximately 3,800 law students uniquely identify the three law schools from which the data was drawn: two in California (University of California, Los Angeles (UCLA) and University of California, Davis (Davis)) and one in Arkansas (University of Arkansas, Little Rock (UALR)).[19] According to S&S, another advantage of the school-specific data is their recency.[20] The BPS focuses on students who entered law

15 Richard Sander & Robert Steinbuch, *Mismatch and Bar Passage: A School-Specific Analysis*, 71 J. Leg. Edu. 716 (2024). The initial draft of the paper was uploaded to SSRN, under the same title, on October 17, 2017, as UCLA School of Law, Public Law Research Paper No.17-40, https://ssrn.com/abstract=3054208. The paper was also identified as forthcoming in the *Journal of Legal Education* by Sander in his *amicus* brief in the Harvard and University of North Carolina affirmative action cases. Sander, Brief in *SFFA*, *supra* note 13, at 29 ("Our analysis of the [school-specific] data . . . has been peer-reviewed and accepted for publication by the *Journal of Legal Education*.").

16 It is debatable whether research examining racial disparities in bar exam performance that focuses on first-time success, rather than eventual success, is consonant with the core claims of mismatch theory. *See infra* note 65 and accompanying text.

17 Linda F. Wightman, Law Sch. Admission Council, LSAC National Longitudinal Bar Passage Study (1998).

18 The BPS groups law schools into six "clusters" based on seven school characteristics: size, cost, selectivity, faculty-student ratio, percentage minority, median LSAT score, and median undergraduate GPA. *Id.* at 8 (explaining that "the purpose of grouping law schools was to consider together those schools that were most like one another on some characteristic or set of characteristics. This option is particularly important when samples within individual schools are small, as was the case with ethnic group data in this study"). Consequently, it was not possible to identify individual schools within the dataset. *See also infra* note 137 (explaining the anonymity concerns that motivated LSAC's decision to aggregate schools into clusters).

19 Sander & Steinbuch, *supra* note 15, at 724. The number of graduating cohorts varies across the three schools: Three cohorts are available for UCLA (three distinguishable cohorts and 752 in-state bar-takers); fifteen cohorts are available for Davis (five distinguishable cohorts and 2,333 in-state bar-takers); and seven cohorts are available for UALR (one distinguishable cohort and 723 in-state bar-takers). *Id.* at 725 tbl.1. For the reader's convenience, I reproduce S&S's Table 1 in this article. *See* Table 1.

20 *Id.* at 717; *see also* Kidder & Lempert, *Mismatch Myth*, *supra* note 14, at 113 (explaining that

school in 1991 and tracks those students for six years.[21] S&S's school-specific analysis focuses on students who graduated between 1997 and 2011, following California's 1996 ban on affirmative action by Proposition 209.[22]

It appears, at least for some scholars, that old habits die hard when testing mismatch.[23] The improvements to their data notwithstanding, S&S still fail to meet the key theoretical and methodological standards of statistical analysis required to shore up their theory of mismatch. Several of the theoretical and methodological mistakes that I both raised with Sander[24] and described in

analysts must use caution when generalizing the results of the BPS data to current law school dynamics).

21 Wightman, *supra* note 17, at 2 n.4 (explaining that the BPS "is the culmination of a massive six-year data collection [and] the study tracked students who entered law school in fall 1991 through three or more years of law school and up to five administrations of the bar examination").

22 Sander & Steinbuch, *supra* note 15, at 725 tbl.1. On November 5, 1996, Californians voted to ban the consideration of race and ethnicity in college/university admissions via a statewide ballot initiative (Proposition 209). The ban took effect in 1997 for University of California law schools and in 1998 for the University of California undergraduate campuses. Danny Yagan, *Supply v. Demand Under an Affirmative Action Ban: Estimates from UC Law Schools*, 137 J. Pub. Econ. 38, 40 (2016); *accord* Richard Sander & Stuart Taylor, Jr., Mismatch: How Affirmative Action Hurts Students It's Intended to Help and Why Universities Won't Admit It 152 (2012) (noting that "post-Prop 209 . . . most graduate programs in the UC system were instructed to adopt race-neutral admissions for classes entering in 1997 and beyond"). Consequently, students in the two California law schools—UCLA & Davis—who enrolled after 1996 were not admitted under an affirmative action regime. In the first year that the affirmative ban took effect, the number of Black and Latinx students enrolled UC Berkeley and UCLA plunged by nearly half. Teresa Watanabe, *California Banned Affirmative Action in 1996. Inside the Struggle for Diversity*, L.A. Times (Oct. 31, 2022). At UC Berkeley School of Law, the "[1998] first-year class of 270 included only six or seven Black students, compared with four times that many in the class two years ahead . . . before Proposition 209." *Id.*

23 *See* Thaxton, *supra* note 1, at 994 ("In the roughly fifteen years that have followed [the initial methodological critiques of his work], Sander has been provided ample opportunities to un-'muddy' the waters, 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"); *see infra* note 529 (noting that researchers often attempt to maximize professional recognition by defending problematic assumptions in order to make bold empirical claims).

24 Sander presented a paper responding to a prior critique of his work by Professors Ian Ayres and Richard Brooks at a UCLA School of Law faculty workshop. Richard Sander, *Replication of Mismatch Research: The Case of Ayres and Brooks*, UCLA School of Law Summer Works-in-Progress Workshop Series (June 20, 2018). Both during and immediately following the workshop, I identified recurring methodological problems with Sander's empirical analyses and provided Sander with sources that supported my view. In the published version of the workshop paper, Sander acknowledged the feedback he received from the participants at that faculty workshop, but the issues I raised with the paper were neither acknowledged nor addressed. Richard Sander, *Replication of Mismatch Research: Ayres, Brooks and Ho*, 58 Int'l Rev. L. & Econ. 75, 88 (2019) [hereinafter Sander, *Replication of Mismatch Research*]. As Andrew Gelman has emphasized, "Even if researchers are innocently unaware of statistical errors in their published work, it is poor behavior for them to refuse to admit they could ever have been wrong [because] we can also make mistakes in good faith." Andrew Gelman, *Ethics and*

my *UCLA Law Review* article, *How Not to Lie About Affirmative Action*,[25] are present in S&S's latest study. Consequently, I cannot agree with S&S's conclusion that "the mismatch effect in [their] models accounts for most of the large disparities in bar passage across racial lines."[26]

I have not had the opportunity to conduct my own independent analysis of the school-specific data (as I did with the BPS data in my *UCLA Law Review* article), because the data are not publicly available; but in this article I identify and discuss—as leading scholars of statistical analysis advise—"statistical oversights which are out there in plain sight that advance claims that do not

---

Statistics: Honesty and Transparency Are Not Enough, 30 CHANCE 37, 39 (2017).

25 Thaxton, *supra* note 1. The words "lie" and "statistics" are often associated with one another. Mark Twain is credited with popularizing the phrase "There are lies, damned lies, and statistics." *See* JOEL BEST, DAMNED LIES AND STATISTICS: UNTANGLING NUMBERS FROM THE MEDIA, POLITICIANS, AND ACTIVISTS 5 (2012). Darrell Huff's HOW TO LIE WITH STATISTICS (published in 1954) was a best-seller and is considered a classic; *see* J. Michael Steele, *Darrell Huff and Fifty Years of How to Lie with Statistics*, 20 STAT. SCI. 205 (2005) ("*How to Lie with Statistics* has sold more copies than any other statistical text."). When responding to a critique of the statistical evidence purportedly supportive of mismatch, Sander himself has remarked that "there is no interpretation or argument: the authors of the ESB [Empirical Scholars Brief in *Fisher*] simply lie." Sander, *Mismatch and the ESB*, *supra* note 8, at 568.

Gary King famously noted that one can knowingly or unknowingly lie with statistics, so bad faith is *not* a prerequisite. Gary King, *How Not to Lie with Statistics: Avoiding Common Mistakes in Quantitative Political Science*, 30 AM. J. POL. SCI. 666, 667 (1986). In fact, as King explains, recurring problems in quantitative social science "often represent important theoretical and conceptual misunderstandings" about statistical inference rather than deliberate misrepresentations. *Id*. at 666. In the same spirit as King's article, I titled my article *How Not to Lie About Affirmative Action*, identified persistent problems in the extant research on mismatch theory, and proposed sensible solutions. Thaxton, *supra* note 1, at 838. To that end, I highlighted conceptual and methodological problems present in research that has been both supportive and unsupportive of mismatch theory. I neither assumed nor claimed that these errors were committed in bad faith unless there was evidence to the contrary.

Across disciplines, quantitative methodologists have endeavored to educate their respective fields on how to avoid "lying" with statistics. *See, e.g.*, Daniel E. Ho & Kevin M. Quinn, *How Not to Lie with Judicial Votes: Misconceptions, Measurement, and Models*, 98 CAL. L. REV. 813 (2010) (discussing core flaws in the statistical analysis of judicial voting); Luc Anselin, *How (Not) To Lie with Spatial Statistics*, 30 AM. J. PREV. MED. S3 (2006) (describing methodological challenges when employing spatial analysis in cancer research); Phillip J. Fleming & John J. Wallace, *How Not to Lie with Statistics: The Correct Way to Summarize Benchmark Results*, 29 COMM. ASSOC. COMPUTING MACHINERY 218 (1986) (explaining problems with the use of measures of central tendency of data with large variances when evaluating computing efficiency); Scott Farrow, *How (Not) to Lie with Benefit-Cost Analysis*, 10 ECON. VOICE 45 (2013) (identifying several key challenges in the practice and interpretation of benefit-cost analysis in welfare economics); Bernice E. Rogowitz & Lloyd A. Treinish, *How Not to Lie with Visualization*, 10 COMPUTERS PHYSICS 268 (1996) (describing how common mistakes in the visual representation of magnetic-resonance-imaging (MRI) data impede interpretation); *see also* Steele, *supra* at 207 ("Indeed, the title [*How to Lie with Statistics*] has done more than simply stand the test of time; it has been honored through the years by other authors who pursued their own variations.").

26 Sander & Steinbuch, *supra* note 15, at 740; *see also* Sander, Brief in *SFFA*, *supra* note 10 at 29 ("Once again, [S&S] have a strong finding that it is preferences, not race, which explain large racial disparities in educational outcomes.").

follow from the data . . . . All of these mistakes are well known and there have been many articles written about them, but they continue to appear."[27] S&S state that they are willing to provide their data and computer code to interested scholars who submit a request, so I look forward to more deeply engaging their work and, ideally, assisting them in addressing the mistakes I describe below.[28]

A. Overview

In this article I identify six problems with *Mismatch and Bar Passage: A School-Specific Analysis*. I also attempt to propose a sensible solution, or set of solutions, to each problem. These problems are not new in this line of research, nor are they news to the authors. Most of these mistakes are recurring and remain unacknowledged–that is, they have already been identified in prior research conducted by Sander.[29] These errors have their origins in ineffective

27 *See generally* Tamar R. Makin & Jean-Jacques Orban de Xivry, *Science Forum: Ten Common Statistical Mistakes to Watch Out for when Writing or Reviewing a Manuscript*, 8 eLife 1 (2019) (discussing common statistical mistakes found in the scientific literature) (quote edited for clarity) [http://elifesciences.org/articles/48175].

28 *But compare* Sander & Steinbuch, *supra* note 15, at 744 ("Scholars interested in reproducing or replicating our results, or exploring the data for other purposes, should contact Sander"), *with* Thaxton, *supra* note 1, at 960 (noting that "the main obstacle to engaging in fruitful discussions about the impact of race-conscious admissions policies . . . [is] 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") *and* Kidder & Lempert, *Mismatch Myth*, *supra* note 20, at 108 ("Sander . . . has acknowledged problems with his data and models, but his concessions have never extended to acknowledging fundamental flaws") *and* Daniel E. Ho, *Affirmative Action's Affirmative Actions: A Reply to Sander*, 114 Yale L.J. 2011, 2013 (2005) (explaining that rather than responding to specific criticisms of his empirical analyses, Sander employed an entirely different analytical framework based on debatable assumptions and changed the outcome variable from eventual bar passage to first-time bar passage in an effort to defend his core claims about mismatch and bar passage); *see also infra* note 38 and accompanying text.

On June 16, 2006, Sander testified before the U.S. Commission on Civil Rights and urged the commission to assemble a panel of social scientists to review and verify his research on the impact of affirmative action on academic performance in law school, bar passage, and long-term legal career success. U.S. Comm'n on Civil Rights, Affirmative Action in Law Schools 9, 35 (2007) (statement of Richard Sander). Several years later, in an *amicus* brief filed in *Fisher v. University of Texas at Austin (Fisher I)*, eleven of the nation's most influential theoretical and applied statisticians stated that "based on over 255 collective years of social-science research experience [we] have concluded the research [supporting] Professor Sander's 'mismatch hypothesis' is unreliable, failing the basic tenets of research design." Brief of Empirical Scholars as *Amici Curiae* in Support of Respondents at 1, *Fisher I*, 570 U.S. 297 (2013) (No. 11-345) [hereinafter Empirical Scholars Brief in *Fisher I*]. Although Sander acknowledged the gravitas of the signatories of the amicus, he responded to the amicus by describing the signatories as "ideological Zealots [who] see their attacks on mismatch as a sort of holy war" and describing their work as "ideology wearing empirical makeup," "laughably inept throughout," and "fraudulent to its core." *See supra* note 8; Thaxton, *supra* note 1, at 846 n.40 & 960.

29 *Compare* Brown et al., *supra* note 8, at 2563 ("[B]y labeling something an error, we declare only its lack of objective correctness, and make no implication about the intentions of those

experimental designs, inappropriate analyses, and flawed reasoning;[30] "involve factual mistakes or veer substantially from clearly accepted procedures in ways that, if corrected, might alter a paper's conclusions";[31] and "there is a reasonable expectation that the scientist[s] should have or could have known better."[32] Moreover, "these problems are often interdependent, such that one mistake will likely impact others, which means that many of them cannot be remedied in isolation."[33] Indeed, the "creation of first-rate empirical research [in legal scholarship requires]… awareness of…[and] compliance with the of rules of inference [and] paying heed to the key lessons of the revolution in empirical analysis that has taken place over the last century in other disciplines."[34]

The identification of errors and the assessment of the impact of those errors in a particular instance are distinct—yet equally important—inquiries in both law[35] and statistics.[36] Errors of statistical inference lead to failed replication or lack of robustness.[37] And the fact that these errors have been repeatedly

making the error.") *with* Gelman, *supra* note 24, at 39 (noting that professionalism requires researchers to acknowledge statistical errors in their work when they are identified by others, even when those errors were made in good faith).

30 *See, e.g.*, Brown et al., *supra* note 8, at 2563, 2567 ("[I]n science, three things matter: the data, the methods used to collect the data (which give them probative value), and the logic connecting the data and methods to conclusions."); *accord* Richard Lempert, *Growing Up in Law and Society: The Pulls of Policy and Methods*, 9 Ann. Rev. L. & Soc. Sci. 1, 13 (2013) ("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.").

31 Brown et al., *supra* note 8, at 2563.

32 *Id.*

33 Makin & Orban de Xivry, *supra* note 27, at 1–2; *accord* Richard McElreath, Statistical Rethinking: A Bayesian Course with Examples in R and Stan 565 (2d ed. 2020) ("Scientific discovery is not an additive process, in which sin in one part can be atoned by virtue in another. Everything interacts.").

34 Lee Epstein & Gary King, *The Rules of Inference*, 69 U. Chi. L. Rev. 1 (2002) (quote edited for clarity).

35 *See, e.g.*, 28 U.S.C. § 2111 ("Harmless Error") (criminal law); Fed. R. Crim. P. 52(a) ("Harmless and Plain Error") (criminal law); Fed. R. Civ. P. 61 ("Harmless Error") (civil law); *see also* United States v. Dominguez Benitez, 542 U.S. 74, 86 (2004) (Scalia, J., concurring) (noting that the Court "has adopted no fewer than four assertedly different standards of probability relating to the assessment of whether the outcome of a trial would have been different if error had not occurred.").

36 Brown et al., *supra* note 8, at 2563 ("[I]f we do not identify and correct errors, science cannot claim to be self-correcting, a concept that has been a source of critical discussion. . . . By errors, we mean actions or conclusions that are demonstrably and unequivocally incorrect from a logical or epistemological point of view (e.g., logical fallacies, mathematical mistakes, statements not supported by the data, incorrect statistical procedures, or analyzing the wrong dataset).").

37 *Id.* at 2565 ("Identification of an error in a paper or reasoning cannot be used to say the conclusions are wrong; rather, we can only say the conclusions are unreliable until further analysis."). Failed replication can result from either failed verification or failed reproduction.

identified in prior work conducted by proponents of mismatch, yet remain unacknowledged and unaddressed, is itself illuminating.[38] I present a concise summary of these problems below and provide greater detail later in my review. Properly unpacking the organizational dynamics contributing to racial differences in law students' bar exam performance is of monumental importance: it assists law schools in their continuing efforts to reassess their admissions practices to ensure that all students can maximally contribute to and gain from the academic experience.

The first problem is the improper interpretation of the effect of race on bar passage. S&S do not show, as mismatch theory hypothesizes, that race has no independent impact on law school grades (LGPA) after accounting for students' degree of mismatch. A core concept of mismatch theory is "learning mismatch"—by which S&S contend that students tend to *learn* less in law school when their undergraduate grades and law school entrance exam scores dip significantly below those of their classmates, which in turn hurts their ability to pass the bar. S&S take LGPA as a proxy for learning mismatch. Given the centrality of "learning mismatch," for an analyst to accurately claim that racial disparities in bar exam performance are explained by mismatch, they must prove as much through a properly constructed test of that theory. Such a test must disentangle the direct and indirect effects (operating through learning mismatch) of race and credential mismatch. Absent an examination of this mediation process, S&S's analysis falls short of causal explanation

Michael A. Clemens, *The Meaning of Failed Replications: A Review and Proposal*, 31 J. Econ. Surveys 326, 327 (2017). Verification pertains to validating the reported results with the same methods and sample and is used to detect measurement error, coding errors, errors in dataset construction, and scientific misconduct. *Id*. Reproduction involves using the same methods on a different sample from the same population. Lack of robustness stems from either a failed reanalysis or a failed extension. *Id*. Reanalysis involves altering the specification of the model on the same or representative sample. Extension pertains to analyzing a sample from a different population with the same model specification. *Id*.

38 *See, e.g.*, Ho, *supra* note 25, at 2013 (explaining that rather than responding to specific criticisms of his empirical analyses, Sander employed an entirely different analytical framework based on debatable assumptions and also changed the outcome variable from eventual bar passage to first-time bar passage in an effort to defend his core claims about mismatch and bar passage); *accord* Richard Lempert et al., *Affirmative Action in American Law Schools: A Critical Response to Richard Sander's—A Reply to Critics* 2, 8 (University of Michigan Law & Economics Olin Working Paper No. 06-001, 2006) (noting that "the claims in Sander's *Reply to Critics* are so at odds with the claims of *Systemic Analysis*, that Sander's reply effectively repudiates much of *Systemic Analysis*," "Sander's reply explicitly or implicitly repudiates much of the methodology and many of the claims he made in *Systemic Analysis*," and Sander shifts his focus from eventual bar passage to first-time bar passage "because, as [Ian] Ayres and [Richard] Brooks show in their reanalysis of the bar passage data and as [Daniel] Ho also found, analyses using eventual bar passage as the test of affirmative action do not reveal a mismatch effect"). *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.").

and, at best, constitutes "black box" causal description, assuming adequate adjustment for confounding variables. Scholars from diverse fields of science have emphasized that credible causal inference requires analysts to posit causal mechanisms linking treatments to outcomes and think carefully about the observable implications of their proposed causal process.[39]

The second problem is the incorrect comparison of the effect of race across nested models. S&S "eyeball" differences in the effect of race across models that introduce mismatch-related variables. On that basis, they conclude that adjusting for mismatch either significantly reduces or eliminates the effect of race on the likelihood of passing the bar. S&S's analytical approach is flawed because (1) the logistic regression coefficients they estimate are not directly comparable across models and must be recalibrated before any meaningful distinction can be made, and (2) eyeballing differences in coefficients can be misleading, and reliable analysis requires a more rigorous approach to comparing coefficients. Consequently, the magnitude and the statistical significance of any change in the effect of racial differences due to mismatch that S&S report may be unreliable.[40]

The third problem is the poor predictive ability of the estimated regression models that S&S construct. There are two related issues here. The first issue is S&S's use of the model fit statistic "Somers' *D*," which is inadvisable because it inflates the predictive capacity of their regression models. The second issue is the lack of incremental predictive improvement for models that account for students' degree of mismatch. In other words, models that include a variable(s) for mismatch are no better at predicting bar passage than models that simply include students' law school admission exam score or a weighted combination of undergraduate grades and the law school entrance exam score.[41]

The fourth problem is the sensitivity of the hypothesized race effect to modest changes in the assumptions concerning how mismatch is both measured and influences bar exam performance. That is, whether a student's level of mismatch can account for racial differences is dependent, in part, on (a) the measurement of the mismatch variable, and (b) whether analysts assume a linear or nonlinear relationship between mismatch and the likelihood of passing the bar in several of S&S's models.[42] Improper measurement injects bias into the analysis by misidentifying treatment and control groups, misjudging the treatment dosage, and failing to balance the treatment and control groups with respect to potential confounding variables. Incorrect functional form assumptions undermine statistical adjustments for confounding and mediating variables when appropriate counterfactual comparisons are either nonexistent or sparse in the observed data. This problem is called "model dependence."

39 *See* Part III.A.

40 *See* Part III.B.

41 *See* Part III.C.

42 *See* Part III.D.

The fifth problem is the interpretation of the role of school environment on bar exam performance. There are two issues here as well. First, racial differences in bar passage rates are eliminated only in models that include school-specific effects. This finding implies that differences in school environment influence whether racial differences remain after accounting for students' level of mismatch. And this result directly contradicts Sander's claim that Black students perform poorly on the bar examination, relative to white students, because of their level of mismatch and not because of racial hostility or alienation experienced at law school. Second, the positive effect of attending a more difficult law school on bar exam performance is larger than the negative effect of credential deficit for the first three levels of students' credential deficits. This finding refutes S&S's claim that students would perform better on the bar exam at a less competitive school for the majority of students with a similarly situated counterpart.[43]

The sixth problem is the incorrect claim that the results from S&S's supplementary analysis presented in Table A—which are intended to demonstrate the reliability of their conclusions—are consistent with the results from the other models in their paper. While some of the parameter estimates and model fit statistics are consistent with the other models examined in the paper, S&S's core claim that mismatch accounts for racial differences in the likelihood of passing the bar is contradicted by their own analyses. Specifically, adjusting for mismatch does not eliminate racial differences in bar passage for Black students in any of S&S's models, and such adjustment eliminates the racial difference for Latinx students in only one of their models.[44]

B. What is Mismatch?

Mismatch theory posits that students perform poorly in law school (including failing to graduate) and on the bar exam because they are elevated into environments where they are academically inferior to their classmates. According to its proponents, mismatch produces these poor outcomes through several hypothesized mechanisms, but the key mechanism is "learning mismatch."[45] In other words, students who are outmatched by their peers on

43 *See* Part III.E.

44 *See* Part III.F.

45 Sander, *Replication of Mismatch Research*, *supra* note 24, at 76 (claiming that, in the law school context, learning mismatch is primarily responsible for student differences in bar exam performance); *accord* Thaxton, *supra* note 1, at 1006 ("Recall that, according to Sander's theory, 'learning mismatch' is the primary mechanism through which outmatched [law] students are harmed by affirmative action.").

When Sander initially articulated mismatch theory, he did not specify the full set of mechanisms (labeled "first-order effects") but did so in later work that responded to critiques of his analyses. *See* Sander, *Stylized Critique*, *supra* note 8, at 1642–43. His claim of three negative consequences of mismatch are: learning mismatch, competition mismatch, and social mismatch. *Id*. However, several scholars have failed to find empirical support for these first-order effects. *See, e.g.*, Camille Z. Charles et al., Taming the River: Negotiating the Academic, Financial, and Social Currents in Selective Colleges and Universities

assessments of undergraduate grades (UGPA) and Law School Admission Test (LSAT) scores learn *less* in law school environments because law professors teach to the "middle" student.[46] Had these students attended law schools where

200 (2009) (reporting that race-conscious affirmative action does not result in social mismatch); Beverly Daniel Tatum, Why Are All the Black Kids Sitting Together in the Cafeteria? (rev. ed. 2003) (describing the dynamics leading to students' so-called self-segregation into racial/ethnic groups in educational institutions that are independent of academic ability); *accord* Uma M. Jayakumar, *The Shaping of Post-College Colorblind Orientation Among Whites: Residential Segregation and Campus Diversity Experiences*, 85 Harv. Edu. Rev. 609 (2015) (noting that white students who were primarily socialized in and accustomed to segregated white environments before college are more inclined toward sticking to white-dominant environments on campus (e.g., Greek organizations) and less likely to choose to engage in cross-racial interactions (compared with their counterparts from racially integrated precollege environments)); Richard Lempert, *Mismatch and Science Desistance: Failed Arguments Against Affirmative Action*, 64 U.C.L.A. L. Rev. Disc. 136, 160–62 (2016) (challenging the competition mismatch hypothesis by showing that a critical mass of minority students at an institution likely explains differential academic performance among science majors rather than school selectivity); *accord* Frederick L. Smyth & John J. McArdle, *Ethnic and Gender Differences in Science Graduation at Selective Colleges with Implications for Admission Policy and College Choice*, 45 Res. Higher Edu. 353, 369 (2004) (finding that "institutional selectivity also failed to produce additional significant parameters" when predicting graduation with science, math, or engineering degrees). *But see* Peter Arcidiacono et al., *Racial Segregation Patterns in Selective Universities*, 56 J.L. & Econ. 1039, 1958–59 (2013) (finding tentative support for social mismatch based on academic background for close friendships, but also acknowledging that an important limitation of the study was that it did not investigate interracial social interactions in other important social settings, such as the classroom and social clubs).

46 Sander & Steinbuch, *supra* note 15, at 716 ("Sander argued that law professors (like most teachers) tend to tailor instruction to the middle range of ability in a classroom."); *accord* Douglas Williams et al., *Revisiting Law School Mismatch: A Comment on Barnes (2007, 2011)*, 105 Nw. L. Rev. 813 n.3 ("The mismatch hypothesis implicitly posits that instructors aim the level of instruction at the middle student."). It is debatable, however, whether the "middle student" is the correct baseline, and Sander even acknowledges that "much of the evidence behind these theories [of teaching to the middle student] is more anecdotal than systematic[.]" Richard Sander, *A Systemic Analysis of Affirmative Action in Law Schools*, 57 Stan. L. Rev. 367, 451 (2004) [hereinafter Sander, *Systemic Analysis*].

There are at least three reasons to question Sander's teaching-to-the-middle-student assumption on which "learning mismatch" relies. First, instruction at the typical law school is much more individualized than Sander would have his readers believe. The Socratic method, which is the dominant style of teaching in U.S. law schools (especially in the first-year curriculum), entails the professor engaging in a back-and-forth dialogue with a single student (or small number of students) for the majority of each class period. Consequently, instruction is more student-based than classroom-based. Granted, the middle student may be able to better follow (or participate in) the Socratic dialogue during exchanges than students in the lower end of the entering credential distribution, but it would be a stretch to claim that the professor is teaching to the middle student. It is also common that professors teaching seminars and clinical courses adopt a quasi-individualized approach to teaching. And, of course, nearly all law professors are required to make themselves available to individual students during office hours.

Second, it is incorrect to assume that the middle student at the law school level is similar to the middle student at the classroom level with respect to entering credentials. After their first year, students do not randomly sort themselves into classes; rather, they typically choose courses that align with their summer employment and post-graduation goals. In fact, only in the first-year law school curriculum are students randomly assigned

they matched their classmates' UGPA and LSAT scores, they would learn *more* and, as a consequence, achieve better law school-related outcomes (i.e., better grades, higher likelihood of graduating and passing the bar exam, and better employment).[47] According to S&S, race-conscious admissions policies "hurt their intended beneficiaries by undermining student learning and subsequent performance on bar exams . . . [and] without preferences, Blacks would have grades comparable to whites and the bar passage gap would substantially narrow."[48]

Mismatch theory can be depicted in a causal diagram, which should assist the reader in developing intuitions about causal inference and help clarify the shortcomings that I highlight in S&S's empirical analyses.[49] Figure 1 shows the

to *most* of their classes. Consequently, the middle student for classes in the second and third years of law school may vary significantly from the overall middle student. This may partly explain why LSAT score has been shown to be a poor predictor of LGPA in upper-division courses. *See infra* note 155. Because classrooms are subunits of the overall law school, Sander engages in an *ecological fallacy*—i.e., where relationships discovered at one particular level are inappropriately assumed to occur in the same fashion at some lower level. Craig Duncan et al., *Context, Composition and Heterogeneity: Using Multilevel Models in Health Research*, 46 Soc. Sci. & Med. 97, 98 (1998).

Third, for "mismatch to result in worse outcomes, low credential students have to learn less than they would have at another school—not less than the middle-range student at their current institution." Katherine Y. Barnes, *Is Affirmative Action Responsible for the Achievement Gap Between Black and White Law Students? A Correction, a Lesson, and an Update*, 105 Nw. L. Rev. 791, 792 n.4 (2011). But simply demonstrating that a student may have earned a higher class rank at a less selective school does not necessarily imply that the student has learned less at the more selective school. As Ho explains, "the same student may receive lower grades in a higher-tier school even if she learned the same amount." Ho, *supra* note 28, at 2012; *see also* Part III.E (noting that S&S's analysis contradicts mismatch theory because their results show that the positive effect of attending a more competitive school trumps the negative effect of credential deficits for the first three levels of mismatch—which constitute the most plausible counterfactual comparisons—even after adjusting for overall credentials).

47 *But see infra* note 155 (describing studies reporting that female law students perform worse than their male counterparts with identical entering credentials).

S&S acknowledge that students with large positive mismatch might be disadvantaged by attending a less competitive school. They suggest that a proper test would require access to students' precise bar exam scores, and not simply whether the student passed the exam, because nearly all positively mismatched students passed the bar. Sander & Steinbuch, *supra* note 15, at 729 nn.29 & 727 tbl.2. Another plausible test of positive mismatch could have been performed using LGPA as the outcome variable, and S&S have such data for all cohorts from Davis & UALR and from some cohorts from UCLA. *Id.* at 725 tbl.1 & 743; *see also* Part III.A (noting that S&S do not examine the relationship between race and LGPA, net level of entering credentials and mismatch, although that inquiry would help illuminate the direct and indirect impacts of race on bar passage).

48 Sander & Steinbuch, *supra* note 15, at 716. *But see supra* note 46 (emphasizing that law school grades may be a poor proxy for legal learning when comparing similarly situated students across institutions of unequal competitiveness).

49 Thaxton, *supra* note 1, at 849 n.51 ("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

relationships between S&S's four core variables: (1) aptitude for legal training (*Aptitude*), (2) law school selectivity/difficulty (*Selectivity*), (3) acquired legal knowledge/reasoning skills (*Knowledge*), and (4) competence for legal practice (*Competence*).[50] *Aptitude*, *Selectivity*, *Knowledge*, and *Competence* are latent variables (i.e., abstract concepts) that can be measured only indirectly. Figure 2 depicts the manifest variables used to measure the latent variables.[51] Student UGPA and Student LSAT are proxies for *Aptitude*; Law School UGPA and Law School LSAT are proxies for *Selectivity*;[52] LGPA is a proxy for *Knowledge*; and bar exam (BAR) is a proxy for *Competence*.[53] Race is excluded from the causal diagram because, according to S&S, race does not have a causal relationship with any of the other variables.[54]

S&S's theoretical model makes four additional assumptions in order to establish a causal relationship between entering credentials, law school grades, and bar exam performance. First, there are no unobserved common causes of Law School UGPA or Law School LSAT (or, more accurately, matriculating to a law school at a certain level of selectivity/difficulty) and BAR. Second, there are no unobserved common causes of LGPA and BAR. Third, there are no unobserved common causes of Law School UGPA, Law School LSAT and LGPA. Lastly, there are no unobserved common causes of LGPA and BAR that are affected by Law School UGPA and Law School LSAT. This fourth assumption, called the "no post-treatment confounders assumption" is especially strong and highly unlikely. It essentially requires that there are no other unmeasured mediators (intervening causes) linking law school competitiveness/selectivity to bar passage that also influence LGPA.[55]

---

analysis.") (citation omitted).

50 *See also* Thaxton, *supra* note 1, at 849 fig.1. Figure 1 accounts for students' absolute and relative (i.e., school-specific) differences in aptitude for legal training by modeling law school selectivity/difficulty.

51 *Id.*, at 850 fig.2. This figure is slightly modified from the one presented in *How Not to Lie About Affirmative Action* because the BPS data use TIER as a proxy for school difficulty/selectivity, whereas S&S's data include the school-level UGPA and school-level LSAT. Figure 2 accounts for students' absolute and relative (i.e., school-specific) differences in entering credentials by modeling law school UGPA and law school LSAT.

52 Law School UGPA and Law School LSAT represent, respectively, the average student UGPA and average student LSAT at a particular law school.

53 Thaxton, *supra* note 1, at 851. *But see infra* note 181 (noting UGPA and LSAT may be racially biased proxies for aptitude for legal learning).

54 *See, e.g.*, Sander, *Systemic Analysis*, *supra* note 46, at 447 ("If there were any other factor that somehow disadvantaged blacks . . . then this would make being black an independently significant causal factor in bar passage rates. But it is not."). *But see infra* notes 156–186 and accompanying text (summarizing the empirical literature finding that race has an impact on LGPA, independent of UGPA, LSAT, and school difficulty/selectivity).

55 Thaxton, *supra* note 1, at 853–54. This assumption is more plausible if LGPA occurs shortly after the exposure to the treatment (i.e., law school difficulty/selectivity). Based on S&S's theoretical framework, the assumption is unlikely to hold because of the temporal distance between what is learned in law school and when a student takes the bar exam. *Id.*

In short, a necessary condition to establish a causal relationship for S&S's models is that one can reasonably rule out all other reasons that similarly situated students differ in law school performance (i.e., LGPA) and bar exam performance besides the influence of differences in degree of mismatch.[56]

II. A Comment on Peer Review & Research Design for Statistical Inference

A. On Peer Review

It has been said that the peer review process "makes the publishing world go around" by identifying the most impactful scholarly contributions and improving the quality of manuscripts that advance through that process.[57] A major challenge to the peer review process is the quality of the review itself, and editors of journals across disciplines routinely lament the difficulty of attracting qualified peer reviewers and establishing dedicated editorial boards.[58] Frequently, peer review misses major errors in submitted papers and is subject to confirmatory biases by peer reviewers.[59] Unsurprisingly, articles that survive the peer review process are challenged by experts in the field, and these challenges are presented in various forms—ranging from articles in peer-reviewed journals, books, and blogs to amicus briefs before the Supreme Court and other tribunals.[60] It might be said, then, that traditional peer review does not validate research findings; rather, it is *one* of several hurdles that must be

56 *Id.* at 848.

57 Paul Djupe, *Peer Reviewing in Political Science: New Survey Results*, 48 Pol. Sci. & Pol. 346, 350 (2015); *accord* Justin Esarey, *Does Peer Review Identify the Best Papers? A Simulation Study of Editors, Reviewers, and the Scientific Publication Process*, 50 Pol. Sci. & Pol. 963 (2017).

58 *See, e.g.*, Charles W. Fox, *Difficulty of Recruiting Reviewers Predicts Review Scores and Editorial Decisions at Six Journals of Ecology and Evolution*, 113 Scientometrics 465 (2017); Kerry-Anne Rye et al., *Working Toward Reducing Bias in Peer Review*, 20 Mol. Cell. Proteomics 1 (2021).

59 Esarey, *supra* note 57; McElreath, *supra* note 33, at 567 ("Peer review [often] selects for hyperbole, since honestly admitting limitations of work only hurts a paper's chances.").

60 Serge Horbach & Willem Halffman, *The Changing Forms and Expectations of Peer Review*, 3 Res. Integrity & Peer Rev. 8 (2018) (describing various forms of nontraditional peer review); Jonathan P. Tennant et al., *A Multi-Disciplinary Perspective on Emergent and Future Innovations in Peer Review*, 6 F1000Research 1151 (2017) (explaining that "[d]ifferent forms of peer review beyond that for manuscripts are also clearly important and used in other contexts such as academic appointments, measurement time, research ethics or research grants"); McElreath, *supra* note 33, at 567 ("What happens to a finding after publication is just as important as what happens before. It is common for invalid analyses to be published in top-tier journals, only to be torn apart on blogs."); Brown et al., *supra* note 8, at 2565 ("Effective postpublication peer review may be particularly useful to mitigate ignorance by using such errors to serve as instructive examples of what not to do."); Liliana M. Garces et al., *Arguing Race in Higher Education Admissions: Examining* Amici's *Use of Extra-Legal Sources in* Fisher, 14 J. Diversity in Higher Edu. 278 (2021) (explaining how amici use research and other types of extralegal sources).

cleared before the scientific community deems a finding, or series of findings, worthy of acceptance.[61]

S&S claim that "several prominent critics [of Sander's empirical analysis of mismatch theory] emerged, though their critiques were not generally published in peer-reviewed journals and have not held up well under closer examination."[62] S&S then identify the peer-reviewed work of Peter Arcidiacono and Michael Lovenheim as supportive of Sander's prior empirical examination of mismatch theory.[63] S&S neglect to mention that Arcidiacono and Lovenheim clearly state that Sander's analysis rested on debatable modeling assumptions, and even when granting those assumptions, that the impact of mismatch on bar exam performance was extremely modest.[64] Specifically, they found that attending a higher-ranked school decreased Black students' chances of passing the bar exam on the first attempt by a mere 2.5% and had no impact on eventual bar passage.[65] Perhaps more interesting is Arcidiacono and

61 Sander has also highlighted methodological problems with articles published in peer-reviewed journals, critiquing two empirical studies and stating that "the publication of these two astoundingly shoddy works . . . should raise the alarm . . . that the editing and peer-review process at these two journals needs overhauling." Richard Sander & Abraham Wyner, *Studies Fail to Support Claims of New California Ethnic Studies Requirement*, Tablet (Mar. 28, 2002), http://www.tabletmag.com/sections/news/articles/studies-fail-to-support-claims-new-california-ethnic-studies-requirement; *see also* Richard Sander, *A Reply to Critics*, 57 Stan. L. Rev. 1963, 1982 n.42 (2005) (claiming that a peer-reviewed study by Timothy Clydesdale reporting the persistence of racial disparities in LGPA even after accounting for entering credentials "survived the peer-review process and found its way into print—gross errors intact") [hereinafter Sander, *Reply to Critics*]. *But cf. infra* notes 165–195 and accompanying text (explaining that Sander exaggerates the alleged flaws in Clydesdale's work and that other scholars who have examined the issue, both before and after the publication of Clydesdale's article, report findings that support Clydesdale's conclusions).

62 Sander & Steinbuch, *supra* note 15, at 716.

63 *Id.* at 716 n.1 (citing Peter Arcidiacono & Michael Lovenheim, *Affirmative Action and the Quality-Fit Trade-Off*, 54 J. Econ. Lit. 3 (2016)). *But compare* Sander, *Reply to Critics*, *supra* note 61, at 1982 ("It is worth noting that a peer review process is no guarantee that a work is methodologically competent.").

64 Arcidiacono & Lovenheim, *supra* note 63, at 19 (discussing Sander's questionable assumptions of no endogeneity bias, his inclusion of historically Black colleges and universities (HBCUs), and his modeling the tier effect continuously rather than discretely).

Although S&S acknowledge that prior empirical work finding support for mismatch theory has been subject to criticism by several prominent scholars, they do not provide any details about those criticisms and to what extent (if any) their current work adequately addresses the myriad theoretical and methodological problems that have been identified in research purportedly finding support for mismatch theory. As discussed below, *see infra* note 74, nearly all independent empirical examinations of mismatch theory have disputed Sander's conclusions.

65 Arcidiacono & Lovenheim, *supra* note 63, at 19; Thaxton, *supra* note 1, at 1000. *Cf.* Andrew Gelman, *Criticism as Asynchronous Collaboration: An Example from Social Science Research*, 11 Stat e464 9 (2022) ("But, remember, honesty and transparency are not enough. If you do a study of an effect that is small and highly variable . . . you've set yourself up for scientific failure: you're working with noise.").

The 2.5% decrease in bar passage was not easily discernible from Sander's article,

Lovenheim's conclusion that, if Sander's analysis is to be "taken at face value . . . 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."[66]

---

and this underscores another recurring issue in Sander's work: the reporting of parameter estimates (e.g., log odds [i.e., logits], odds ratios, and standardized coefficients) that do not measure what they appear and that can be highly misleading. Thaxton, *supra* note 1, at 997–1000 & 1020. This problem is also present is S&S's school-specific analysis. S&S should have reported marginal effects, which reflect changes in predicted probability of bar passage, rather than odds ratios; uncertainty in those effects—based on sampling effects and potential model misspecification—should have also been reported. *Id.* at 858 n.75, 893 n.211, 973–76; *see generally* Lee Epstein et al., *On the Effective Communication of the Results of Empirical Studies, Part II*, 60 Vand. L. Rev. 801, 837 (2007) (explaining that legal scholars engaged in quantitative empirical work have an obligation to communicate substantive effects (not merely statistics) and the level of uncertainty about those effects); *see also infra* note 159.

It is important to note that Sander changed his dependent variable from "whether a person passes the bar on one of her first two attempts," Sander, *Systemic Analysis*, *supra* note 46, at 444, to first-time bar passage in response to multiple critics' reanalysis of the BPS data, which failed to reveal any support for mismatch. Lempert et al., *supra* note 38, at 8 (citing Ho, *supra* note 28, *and* Ian Ayres & Richard R. W. Brooks, *Does Affirmative Action Reduce the Number of Black Lawyers?*, 57 Stan. L. Rev. 1807 (2005)). As Ho explains, the BPS manual "indicates that this variable [i.e., passing the bar on one of her first two attempts] actually represents eventual bar result." Ho, *supra* note 28, at 2013 n.15. Nearly fifteen years later, Sander claimed that Ho's use of the eventual bar passage variable and Ayres & Brooks' use of the same, rather than first-time bar passage variable, was "not a particularly probative way of testing mismatch" and "an odd anomaly." Sander, *Replication of Mismatch Research, supra* note 24, at 80 & 86. Yet there was nothing "odd" or nonprobative about Ho, alongside Ayres and Brooks, analyzing the same variable that Sander himself used in the analysis that was the focus of their initial critiques. As I have noted in prior work, "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 other methodological flaws with his analyses." Thaxton, *supra* note 1, at 970; *accord* Lempert et al., *supra* note 38, at 8. Sander claimed that first-time bar passage, rather than eventual bar passage, is a better reflection of how much a student has learned in law school. But he does not offer any evidence supporting his claim, and there is good reason to believe that first-time bar passage reflects contingencies in addition to what was learned in law school, such as the particular subjects a student took in law school, whether the student took a bar exam prep course, the quality of the prep course, and the amount of time the student was able to devote to studying for the bar. Thaxton, *supra* note 1, at 970–71. Most important, however, is the irrelevance of first-time bar passage to the central question animating Sander's initial inquiry—i.e., whether affirmative action harms Black law students' chances of ever becoming lawyers. To adequately answer this question, the variable of primary interest is eventual bar passage. *See generally* Joel Best, Stat-Spotting: A Field Guide to Identifying Dubious Data 45 (2013) (cautioning that "altering definitions between two measurements can produce very misleading statistics").

66 Arcidiacono & Lovenheim, *supra* note 63, at 17; *see also* Thaxton, *supra* note 1, at 839 n.7. *But see* Ho, *supra* note 28, at 2012–13 ("[T]he fact that law school GPA is the variable with the biggest effect in Sander's regressions is not surprising at all. In fact, it is an indicator of the problem: This coefficient reflects the fact that his regression is controlling away the consequences of law school tier, the key causal variable.").

S&S also identify the peer-reviewed article by Douglas Williams as providing support for mismatch,[67] but Arcidiacono and Lovenheim evaluate Williams's article as well and conclude that "the estimated effects are so large as to not be plausible."[68] Williams's work has also been discredited by a group of world-renowned theoretical and applied statisticians.[69] Furthermore, my own analysis of Williams's article,[70] as well as William Kidder and Richard Lempert's evaluation,[71] similarly found that Williams's results do not withstand scrutiny.

67 Sander & Steinbuch, *supra* note 15, at 716 n.1 & 722 n.14 (citing E. Douglas Williams, *Do Racial Preferences Affect Minority Learning in Law Schools?*, 10 J. Empirical Legal Stud. 171 (2013)).

68 *Compare* Arcidiacono & Lovenheim, *supra* note 63, at 20 n.30 *with* Sander, *Mismatch and the ESB*, *supra* note 8, at 568 (describing Williams's article as "detailed and masterful"). *See also* Thaxton, *supra* note 1, at 871. Although Sander appears unconcerned with the implausibly large effect sizes reported by Williams, he is more circumspect about the large effect sizes reported in two different studies on the effect of taking an ethnic studies course on high school GPA. Sander and a coauthor caution that "the methods employed *appear* to be highly sophisticated, but . . . the fundamental absurdity of reporting such large effect size is ignored." Sander & Wyner, *supra* note 61 (emphasis in original). *See generally* Andrew Gelman, *Before Data Analysis: Additional Recommendations for Designing Experiments to Learn about the World*, 34 J. Consumer Psych. 187, 190 (2024) ("Overestimation of effect sizes leads to overconfidence in design, with researchers being satisfied by small sample size and sloppy measurements in a mistaken belief that the underlying effect is so large that it can be detected even with a very crude inference.").

Arcidiacono's cautionary assessment of Sander's and Williams's analyses is telling because Arcidiacono has published several papers alleging deleterious effects of race-conscious admissions policies, including mismatch effects. *See* Thaxton, *supra* note 1, at 966 n.432. Arcidiacono's analyses have been plagued by several of the same methodological problems identified in Sander's and Williams's work, and sixteen leading economists, including a Nobel laureate, filed an amicus brief identifying significant methodological shortcomings in Arcidiacono's statistical analyses submitted in litigation over Harvard's undergraduate admissions process. *Id.* (citing 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)). Of the sixteen signatories, only one of them (Guido Imbens of Stanford University) was also a signatory on the Empirical Scholars Brief in *Fisher I*.

Arcidiacono has also claimed that banning affirmative action had a positive effect on graduation rates for minorities in California, *see* Peter Arcidiacono et al., *Affirmative Action and University Fit: Evidence from Proposition 209*, 3 IZA J. Labor Econ. 1, 21 (2014), but his results have not withstood scrutiny, *see* William C. Kidder, *Fact Check and Research Synthesis: Affirmative Action, Graduation Rates and Enrollment Choice at the University of California*, Civil Rights Project, University of California, Los Angeles (Sept. 7, 2020) (summarizing several studies revealing fatal flaws in Arcidiacono's methodology).

69 Empirical Scholars Brief in *Fisher I*, *supra* note 28, at 19; *see also* Thaxton, *supra* note 1, at 961. Richard Lempert and a co-author also sought to publish a reply to Williams's article in the same journal, but the journal's editor refused to publish Lempert's reply, stating that the journal did not publish replies that simply documented flaws in published articles. Brief *Amicus Curiae* for Richard Lempert in Support of Respondents at 9–10 n.5, *Fisher v. Univ. of Tex. at Austin* (*Fisher II*), 136 S. Ct. 2198 (2016) (No. 14-981).

70 Thaxton, *supra* note 1, at 922.

71 Kidder & Lempert, *Mismatch Myth*, *supra* note 14, at 109 (describing bias in Williams's analysis resulting from his choice of an invalid instrument and his removal of the third and fourth

S&S also note that Sander recently published a defense of his prior work in the peer-reviewed journal *International Review of Law and Economics*.[72] However, two responses to that article have been published by the same journal, and both conclude that Sander's article does not establish the empirical support for mismatch theory that he claims.[73] Kidder and Lempert reviewed the extant research on mismatch theory and remarked, "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."[74]

B. On Research Design for Statistical Inference

The aim of empirical research in the social sciences is to learn what conclusions can and cannot be drawn about a population of interest given empirically relevant combinations of assumptions and data.[75] Data alone are insufficient to draw conclusions; inference requires assumptions that relate the data to the population of interest.[76] Moreover, "[t]he credibility of statistical inference *decreases* with the strength of the assumptions maintained…. [And]

---

law school tiers).

72 Sander & Steinbuch, *supra* note 15, at 716 n.1 (citing Sander, *Replication of Mismatch Research*, *supra* note 24).

73 David Bjerk, *Replication of Mismatch Research: Ayres, Brooks and Ho (Comment)*, 58 Int'l. Rev. L. & Econ. 3 (2019); Ian Ayers & Richard Brooks, *Affirmative Action Still Hasn't Been Shown to Reduce the Number of Black Lawyers: A Response to Sander*, 69 Int'l. Rev. L. & Econ. 106032 *1 (2021); *see also* Thaxton, *supra* note 1, at 843 n.25 (noting "in his [peer-reviewed] response to [Daniel] 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)").

74 Kidder & Lempert, *Mismatch Myth*, *supra* note 14, at 106–08 (noting that researching claiming strong support for mismatch has been limited to Sander and "predominantly people tied to him in some way"); *accord* Williams, *supra* note 67, at 172 ("With the exception of Sander (2004), previous research using the BPS has concluded that the balance of evidence is against the mismatch hypothesis."); Empirical Scholars, Brief in *SFFA*, *supra* note 5 (noting that research providing support for mismatch theory has been produced by a very small group of scholars); *see also* Charles et al., *supra* note 45, at 189 ("Although this [mismatch] hypothesis makes intuitive sense, it likewise has not been supported empirically.").

75 Manski, *supra* note 7, at 3; *accord* Elie Tamer, *Partial Identification in Economics*, 2 Ann. Rev. Econ. 167, 192 (2010) ("There is no better time for empirical economists to be clear about what inferences can be made with what assumptions. This will lead to a better empirical program, one that is clear and transparent, combining both the data and valid economic assumptions. This is exactly what is required from any serious scientific program.").

76 Guillermina Jasso, *Formal Theory*, *in* Handbook of Sociological Theory 45–47 (Jonathan H. Turner ed., 2006) ("A theory begins with an assumption. . . . The assumptions of the theory are general and abstract . . . but the theory ends at the most particular and concrete levels. For if it is a good theory, its implications include observable, particular phenomena and relationships."); *see also* Richard B. Braithwaite, Scientific Explanation: A Study of the Function of Theory, Probability and Law in Science ix (1953) (stating that "theoretical concepts . . . are connected to observable facts by complicated logical relationships").

analysts face a dilemma as they decide what assumptions to maintain: Stronger assumptions yield inferences that may be more powerful but less credible."[77] Consequently, it is "tempting for researchers to maintain assumptions far stronger than they can persuasively defend in order to draw strong conclusions because the scientific community rewards researchers who produce strong novel findings and the public, impatient for solutions to its pressing policy concerns, rewards researchers who offer simple analyses leading to unequivocal policy recommendations."[78]

The most important (and sobering) rule of statistical inference is that available data may be inadequate to answer important questions concerning the impact of various policy interventions.[79] One notable statistician famously remarked, "The data may not contain the answer. The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data."[80] Another prominent statistician cautioned, "Often the dataset being used is so obviously deficient with respect to key covariates that it seems as if the researcher was committed to using that dataset no matter how deficient."[81] And while concerns about data inadequacy are relevant to both experiments and observational (i.e., nonexperimental) studies,[82] special care must be taken when analyzing observational data because they are generally fraught with problems that compromise claims of objectivity and revealing scientific truths. "Judgment is the critical input into the analysis of nonexperimental data. Systematic errors

77 Charles F. Manski, Public Policy in an Uncertain World: Analysis and Decisions 12 (2013); *see also* Rebecca B. Morton, Methods and Models: A Guide to the Empirical Analysis of Formal Models in Political Science 38 (1999) ("[Theorists] like to begin with the simplest possible model that can capture the research question before building to a larger, more general, and perhaps less tractable model.").

78 Manski, *supra* note 7, at 2 (quote edited for clarity).

79 Donald B. Rubin, *For Objective Causal Inference, Design Trumps Analysis*, 2 Annals of Appl. Stat. 808 (2008); *see* Moritz Hardt & Benjamin Recht, Patterns, Predictions, and Actions: Foundations of Machine Learning 205 (2022) ("There is no way statistical estimators can recover from poor design. If the design does not permit causal inference, there is simply no way that a clever statistical trick could remedy the shortcoming.").

80 John W. Tukey, *Sunset Salvo*, 40 Am. Stat. 72, 74–75 (1986); *accord* Saul Levmore, *The Eventual Decline of Empirical Law and Economics*, 38 Yale J. on Reg. 612, 614 (2021) ("Some data is not necessarily better than no data if the available data is inspected for the wrong thing.").

81 Rubin, *supra* note 79, at 817; McElreath, *supra* note 33, at 582 n.235 ("The answer is not always in the data. But if you torture the data long enough, it will confess."). *See generally* Lempert, *supra* note 30, at 13 (identifying several (in)famous examples of methodologically unsound empirical studies, including Sander's research on mismatch theory, and noting that 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 regardless of weaknesses in the science").

82 *See* Thaxton, *supra* note 1, at 958–59 (describing flaws that undermine causal inference in both experimental and observational studies).

in the formation of judgment may lead to significant systematic errors in the interpretation of evidence."[83]

Judging the credibility of assumptions is subjective, and analysts make assessments on their own terms; nonetheless, a community of researchers may agree, overwhelmingly, on the credibility of certain assumptions, and this agreement creates a "scientific consensus."[84] It is this consensus that distinguishes a reasonable belief from a mere opinion.[85] The Empirical Scholars Brief in *Fisher I*, signed by eleven theoretical and applied statisticians, begins by stating, "Based on over 255 collective years of social-science research experience, *amici* have concluded that the research on Professor Sander's 'mismatch' hypothesis is unreliable, failing basic tenets of research design."[86] Similarly, the Empirical Scholars Brief in *SFFA*, signed by eight theoretical and applied statisticians, explains that "Sander's research has major methodological flaws—misapplying basic principles of causal inference . . . [and] fails to meet the basic tenets of rigorous social science research. The more recent work of other researchers that Sander relies on in his *amicus* brief does not support the mismatch hypothesis either."[87] It is revealing that the Empirical Scholars Briefs in *Fisher I* and *SFFA* thoroughly engage both the seminal methodological literature establishing the rules for credible statistical inference *and* studies alleging both support and lack of support for mismatch.[88] By contrast, Sander's briefs in *Fisher I* and *SFFA* selectively (and

83 Edward E. Leamer, Specification Searches: Ad Hoc Inference with Nonexperimental Data 307 (1978); *accord* Ben Jones, Avoiding Data Pitfalls: How to Steer Clear of Common Blunders When Working with Data and Presenting Analysis and Visualizations 3 (2019) ("We often approach data with the wrong mind-set and assumptions, leading to errors all along the way, regardless of what [methods] we choose.").

84 Manski, *supra* note 77, at 12.

85 Jones, *supra* note 83, at 3; *accord* Morton, *supra* note 77, at 38 ("Some of the assumptions in [theories] are considered to be true—or there is a large amount of evidence that they are true.").

86 Empirical Scholars Brief in *Fisher I*, *supra* note 28 and accompanying text (quote edited for clarity). Sander referred to the signatories of the Empirical Scholars Brief in *Fisher I* as "an eminent group" and acknowledged "the prestige they brought to the enterprise." Sander, *Mismatch and the ESB*, *supra* note 8, at 566; *see also* Thaxton, *supra* note 1, at 956 ("Many of the signatories [of the Empirical Scholars Brief in *Fisher*] 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.").

87 Empirical Scholars Brief in *SFAA*, *supra* note 5 and accompanying text. Sander has speculated that "most of the signatories of the ESB [in *Fisher I*] 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." Sander, *Stylized Critique*, *supra* note 8, at 1664. Sander's speculation has been refuted by the fact that most of the signatories of the Empirical Scholars Brief in *Fisher I* were amici, ten years later, in the Empirical Scholars Brief in *SFFA*.

88 Empirical Scholars Brief in *Fisher I*, *supra* note 28; Empirical Scholars Brief in *SFAA*, *supra* note 5; *see generally* James N. Druckman, Experimental Thinking 161 (2022) (noting that credible research design and implementation requires "an exhaustive search of research on

incompletely) engage studies that identify fundamental errors in research proclaiming empirical support for mismatch theory and fail to engage the core social science literature that developed the conceptual and methodological toolkit for credible statistical inference.[89]

In determining the adequacy of data and the credibility of assumptions for the purposes of assessing the impact of race on bar exam success (as well as other law school-related outcomes), the analyst must carefully consider the design of the (quasi-)experiment.[90] Such consideration is important because "it is crucial to ensure that research design and analysis are sufficiently rigorous to allow for causal interpretation of results."[91] Indeed, at the foundation of credible statistical inference is thinking clearly about what the data can reasonably tell us "before attacking it with mathematical analysis or available computer programs because the blind use of complicated statistical procedures is doomed to lead to absurd conclusions."[92] Economist Joshua Angrist has

---

the topic and detailed descriptions of specific studies . . . across fields and methods to assess the state of knowledge and the strength and weaknesses of different approaches").

89 *See* Brief *Amici Curiae* for Richard Sander & Stuart Taylor, Jr. in Support of Neither Party, *Fisher I*, 570 U.S. 297 (2013) (No. 11-345) [hereinafter Sander, Brief in *Fisher I* (Revised)]; Sander, Brief in *SFFA*, *supra* note 13. *Compare* Kidder & Lempert, *Mismatch Myth*, *supra* note 14, at 108 (noting that Sander "acknowledged problems with his data and models, but [that] his concessions have never extended to acknowledging fundamental flaws"); *accord* Thaxton, *supra* note 1, at 994, 960 ("Sander's. . . work . . . nibbl[es] at the margins of the scholarship of those who have revealed persistent and seemingly fatal flaws in his research, while avoiding tackling the fundamental shortcomings that have plagued the entire enterprise. . . . [T]he 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.").

The lack of meaningful engagement with the relevant methodological literature establishing the rules and best practices of statistical inference to justify his experimental design, analyses, or reasoning characterizes nearly all of Sander's prior responses to critiques. *See, e.g.,* Sander, *Reply to Critics*, *supra* note 61 (responding to several critiques of "Systemic Analysis" without including any discussion of the core literature establishing the rules of credible statistical inference); *accord* Sander, *Mismatch and the ESB*, *supra* note 8 (replying to the Empirical Scholars Brief in *Fisher I* but neither referencing nor discussing the key methodological literature). To be clear, Sander references and discusses prior empirical studies that examine mismatch; however, these references/discussions fail to establish that Sander's work, or the work of others who report findings supportive of mismatch, adheres to the basic rules and best practices of credible statistical inference.

90 *See generally* Gerald van Belle, Statistical Rules of Thumb 2 (2008) ("Any statistical [inquiry] must address the following questions: What is the question? Can it be measured? Where, when, and how will you get the data? What do you think the data are telling you?").

91 Fredrik Falkenström et al., *Using Copulas to Enable Causal Inference from Nonexperimental Data: Tutorial and Simulation Studies*, 28 Psychological Methods 301, 302 (2023).

92 Thaxton, *supra* note 1, at 967; *accord* Jim Granato et al., *A Framework for Unifying Formal and Empirical Analysis*, 54 Am. J. Pol. Sci. 783, 784 (2010) (arguing that the identification of causal mechanisms has been negatively impacted by "the use of hand-me-down applied statistical techniques . . . and an overall inattention to relating theoretical specifications to applied statistical tests"); Edward E. Leamer, *Tantalus on the Road to Asymptopia*, 24 J. Econ. Persp. 31,

noted that "the primary force driving the credibility revolution [in empirical economics] has been a vigorous push for better and more clearly articulated research designs."[93]

A full discussion of research design is beyond the scope of this article,[94] but it is useful to provide a brief overview of four important components of credible research design. This overview will allow readers to better situate the ensuing assessment of S&S's study: (1) identification of the exact outcome and treatment conditions; (2) suitable sample sizes for comparison of the treatment and control groups; (3) collection and proper measurement of key covariates; and (4) balance on key covariates across the treatment and control groups that are to be compared. Each component is discussed in turn below.

(1) Component 1: Identification of the exact outcome and treatment conditions

The first component is the identification of the exact outcome variable and treatment conditions—e.g., the real or hypothetical experimental manipulation around questions of race.[95] There are two related problems with S&S's study as it pertains to this component.[96] The first problem is that the study does not examine the impact of race-conscious affirmative action on racial differences in students' entering credentials or other mismatch-related outcomes.[97] Nonetheless S&S claim that "the racial effects we have been discussing occur because current admissions practices tend to afford preferences more often to

33 (2010) (offering "some warnings about the willingness of applied economists to apply push-button methodologies without sufficient hard thought regarding their applicability and shortcomings").

93 Joshua D. Angrist & Jörn-Steffen Pischke, *The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics*, 24 J. Econ. Persp. 3, 10 (2010); *accord* Cristobal Young, *Model Uncertainty and the Crisis in Science*, 4 Socius 1, 5 (2018) ("The 'crisis in science' [i.e., replication crisis] today is rooted in genuine problems of model uncertainty and lack of transparency.").

94 For a detailed discussion of the core requirements for credible statistical inference in studies of mismatch effects, *see* Thaxton, *supra* note 1, at 984–93.

95 Prior research conducted by Sander also suffers from failing to exactly identify the treatment condition, thereby undermining the inferences he has drawn about the relationship between race, mismatch, and bar passage. Thaxton, *supra* note 1, at 940 (discussing problems with the first-choice analysis that Sander called the best test of mismatch using BPS data, despite clear evidence of a significant problem in the manner in which the treatment condition was defined).

96 *See generally* Miguel A. Hernan & James M. Robins, Causal Inference: What If vii (2023) ("Unfortunately, the scientific literature is plagued by studies in which [1] the causal question is not explicitly stated and [2] the investigators' unverifiable assumptions are not declared. This casual attitude towards causal inference has led to a great deal of confusion.").

97 Druckman, *supra* note 88, at 164 ("The researcher should directly connect the design and measures to the hypotheses.").

racial minorities."[98] California's ban on race-conscious admissions policies took effect for the 1997 entering classes, so students admitted to the two California law schools in their study after 1996 (i.e., graduating cohorts after 1999)—which includes all but one of the cohorts in their data—were not admitted under a race-conscious affirmative action regime.[99] In other words, there is no variation in affirmative action rules to identify the effects of the policy. That is, there is no "treatment" group here. As a consequence, no direct inferences about the effects of affirmative action can be made between the California schools for seven of the eight cohorts.[100] To be sure, the third school in S&S's study, located in Arkansas, did not ban affirmative action during the time the data were collected. However, any comparison of students in California with students in Arkansas to directly test the impact of affirmative action policies is inappropriate given the incommensurability of both bar-takers and outcomes (counterfactuals).[101]

S&S investigate the effect of disparities in entering credentials on bar exam performance rather than test the effect of affirmative action. The effect that S&S examine may have policy relevance for law school admissions practices—provided S&S adhere to the other rules of statistical inference—but it cannot support S&S's conclusion that racial preferences in admissions practices, not race, produce mismatch effects that, in turn, explain large racial disparities in bar passage.[102] To justify such an inference, S&S would have to test whether

98 Sander & Steinbuch, *supra* note 15, at 739; *see also* Sander, Brief in *SFFA*, *supra* note 13, at 29. (stating that "we [S&S] have a strong finding that it is preferences, not race, which explain large racial disparities in educational outcomes.").

99 *See supra* note 22. For UCLA, S&S obtained six years of data (2000–2005) and analyzed three graduating years (2000, 2001 & 2005); therefore, all three of the cohorts enrolled after the ban on race-conscious affirmative action took effect in 1997. Sander & Steinbuch, *supra* note 15, at 743. For Davis, S&S obtained fifteen years of data (1997–2011) grouped into three-year admissions blocks (for a total of five cohorts); therefore, only the first cohort grouping (1997–1999) was admitted before the ban on race-conscious affirmative action. *Id*.

100 *See* Daniel E. Ho, *Why Affirmative Action Does Not Cause Black Students to Fail the Bar*, 144 Yale L.J. 1997, 1998 (2005) (noting that Sander's 2004 study of schools in the BPS all employed some system of affirmative action, so the study could not draw conclusions about effects of the policy).

101 *Compare* Sander & Steinbuch, *supra* note 15, at 724 ("We obtained data from three law schools . . . on the credentials of each student at the school who sat for the *in-state* bar exam over multiple years . . . . [A]t each of these schools the vast majority of graduates take the *in-state* bar exam, and we know the pass/fail outcome of those exams.") (emphasis added) *with* Empirical Scholars Brief in *Fisher I*, *supra* note 28, at 21 ("The estimate for how a *black* student at Yale Law School would have performed at a lower-tier school might be based on a white student at the University of Alabama Law School . . . violates the principle of creating groups that are comparable in all pre-existing respects except for law school [eliteness]."); *see also* Thaxton, *supra* note 1, at 1021 ("[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?") (citation omitted).

102 Sander & Steinbuch, *supra* note 15, at 740; *accord* Sander, Brief in *SFFA*, *supra* note 13, at 29.

race-conscious affirmative action produces racial disparities in entering credentials which, in turn, cause Black and Latinx students to fail the bar because of mismatch theory's posited causal mechanism: learning mismatch.[103]

Other key factors in law school admissions policies are relevant, which S&S leave aside in their testing. Following the ban on race-conscious affirmative action in California, law schools were still permitted to consider student-specific factors that were correlated with race, such as "a personal or family history of cultural, educational, and socioeconomic disadvantage."[104] Consequently, it does not necessarily, or even logically, follow that observed racial disparities in entering credentials stem from the consideration of race, per se, in law school admission decisions.[105] As noted earlier, nearly all of the students enrolled in the two California law schools were admitted after the ban on race-conscious affirmative action was enacted.[106]

In prior work, Sander made different assumptions about school selectivity based on the race of the students he assessed. Namely, Sander implicitly assumed that when Black and Latinx students attend a more selective school than white students with the same entering credentials, the higher selectivity for minority students is due to racial preferences; however, when white students attend a more selective school than other white students with the same entering credentials, this higher selectivity for undifferentiated white students is attributable to unobserved quality differences between students.[107] This divergent interpretation underscores the second problem with S&S's study: their conceptualization of preferences. S&S's definition of preferences as "credential deficits" is both over- and underinclusive at the student- and school-level, because (1) admissions committees engage in holistic assessments of applicants' files and (2) university quality encompasses a variety of measures related to selectivity and school resources. At the student level, each law school applicant represents a bundle of attributes that admissions committees

103 *See* Part III.A.

104 *See* Yagan, *supra* note 22, at 40. As noted at the outset, the Supreme Court's majority opinion in *SFFA* reiterated that universities are still permitted to consider how an applicant's race may have specifically influenced her life in ways that the institution deems important. *See supra* note 3.

105 Sander has been a vocal proponent of class-based affirmative action. *See* Richard Sander, *Class in American Legal Education*, 88 Denver U. L. Rev. 631 (2011); Richard Sander, *Experimenting with Class-Based Affirmative Action*, 47 J. Legal Educ. 472 (1997).

106 In an earlier draft of their article, *see infra* note 333, S&S present a table with (somewhat) racially disaggregated data—Black versus non-Black—on entering credential disparities and bar exam success (comparing one of the California schools and the Arkansas school). These data reveal that a considerable number of non-Black law students have sizable entering credential deficits. S&S did not disaggregate white from Latinx and Asian bar-takers. This would have been helpful in determining whether the non-Black bar-takers with large entering credential deficits were white or members of some other racial group (e.g., Latinx or Asian).

107 Ayres & Brooks, *supra* note 65, 1854 n.82.

consider in an effort to create class cohorts with a diversity of legally relevant perspectives, skills, interests, and objectives. Moreover, undergraduate grades (UGPA) and entrance exam scores (LSAT) have been shown to be poor criteria of merit and poor predictors of law school success.[108] Continued reliance on these metrics is driven by considerations other than student success in the classroom, on the bar exam, and in the legal profession.[109]

At the school level, institutional quality is understood to include a wide range of factors in addition to median student entering credentials—such as tuition amounts, acceptance rates, public/private status, faculty-student ratios, faculty salaries, *et cetera*.[110] There is also a growing literature documenting the persistence of racial disparities in UGPA and LSAT, even after taking into account relevant variables.[111] The fact of such disparities suggests that students' entering credentials are themselves impacted by race and, consequently, interracial comparisons that hold students' level of mismatch constant are invalid.[112] S&S's reliance on the UGPA and LSAT of the median student as a baseline for both merit and aptitude for legal learning misrepresents the actual admissions calculus, and Sander has acknowledged that law schools consider factors beyond UGPA and LSAT when making admissions decisions.[113] Moreover, S&S's recognition that "there are many Black and Hispanic

108 *See* Part III.A.

109 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, 1065 & 1121 (2001) (noting that an LSAC study discovered that the majority of law schools misuse the LSAT in law school admissions and that law school rankings, such as those of the *U.S. News & World Report*, create a feedback loop that encourages overreliance on the LSAT in ways that were not justified given the test's very limited predictive ability of first-year LGPA); *accord* Alex M. Johnson Jr., *The Destruction of the Holistic Approach to Admissions: The Pernicious Effects of Rankings*, 81 Ind. L. J. 309, 324-348 (2006) (explaining that, from its inception, the LSAT was not intended to be used to the exclusion of other relevant admissions-related factors but that *U.S. News* rankings create strong incentives for law school deans to place disproportionate weight on the LSAT).

110 Eleanor Wiske Dillon & Jeffrey Andrew Smith, *Determinants of the Match Between Student Ability and College Quality* 35 J. Labor Econ. 45, 46 (2016); *accord* Thaxton, *supra* note 1, at 931–32.

111 *See infra* note 181 and accompanying text.

112 In statistics parlance, UGPA and LSAT are endogenously determined, meaning that they are determined, at least in part, by other variables in the model (i.e., race). *See generally* Tyler J. VanderWeele & Whitney R. Robinson, *On the Causal Interpretation of Race in Regressions Adjusting for Confounding and Mediating Variables*, 25 Epidemiology 473, 476 (2014) (explaining that isolating the effect of race entails controlling for other variables that correlate with race and the outcome, but that are not themselves affected by race).

113 *Compare* Sander, *Reply to Critics*, *supra* note 61, at 1971–72 ("I realized that I had made two fundamental errors in my analyses. First, I had overlooked the problem of 'unobserved [student] characteristics'.") *with* Sander, *Systemic Analysis*, *supra* note 46, at 409 ("The academic index for applicants—however it might be constructed by individual schools—is always the dominant factor in admissions within each racial group; other 'soft' factors play a prominent role only for those relatively few cases that are on the academic score boundary between 'admit' and 'reject'.").

students who attend law schools without a preference, and there are some whites and Asians who receive large preferences,"[114] simply underscores, rather than remedies, the defect in their definition of an admissions preference.

(2) Component 2: Suitable sample sizes for comparison of the treatment and control groups

The second component of a credible research design to test the relationship between race, mismatch, and bar passage is that sample sizes for the different racial groups that are to be compared must be sufficient to learn anything of interest from the data. Sample sizes must allow the statistical test to have enough power to reject the null hypothesis. A large *p* value (i.e., a statistically insignificant finding) can occur either because the null hypothesis is "true" or because the null is false and the test is not powerful enough to detect any meaningful difference.[115] In addition, the ratios of the sample sizes needed to obtain well-matched samples is relevant. Given the very small number of Black bar-takers in most of the mismatch groupings in S&S's study, it is implausible that much can be learned from the comparisons.[116]

(3) Component 3: Collection and proper measurement of key covariates

The third component is the collection and proper measurement of key explanatory variables (i.e., covariates). Lack of information on important covariates[117] or poorly measured/coded covariates distorts the relationship between the key variable(s) of interest and the outcome variable.[118] This same lack of information results in an inadequate account of the dynamics that generated the observed data—what empiricists refer to as "poor model fit."[119] S&S fail to provide or assess data on important (and properly measured) covariates in ways that seriously challenge their claims of model fit necessary to support their conclusions.

S&S claim that "two generations of aggressive affirmative action . . . is the most plausible explanation for a substantial proportion of the [racial] gap" in bar passage rates (particularly for Black students),[120] and they use race as proxy for race-conscious affirmative action, as Sander has done in prior

114 Sander & Steinbuch, *supra* note 15, at 739.

115 *See infra* Part III.D; *see also infra* notes 367–370 and accompanying text.

116 *See infra* note 333; *see also* Thaxton, *supra* note 1, at 900 n.238 & 961–63 (noting that analyses of mismatch theory based predominately on white law students and lawyers in the sample may be inapplicable to Black students and lawyers).

117 David P. MacKinnon et al., *Equivalence of the Mediation, Confounding and Suppression Effect*, 1 Prevention Sci. 173, 173–74 (2000) (explaining that covariates take the form of confounders, suppressors, and mediators).

118 *See infra* Part III.D; *see infra* notes 321–324 and accompanying text.

119 *See infra* Part III.C.

120 Sander & Steinbuch, *supra* note 15, at 740.

research. It has been well established that using an immutable characteristic, such as race, when making comparisons between law students can bias causal inference.[121] The immutability of race has been raised in prior critiques of Sander's work using different data, including in the assessment provided in the Empirical Scholars Brief in *Fisher I*.[122] And Sander himself has both modified his analyses in response to this critique and spoken favorably about another study that purportedly avoided making interracial comparisons when examining the effect of mismatch on bar exam passage.[123] Still, S&S claim that the relationship between race and bar passage is attributable to mismatch, and when comparing Black students with white students who are similar in their degree of mismatch, any effect of race disappears.

The problem with their assertion—as scholars have observed—is that "Black students could differ from white students in all sorts of respects that are not due to affirmative action or mismatch,"[124] and "the existence of important unobserved differences between these students that affect bar performance invalidates the estimates."[125] Because S&S include a very limited number of potentially confounding variables in their analysis,[126] their results are based on the implausible assumption that Black and Latinx students are similar to white students with respect to all relevant pre-mismatch characteristics.[127] To the contrary, multiple studies that examined the BPS data discover that

121 Some argue that immutable characteristics, such as race and gender, cannot "cause" exposure to discrimination because it cannot be subject to human manipulation, but this perspective about causality leads to counterintuitive ideas about causation—e.g., the moon cannot cause the tides, and natural disasters (e.g., tornadoes, hurricanes, and earthquakes) do not cause destruction of property, etc. Kenneth A. Bollen, Structural Equations with Latent Variables 41 (1989). Human manipulation can aid in causal inference, but it is by no means a requirement. *Id*. Three conditions are necessary for causal inference: (1) conditional independence (i.e., nonspuriousness), (2) association (i.e., empirical correlation), and (3) direction of influence (i.e., cause must precede the effect in time). *Id*.; *accord* Thaxton, *supra* note 1, at 848 n.49.

122 Brief in *Fisher I*, *supra* note 28, at 20 ("[T]he primary comparison that Sander and Williams employ is that of black and white students . . . . The existence of important unobserved differences between these students that affect bar performance invalidates the estimates.").

123 Sander, Brief in *Fisher II*, at 27 (acknowledging that interracial comparisons may fail to control for important interracial differences that impacted Sander's initial examination of mismatch effects and identifying another study of mismatch by Doug Williams, *supra* note 67, as avoiding this problem by only making intraracial comparisons); *see also* Richard H. Sander, *Mismeasuring the Mismatch: A Response to Ho*, 114 Yale L.J. 2005, 2007 (2005) (responding to Daniel Ho's critique of the use of interracial comparisons with a reanalysis that made only intraracial comparisons) [hereinafter Sander, *Response to Ho*].

124 Ho, *supra* note 28, at 2012.

125 Brief in *Fisher I*, *supra* note 28, at 20.

126 *See infra* Part III.C; *see infra* note 118 and accompanying text.

127 S&S's analysis of a very limited set of control variables is odd considering that Sander previously praised the BPS data for having "hundreds of variables on tens of thousands of law students." Sander & Taylor, at *supra* note 22, at 57.

any alleged mismatch effects disappear after properly accounting for relevant confounding variables.[128]

Finally, bias resulting from S&S's failure to account for important explanatory variables in their models is exacerbated by S&S's improper adjustment for the consequences of race. These consequences of race are called "post-treatment variables" and the bias stemming from the incorrect handling of these variables is commonly referred to as "post-treatment bias." Because race is determined at a person's birth, all measured variables specific to a person (e.g., UGPA, LSAT, and LGPA) are, by definition, post-treatment variables.[129] Political scientists Jacob M. Montgomery, Brendan Nyhan, and Michelle Torres emphasize that "concerns about post-treatment bias are not really (or only) about the post-treatment variable itself. The problem is that by conditioning on a post-treatment variable, we have unbalanced the treatment and control groups with respect to every other possible [unobserved] confounder."[130] To properly measure the impact of race on bar exam success, S&S must take into account all covariates that are predictive of both bar exam success and the intermediate outcomes.[131]

128 *See infra* note 323.

129 Vanderweele & Robinson, *supra* note 112, at 476 ("To isolate the effect of race . . . we would want to control for all attributes occurring before conception, but not after conception because anything occurring after conception would be affected by the variables constituting race."). Post-treatment variables, "in this context, [are] any factor, attribute, personality trait, or personal or professional experience that could potentially be a consequence of race[.]" Maya Sen & Omar Wasow, *Race as a Bundle of Sticks: Designs that Estimate Effects of Seemingly Immutable Characteristics*, 19 Ann. Rev. Poli. Sci. 499, 505 (2016). This does not imply, however, that all post-treatment variables are impacted by race or that one cannot measure the impact of race on an outcome that operates, in part, through a post-treatment variable. Thaxton, *supra* note 1, at 946 n.353; Constantine E. Frangakis & Donald B. Rubin, *Principal Stratification in Causal Inference*, 58 Biometrics 21, 27 (2002) (describing a framework "for comparing treatment effects on outcomes [when] adjusting for posttreatment variables"). But naïvely adjusting for a post-treatment variable that both is impacted by race and influences the outcome will bias the measurement of a race effect on the outcome. Thaxton, *supra* note 1, at 1013–16. There is substantial evidence that students' law school grades are affected by race, independent of entering credentials and law school difficulty/selectivity, so interracial comparisons that naïvely hold students' LGPA constant are invalid. *See infra* notes 156–186 and accompanying text.

130 Jacob M. Montgomery et al., *How Conditioning on Posttreatment Variables Can Ruin Your Experiment and What to Do about It*, 62 Am. J. Pol. Sci. 760, 762–63 (2018); *accord* Herbert I. Weisberg, Bias and Causation: Models and Judgment for Valid Comparisons 205 (2010) (explaining that statistically adjusting for a post-treatment variable generally makes the treatment and control groups non-exchangeable); *see also* 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.").

131 Thaxton, *supra* note 1, at 859. Additionally, S&S's coding of the mismatch variable (i.e., measurement) is not properly justified in their article and likely capitalizes on noise in the data. *See infra* Part III.C; *see also infra* notes 345–354 and accompanying text.

(4) Component 4: Balance on key covariates across the racial groups that are to be compared

The fourth component is balance on key covariates across the racial groups that are to be compared.[132] With adequate balance on the covariates, racial groups are similar with respect to all confounders so that the comparisons (i.e., counterfactuals) are driven by the data. Without such balance, any inference will be based on assumptions that are unsupported by the data, and any results will be highly sensitive to modest alterations in those assumptions or in the examination of another representative sample drawn from the same population. S&S do not offer evidence that the racial groups in their data are balanced on the limited number of variables they include in their models. And I as explain, *infra*, an earlier draft of S&S's article reveals that there are very few comparable Black students at the various levels of mismatch.[133] One scholar has cautioned, "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."[134]

To be sure, there is often tension between minimizing confoundedness and minimizing covariate imbalance because achieving balance is more difficult as more potential confounding variables are included in the analysis. The root of the problem is that as more explanatory variables are included in the model, stronger assumptions about the interrelatedness of those variables and the outcome of interest are needed to properly estimate the effects of those variables. This tension often grows taut when it comes to race. When racial groups differ sharply with respect to the distributions of values for confounding variables, statistical analysis may not credibly adjust for confounding variables. "An extreme case of imbalance occurs when the ranges of data values of the two covariate distributions by [group] differ . . . . More typical, even if the ranges of data values of the covariate distributions [across the two groups] are identical, there may be substantial differences in the shapes of the covariate distributions by [group] status."[135] Nonetheless, credible statistical inference about the impact of race on law school-related outcomes requires that the racial groups being compared are similar along all relevant dimensions *and* that the relevant comparisons are present in the data under investigation. Otherwise, it is implausible that the analysis can meaningfully distinguish between the

132 Levmore, *supra* note 80, at 614 ("Econometrics is a game of looking for treatment effects, but doing so requires that the groups being compared are truly comparable, so that the groups are exchangeable and attention can be focused on the variable being studied.").

133 *See infra* note 333 and accompanying text.

134 Joseph S. Nye, Jr., Understanding International Conflicts: An Introduction To Theory And History 52 (4th ed. 2003).

135 Guido W. Imbens & Donald B. Rubin, Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction 337 (2015). *See infra* Part III.C; *see infra* notes 404–405 and accompanying text.

effect of race and the effects of other confounding variables with the available data.

* * *

Taking the aforementioned components of credible research design seriously is a *must* if the analyst desires to "steer clear of common blunders when working with data and presenting analysis."[136] S&S note that the BPS data are the primary data used for many articles examining mismatch in law schools, which is understandable given that Sander centered his earlier work on the BPS. Acknowledging the problem with the prior data for their research design, S&S posit that these prior studies were negatively impacted by use of a crude (i.e., non-school-specific) measure of mismatch, and that this limitation biased findings of the impact of mismatch toward zero.[137] S&S argue that a school-specific measure of mismatch would likely produce strong support for their mismatch hypothesis.[138] As I have explained previously, it is incorrect, as a

136 Jones, *supra* note 83.

137 Sander & Steinbuch, *supra* note 15, at 718. *But cf.* Eric Loken & Andrew Gelman, *Measurement Error and the Replication Crisis*, 355 Sci. 584, 585 (2017) (noting that "results from. . . studies should not be defended by saying that they would have been even better with improved measurement").

S&S do not mention that Sander's assumption of a tier hierarchy in the BPS data that roughly reflects the selectivity of the schools and the academic strength of the students attending those schools is demonstrably false. Thaxton, *supra* note 1, at 933 (noting that the typical law schools in "the second and third tiers are revealing: The[se] two tiers have identical LSAT scores (39; 76th percentile), and the third tier centroid [i.e., typical school] 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."). Moreover, Sander's incorrect assumption leads to bias *in favor* of mismatch, not against it. *Id.* at 935 ("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."); *see also* Part III.E (explaining that the school-level effects influence bar exam performance independent of mismatch and have a stronger impact than mismatch when examining the most plausible counterfactuals).

S&S acknowledge that the coarseness of the measure of mismatch in the BPS data may have been necessary to protect student anonymity and shield law schools from possible litigation. Sander & Steinbuch, *supra* note 15, at 721 n.11. Nonetheless, they also suggest that the BPS was "sanitized, as it were, at its inception," "engineered to be a second-rate research tool," and "cripple[d] [by the LSAC]" in order to, at least in part, prevent researchers from directly examining mismatch effects on bar passage. *Id.* at 721.

But the concerns over anonymity should not be downplayed. When rejecting Sander's request for California bar admissions data, the California Court of Appeals noted that Sander's own experts were unable to develop a plan for access to the relevant data that would reasonably protect bar applicants' privacy interests. *Sander v. State Bar of Cal.*, 237 Cal. Rptr. 3d 276, 291–92 (Cal. Ct. App. 2018); *see also supra* note 18 (noting that the BPS deemed grouping schools necessary because the samples within individual schools were so small for ethnic groups as to preclude meaningful cross-racial comparisons).

138 Sander & Steinbuch, *supra* note 15, at 718 (asserting that "the weaker measures of mismatch used before [with the BPS data] would tend to bias mismatch downward") (citing

statistical matter, to conclude that an error-prone measure of mismatch would necessarily bias estimates of the mismatch effect toward zero.[139] In multivariate settings, the direction and magnitude of such bias depend on other factors.[140]

Assuming, *arguendo*, that S&S are correct about the problems with the BPS data when examining mismatch effects (i.e., downward bias), their analysis of the school-specific data is still incapable of supporting the conclusions they advance because of the deficiencies with their research design and data (i.e., ineffective experimental design), deficiencies with their methods (i.e., inappropriate analyses), and deficiencies in their logic connecting the data and methods to their conclusions (i.e., flawed reasoning).[141] In the following sections, I describe how and why these various shortcomings undermine S&S's claim of disentangling "race effects" from "mismatch effects" when examining students' bar exam performance.

III. Six Methodological Flaws Impacting S&S's Analyses

A. Unpacking Racial Disparities in Bar Exam Success

Pioneering statistician John Tukey remarked, "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."[142] S&S's analytical approach is inadequate for the task of explaining racial differences in bar exam performance because, fundamentally, (1) they neglect to empirically evaluate critical assumptions about the direct effect of race on law school grades (LGPA) and the indirect effect of race on bar exam performance that operates through LGPA;[143] and

Arcidiacono & Lovenheim, *supra* note 63, at 20).

139 Thaxton, *supra* note 1, at 928; *see also* Loken & Gelman, *supra* note 137, at 584–85 (explaining that effect sizes, in the presence of measurement error, may be biased upward or downward because researchers can engage in data-mining to find the desired effect and neglect to report unsupportive effects).

140 Thaxton, *supra* note 1, at 928; *see also* Raymond J. Carroll et al., Measurement Error in Nonlinear Models: A Modern Perspective 41 (2d ed. 2006) ("[T]he 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."); *accord* Loken & Gelman, *supra* note 137, at 585 (noting that the impact of measurement error "becomes more complicated with multiple predictors, or with non-independent errors").

141 *See supra* note 30.

142 John W. Tukey, *The Future of Data Analysis*, 33 Annals Mathematical Stat. 1, 13–14 (1962).

143 *See generally* Morton, *supra* note 77, at 145 ("Researchers discover the limits of theory when they evaluate nonverified assumptions, and such analysis can be quite valuable in the quest to understand [social phenomena] more fully."). Race potentially has a "conditional indirect effect" on bar exam performance via LGPA—i.e., the mediation effect varies in strength conditional on the race of the student. In such situations, the path from the intervention

(2) they do not assess mismatch theory on its own terms, which requires the careful examination of the relevant links in the hypothesized causal chain of events: *Aptitude Deficit* → *Knowledge Deficit* → *Competency Deficit* (see Figure 1).[144] These are the right questions to ask.

S&S ignore these causal processes and focus on the much less informative question about the alleged "black box" effect of aptitude deficit on bar exam performance: *Aptitude Deficit* → *[Black Box]* → *Competency Deficit* (see Figure 3).[145] This is the wrong question to ask. In this section, I explain why S&S's analytical framework is incapable of providing meaningful answers to understanding either racial disparities in bar exam performance or mismatch effects, more generally.[146]

---

to the mediator is constant (RACE → LGPA), whereas the effect of the mediator on the outcome depends on the level of the intervention (RACE × LGPA → BAR). *See infra* note 525 and accompanying text.

Scholars studying these effects have also consistently discovered that race exerts an impact on UGPA and LSAT, independent of relevant third variables, such that the racial gap in entering credentials may be biased proxies for racial differences in aptitude for legal learning. *See infra* note 181.

144 Lempert et al., *supra* note 38, at 4 ("The three-step argument that Sander offers to support his argument against affirmative action—(1) LSAT and UGPA are important in explaining first-year law school grades; (2) Grades are important in explaining bar passage; and (3) Therefore, affirmative action harms African Americans by putting many in law schools where their LSATs and UGPAs will cause them to receive low grades and fail the bar exam."); *accord* Sander, *Response to Ho*, *supra* note 123, at 2006. *See generally* Granato et al., *supra* note 92, at 784 (stating that the disconnect between theoretical and empirical analysis stems from "an overall inattention in relating theoretical [models] to applied statistical tests" and that "the distinctive feature of causal models that each variable is determined by a set of other variables through a relationship (called a 'mechanism')"). Sander has employed a causal process approach in prior work when examining mismatch effects, *see* Sander, *Mismeasuring Mismatch*, *supra* note 123, at 2007-2008, so it is odd that S&S did not adopt this analytical framework.

Competency Deficit (proxied by bar exam performance) and *Knowledge Deficit* (proxied by LGPA) are, respectively, primary and secondary outcomes in S&S's analytical framework. A primary outcome is the behavioral change attributable to the intervention—it is the motivating question of the analysis. In contrast, a secondary outcome typically informs theories/hypotheses and assists in the interpretation of findings. Secondary outcomes also aid in the assessment of the overall effects of the intervention and are often intermediate (i.e., a mediator/mechanism) to the primary outcome. These secondary outcomes indicate whether the intervention is acting in the expected manner. *See generally* Field Trials of Health Interventions: A Toolbox 201-206 (Peter G. Smith et al. eds., 3d ed. 2015) (noting that secondary outcomes "can be very useful to generate further hypotheses" and that "changes in knowledge or attitudes are sometimes an important initial step before a behavior is changed") [hereinafter Field Trials Toolbox]; *see also infra* note 263 and accompanying text.

145 Figure 3 accounts for students' absolute and relative (i.e., school-specific) differences in aptitude for legal training by modeling law school selectivity/difficulty.

146 *See generally* Charles C. Ragin, Analytic Induction for Social Research 1 (2023) ("Social scientists ask diverse kinds of questions. Usually, each such question calls for application of a specific analytical strategy to empirical evidence.").

Sander's key claim in *Systemic Analysis*,[147] which he repeated in his book-length treatment of the subject, *Mismatch*,[148] is that racial disparities in law school-related outcomes can be explained by mismatch theory. For him, after accounting for a student's degree of mismatch, racial differences in law school-related outcomes–such as law school grades, likelihood of graduation, and the likelihood of bar passage–disappear.[149] S&S's current article repeats this claim: "mismatch is not about race, but about large admissions preferences that create big credential gaps between the students who receive preferences and their classmates."[150] Specifically, S&S contend (as noted above) that "learning mismatch"–proxied by LGPA–links "credential gaps" to bar exam performance in a causal chain because law professors teach to the "middle" student, such that students with large credential deficits are outmatched by their peers and learn less.[151] As explained below, their claim is not supported by available evidence.

It is true that Black, Native American, and Latinx students, on average, have lower UGPAs and LSAT scores than white and Asian students.[152] But it is equally true that nearly every credible study conducted at both specific law schools (e.g., Berkeley, Columbia, Michigan, and Penn) and on clusters of law schools (BPS[153] and National Survey of Law Student Performance[154] (NSLSP))

147 Sander, *Systemic Analysis*, *supra* note 46.

148 Sander & Taylor, at *supra* note 22, at 86.

149 Thaxton, *supra* note 1, at 866 n.113 & 1013; *see also* Sander, *Mismeasuring, supra* note 123, at 2007 (stating that he "is concerned about independent variables indirectly affecting one another"). *But compare* Theodore Cross, *The Good News that the Thernstroms Neglected to Téll*, 42 J. Blacks in Higher Edu. 80, 89 (2003) (noting that the graduation rates of white and Black students were practically identical at the most competitive schools although fewer than 1% of Black students at these schools score at 165 or better on the LSAT (compared with over 8% of non-Black students)).

150 Sander & Steinbuch, *supra* note 15, at 739.

151 *See supra* notes 45 & 46.

152 Thaxton, *supra* note 1, at 909. *But see infra* note 181 (noting that Black, Latinx, and Native American students score lower on the LSAT than expected based on, inter alia, UGPA, college major, and undergraduate institution).

153 *See supra* note 18 and accompanying text for description of BPS data.

154 Sander and colleagues compiled the NSLSP data with "the "principal motivation for the study [being] the investigation of whether there was a significant 'gender gap' in law school." Kris Knaplund, Kit Winter & Richard Sander, *1995 National Survey of Law Student Performance* *1 (1995), www.seaphe.org/databases/Nat95/Introduction.pdf. The NSLSP includes approximately 4,000 first-year law students from twenty-one schools (78% response rate). Data were collected on gender, race, socioeconomic status, UGPA, LSAT, and first-semester LGPA. UGPA, LSAT, and LGPA were standardized within schools, thereby representing each student's deviation from the school average for each variable (on the same scale across variables). *Id.* at *1 (1995) (noting "the data we report here is in all cases standardized by school").

Scholars have examined gender differences in law school performance with the BPS data, as well as with school-specific data from Harvard, the University of

have revealed that entering credentials neither meaningfully explain law school performance nor racial differences in law school performance. Specifically, (1) entering credentials (i.e., UGPA and LSAT) explain less than a quarter of the variation in LGPA,[155] and/or (2) Black and Latinx students receive lower law school grades than white students—first semester, first year, and cumulatively—even after accounting for entering credentials and law school difficulty/selectivity.[156] These two robust empirical findings undermine S&S's assertions that credential gaps are primarily responsible for learning mismatch and that racial differences in legal learning, in turn, result from credential gaps.

Racial disparities in LGPA that persist even after accounting for UGPA, LSAT, and school difficulty/selectivity merit further discussion in light of S&S's interpretation of the relevant evidence. S&S report that "the [NSLSP] established the following . . . controlling for credentials[:] Blacks received roughly the same grades as whites at the same school with similar credentials."[157] This statement misrepresents the data. In *Systemic Analysis*, Sander analyzes the NSLSP data and reports that there are no racial differences in first-semester law grades after controlling for UGPA and LSAT.[158] However, this result was only possible because Sander elected to code all students with missing data on race as "white." Meanwhile, he acknowledged that race does exert an

Pennsylvania (Penn), Stanford, and the University of Texas at Austin (UT-Austin); these scholars discovered that female students perform worse than their male counterparts with identical entering credentials. Thaxton, *supra* note 1, at 980–81 nn.498 & 499 (citing 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 (1996) (BPS schools); Lani Guinier, Michelle Fine & Jane Balin, Becoming Gentlemen: Women, Law School, and Institutional Change (2d ed. 1997) (Penn); 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) (Harvard); Harvard Law Sch. Working Grp. on Student Experiences, Study on Women's Experiences at Harvard Law School (2004) (Harvard); Daniel E. Ho & Mark G. Kelman, *Does Class Size Affect the Gender Gap? A Natural Experiment in Law*, 43 J. Legal Stud. 291 (2014) (Stanford); Allison L. Bowers, *Women at The University of Texas School of Law: A Call for Action*, 9 Tex. J. Women & L. 117 (2000) (UT-Austin)).

155 Thaxton, *supra* note 1, at 1006–1007; *see also* Andrea A. Curcio et al., *Measuring Law Student Success from Admissions through Bar Passage: More Data the Bench, Bar and Academy Need to Know* (Georgia State University College of Law Research Report No. 2890, 2019) (describing research revealing that LSAT is a weaker predictor of grades in experiential learning and writing courses than in doctrinal courses); William D. Henderson, *The LSAT, Law School Exams, and Meritocracy: The Surprising and Undertheorized Role of Test-Taking Speed*, 82 Tex. L. Rev. 975, 1043–44 (2004) (noting that the LSAT is correlated with 1L in-class exam scores, but not scores on take-home exams and papers in the first-year curriculum).

156 Thaxton, *supra* note 1, at 900 n.240, 977–78.

157 Sander & Steinbuch, *supra* note 15, at 723.

158 Sander, *Systemic Analysis*, *supra* note 147, at 429 ("When we control for the LSAT and UGPA variables, none of the 'race' variables (or the gender variable) is even close to being statistically significant (all the *p*-values are well above .05). This means that when we control for academic credentials, blacks, whites, Hispanics, and Asians all get pretty much the same grades.").

independent effect on first year law grades "if one does not include those not reporting race with white students."[159]

It has been widely known since the 1970s that Sander's approach to handling missing data typically leads to biased estimates.[160] Ian Ayres and Richard Brooks reexamined the NSLSP data using several sensible methods to account for missing information and discovered that racial differences in grades persist even after accounting for UGPA and LSAT.[161] S&S claim that Ayres and Brooks' analysis shows that Black students "performed slightly worse" than similarly situated white students,[162] but this is misleading because the effect size for race in Ayres and Brooks' analysis is nearly six times larger than the effect size reported by Sander.[163] Richard Lempert and colleagues'

159 Sander, *Systemic Analysis*, *supra* note 147, at 429 n.175. Perhaps it is telling that Sander includes this important qualification in a very lengthy footnote that could easily be overlooked by the reader, and this is not the only time this has happened when Sander's results provide little or no support for his claims. As I have explained:

> 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; 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 . . . . [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 . . . . 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 [because it is a categorical variable], but it is also the most policy relevant metric. 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.

Thaxton, *supra* note 1, at 998–99; *see also supra* note 65.

160 *See generally* Roderick J.A. Little & Donald B. Rubin, Statistical Analysis with Missing Data 47–50, 67–69 (3d ed. 2019) (explaining that casewise deletion (i.e., omitting cases with missing data on race) and mean substitution (i.e., recoding all students with missing data on race as "white") are known to produce biased estimates).

161 Ayres & Brooks, *supra* note 65, at 1828. Sander's problematic handling of missing data on race was also highlighted in his analysis of data from the University of Michigan Law School, and when missing race data were coded properly, the results directly contradicted Sander's conclusion regarding Black students' bar passage rates. *See* Thaxton, *supra* note 1, at 992–93 (discussing Richard Lempert, *University of Michigan Bar Passage 2004–2006: A Failure to Replicate Professor Sander's Results, with Implications for Affirmative Action* (July 23, 2012)).

162 Sander & Steinbuch, *supra* note 15, at 723 n.17.

163 Ayres & Brooks, *supra* note 65, at 1830 tbl.3 (noting that Sander's estimate of the effect of being Black on LGPA, net of UGPA and LSAT, is -0.007, whereas the models using widely

independent analysis of the NSLSP data corroborates Ayres and Brooks' findings about the direction, magnitude, and statistical significance of the impact of race on LGPA.[164]

In *Systemic Analysis*, Sander does not conduct an analysis of the impact of race on law school grades for LSAC's BPS data, explaining that entering credentials are not "standardized" for each school.[165] In his view, this lack of standardization makes regression results "meaningless at best and highly misleading at worse," without further explanation as to how or why.[166] S&S repeat this critique of the BPS data: "[A]lthough LSAC did standardize law school grades within each school (which also made schools more anonymous), it did not standardize LSAT scores and UGPAs within schools, so that one could not compare the entering credentials of students against their peers."[167] Four years before the publication of *Systemic Analysis*, LSAC examined the impact of race on LGPA, accounting for students' UGPA and LSAT.[168] The author of the LSAC study, Linda Wightman, found that nonwhite students received lower grades than white students with the same credentials when examining both individual law schools and groupings of law schools (i.e., clusters) in the BPS data.[169] In other words, the relationship between race and law grades remained irrespective of whether entering credentials were

endorsed approaches to missing data reveal that the effect of being Black is -0.041).

164 Lempert et al., *supra* note 38, at 17–18 tbl.1 (reporting the effect size for "Black" as statistically significant and 4.3 to 5.3 times larger than the effect reported by Sander when missing race data are handled properly).

165 *See supra* note 154 (noting that the NSLSP data were standardized for each of the twenty-one schools and finding a statistically significant race effect on LGPA after adjusting for UGPA and LSAT).

166 Sander, *Systemic Analysis*, *supra* note 147, at 428 n.172. Rothstein and Yoon analyzed the BPS data and discovered that Black students had lower LGPAs [class rank] than white students with identical Index scores and that the size of the racial gap in LGPA was similar across Index percentiles. Jesse M. Rothstein & Albert H. Yoon, *Affirmative Action in Law School Admissions: What Do Racial Preferences Do?* U. Chi. L. Rev. 649, 693–95 (2008) (disaggregating the Index distribution into percentiles and finding a similar Black-white gap in law school grades).

167 Sander & Steinbuch, *supra* note 15, at 721.

168 Linda F. Wightman, Beyond FYA: Analysis of the Utility of LSAT Scores and UGPA for Predicting Academic Success in Law School (2000); *see also* Timothy T. Clydesdale, *A Forked River Runs Through Law School: Toward Understanding Race, Gender, Age, and Related Gaps in Law School Performance and Bar Passage*, 29 Law & Soc. Inquiry 711, 736 (2004) (analyzing the BPS data and reporting that all racial minority students report lower LGPAs, relative to white students, even when adjusting for entering credentials).

169 Wightman, *supra* note 168, at 7 ("All analyses were conducted separately within each law school. However, in most instances, data presented in this report are summarized by law school cluster."). A similar dynamic was observed when examining the relationship between race and UGPA—i.e., the magnitude of the impact of race on UGPA, independent of relevant third variables, was similar whether schools were examined individually or collectively. *See infra* note 181.

standardized within schools. Sander does reference other studies of the BPS data conducted by Wightman on behalf of LSAC,[170] but he does not mention the aforementioned LSAC study, although the findings are directly relevant to his concern about the utility of the clustered BPS data for understanding the relationship between race and LGPA.[171]

In his reply to critics of *Systemic Analysis*, Sander identifies Lisa Anthony and Mei Liu's analysis of the impact of race on LGPA using school-specific data from three cohorts (1996–1998) collected by LSAC.[172] Sander acknowledges that Anthony and Liu's analysis contradicts mismatch theory—i.e., they reveal that race influences first-year law grades independent of UGPA and LSAT, with the largest effect being observed for Black students. Yet Sander (1) focuses his attention on the modest effect sizes for race, (2) claims that race may account for a "small portion of the Black-white differential in graduation rates and bar passage," and (3) faults his critics for not directly "deal[ing] with the implications of law grades."[173]

Sander's characterization of Anthony and Liu's findings reveals that he does not adequately appreciate the relevance of a consistent finding of a race effect on LGPA for at least three reasons. First, as noted earlier, undergraduate grades and entrance exam scores (UGPA and LSAT) explain only a modest proportion of variation in law school grades (LGPA).[174] In the NSLSP data, UGPA and LSAT explain 14% of the variance in LGPA (while gender and race explain about 6% of the variability), meaning that 86% of the variability

170 Sander, *Systemic Analysis*, *supra* note 147, at 471 (citing 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 Admissions Decisions*, 72 N.Y.U. L. Rev. 1 (1997) and Linda F. Wightman, *The Consequences of Race-Blindness: Revisiting Prediction Models with Current Law School Data*, 53 J. Legal Educ. 229 (2003)).

171 *See also* Robert L. Linn, *Ability Testing: Individual Differences and Differential Prediction*, *in* Ability Testing: Uses, Consequences, and Controversies Part II 335–88 (Alexandra K. Wigdor & Wendell R. Garner eds., 1982) (describing results from a school-specific analysis of forty-two law schools, which revealed that Black students had lower grades than white students with identical UGPA and LSAT in thirty-eight of the schools).

172 Sander, *Reply to Critics*, *supra* note 61, at 1968 n.15 (citing Lisa C. Anthony & Mei Liu, Analysis of Differential Prediction of Law School Performance by Racial/Ethnic Subgroups Based on the 1996–1998 Entering Law School Classes (2003) (analyzing data from 167 law schools that had ten or more students from at least one minority group)).

173 Sander, *Reply to Critics*, *supra* note 61, at 1968–69.

174 *See supra* note 155. Sander reports analyzing the relationship between entering credentials and LGPA for nine cohorts of law students at UCLA and lists R^2 values of 0.35 for first-semester grades and 0.39 for second-semester grades. Sander, *Systemic Analysis*, *supra* note 46, at 421 n.149. Sander characterizes these model-fit statistics as demonstrating strong predictive power, *id.*, but empirical scholars' views on an acceptable cutoff for the R^2 statistic varies considerably. According to one scholar, "A very high R^2 (say about 0.9) is essential if our predictions are to be accurate." Michael Lewis-Beck, Applied Regression: An Introduction 24 (1980). Another researcher noted some statisticians "will dismiss a model as 'weak' or 'unreliable' and 'lacking predictive power' if the reported R-squared of the model is below 0.6." Peterson K. Ozili, *The Acceptable R-Square in Empirical Modelling for Social Science Research*, *in* Social Research Methodology and Publishing Results 134 (Candauda A. Saliya ed., 2023).

in LGPA is unexplained by entering credentials.[175] Another analysis of four law schools (two private and two public) discovered that UGPA and LSAT explain about 20% of the variation in LGPA,[176] thereby leaving 80% of the variability in LGPA unexplained by entering credentials. Second, UGPA, LSAT, and LGPA only explain a modest proportion of the variation in graduation and bar passage rates.[177] David Chambers and colleagues note, for example, that Sander's models including UGPA, LSAT, law school tier, LGPA, and race explain 26% of the variation in graduation rates and 33% of the variation in bar passage.[178] Thus the majority of variation in law school graduation and bar exam performance is left unaccounted for by entering credentials, law school grades, and mismatch in Sander's models. Third, modest racial differences in LGPA (that are not explained by mismatch) may meaningfully contribute to an overall race effect on bar exam performance via an indirect race effect that is transmitted through LGPA.[179] Consequently, as critics of Sander's analyses have emphasized, when entering credentials explain only a very modest amount of variation in LGPA, mismatch is incapable of explaining the gap in

175 Lempert et al., *supra* note 38, at 17–18. Sander's "first choice" analysis, which offers an alternative analysis of the impact of mismatch on LGPA, reports that mismatch explains 9%–15% of the variation in first-year LGPA and 14%–18% in overall LGPA. Sander, *Replication of Mismatch Research, supra* note 24, at 79 tbls.3 & 4. *See also* Thaxton, *supra* note 1, at 977 (finding that race is strongly predictive of grades in the BPS, increasing explained variation by 136% over models including only UGPA and LSAT).

176 Stephen P. Klein & Roger Bolus, *The Size and Source of Differences in Bar Exam Passing Rates Among Racial and Ethnic Groups*, Bar Examiner 13–15 (Nov. 1997) (reporting that UGPA and LSAT explain about 4% and 16%, respectively, of the variation in LGPA); *see also* Lisa C. Anthony et al., Predictive Validity of the LSAT: A National Summary of the 1995–1996 Correlation Studies 6 tbl.2 (1999) (noting that UGPA and LSAT explain 25% of the variance in first-year law school grades).

177 Thaxton, *supra* note 1, at 1006; *see also* Part III.C for a discussion of the low explanatory power of S&S models for predicting bar passage.

178 *See* David L. Chambers et al., *The Real Impact of Eliminating Affirmative Action in American Law Schools: An Empirical Critique of Richard Sander's Study*, 57 Stan. L. Rev. 1855, 1870 n.51 & 1881 n.99 (2005); *see also* Klein & Bolus, *supra* note 176, at 13 (reporting LGPA explains about 50% of the variation in bar exam scores). Sander's first choice analysis reveals that mismatch explains 1%–11% of the variance in first-time and eventual bar passage. Sander, *Replication of Mismatch Research, supra* note 24, at 81–82, tbls.5 & 7.

179 One common measure of the indirect effect of race is the multiplicative product of the direct effect of race on LGPA and the direct effect of LPGA on the likelihood of graduation/bar passage. The stronger the relationship between LGPA and graduation/bar passage, the stronger the total race effect, because part of the influence of race on graduation/bar passage is transmitted through LGPA for reasons unrelated to mismatch. *See infra* notes 182–186 (discussing the overall effect of race on bar passage via direct and indirect effects); *accord* Thaxton, *supra* note 1, at 1013; *see also* Clydesdale, *supra* note 168, at 747. Not only may race impact graduation/bar performance indirectly through LGPA (mediation), but race may also condition the impact of LGPA on graduation/bar performance (moderation)—i.e., a conditional indirect effect. *See supra* note 143 & *infra* note 525 and accompanying text.

LGPA or bar exam performance in any given school, within or between racial groups.[180]

Of greater importance, however, is the fact that analyses of the school-specific NSLSP data, the BPS data (both aggregated and school-specific), and LSAC's three-cohort school-specific data all provide evidence that race impacts law school grades independent of mismatch. Given the volume of data supporting that finding, it is difficult to understand how any reasonable reading of these studies would lead S&S to conclude that mismatch accounts for racial differences in LGPA.[181]

180 Lempert et al., *supra* note 38, at 4, 17 & 23–25 ("Sander hardly has a theory. What he has is a series of empirical results and an explanation for them, but since his empirical analyses are poorly executed, no theory can be grounded in their explanation."); *accord* Ho, *supra* note 28, at 2014 ("The marginal effect of going to a higher-tier law school in this model is roughly a 2.4 [percentage point] increase in the probability of first-time bar passage . . . . This effect is neither statistically distinguishable from zero nor close to explaining the substantial Black/white bar passage gaps as Sander originally claimed.").

181 Sander acknowledges that research on undergraduate students reveals that racial minorities receive lower grades than white students with the same entering credentials, but he doubts the validity of these studies by claiming that they suffer from methodological shortcomings. Sander, *Systemic Analysis*, *supra* note 147, at 429 n.174.

Sander claims that this prior research does not include important variables about students' pre- and post-enrollment characteristics/experiences; he also mentions the fact that students belonging to different racial groups who are similarly situated with respect to high school class rank and SAT scores may differ in important ways related to mismatch theory. It is worth noting that the specific study that Sander cites by William Bowen and Derek Bok, *infra*, acknowledges those potential problems with their data, but that study also includes a measure of socioeconomic status, which is likely to serve as a proxy for some of those unmeasured variables. Moreover, Bowen and Bok report UGPA underperformance for *all* racial minorities (i.e., Black, Latinx, Asian, and "Other Race"); consequently, for Sander's critique to have merit, these mismatch-related unmeasured variables must impact *all* racial minorities, including those groups that Sander argues do not generally benefit from race-conscious affirmative action (e.g., Asians). Perhaps of greatest importance is that Sander's own analysis does not account for any of these alleged unobserved factors, yet he claims that "the collectively poor performance of black students at elite schools does not seem to be due to their being 'black' (or any other individual characteristic, like weaker educational background, that might be correlated with race). For him, the poor performance seems to be simply a function of disparate entering credentials," Sander, *Systemic Analysis*, *supra* note 147, at 429. He adds that, "since [his] data does not show any net underperformance by blacks, [he] will not belabor the potential measurement problems that sometimes show up in other data sets," *id*. at 429 n.174. It appears that Sander does not have a problem with using limited measures of academic preparedness when the results seem to support his claim, but he is dismissive of others' findings using limited measures when their results contradict his claims. As noted earlier, Sander's analysis of the relationship between race and grades is flawed because of the incorrect way he codes students who do not report their race. *See supra* note 159. When this problem is corrected, the results show that racial minorities underperform their white counterparts with the same entering credentials. *See supra* note 161. As noted below, studies of racial differences in UGPA conducted after Sander published *Systemic Analysis* that included a wider range of pre- and post-enrollment variables discovered that Black and Latinx students receive lower grades than similarly situated white students. *See, e.g.*, Douglas S. Massey et al., The Source of the River: The Social Origins of Freshmen at America's Selective Colleges and Universities 188 tbl.9.1 (2010); Charles

I explain in *How Not to Lie About Affirmative Action* that analyses that "control" for LGPA and report no statistically significant racial differences in first-time bar passage rates are flawed because race exerts an effect on LGPA independent of entering credentials and law school difficulty.[182] This error of inference when assessing the potential impact of race is a classic example of post-treatment bias, because controlling for LGPA suppresses the portion of the influence that race

---

ET AL., *supra* note 45, at 63 tbl.2.17.

Researchers have also discovered that race exerts an impact on LSAT, independent of applicants' age, undergraduate institution, college major, UGPA, and grade inflation. The gap is largest for Black (9.2) and Latinx (6.8) students. This suggests that race may indirectly impact bar passage through *both* entering credentials and LGPA, for reasons other than mismatch. Thaxton, *supra* note 1, at 929 n.299 (citing Kidder, *supra* note 109, at 1074–79). Consequently, the racial gap in entering credentials may not accurately capture differences in aptitude for legal learning. Moreover, by accounting for LSAT—and, by extension, degree of mismatch—S&S's analysis does not measure the total race effect. And as I explain below, the same may be true when controlling for UGPA.

Other studies have identified a similar relationship between race, high school GPA (HSGPA), SAT score, and UGPA. *See* Jesse M. Rothstein, *College Performance Predictions and the SAT*, 121 J. ECONOMETRICS 297, 313 (2004) (examining data from over 22,000 students enrolled at seven University of California campuses and finding evidence that (1) race impacts first-year college GPA, independent of SAT, HSGPA, and undergraduate major, and (2) the SAT explains less than 3% of the variance in first-year UGPA when removing the effect of high school demographic characteristics); WILLIAM G. BOWEN ET AL., CROSSING THE FINISH LINE: COMPLETING COLLEGE AT AMERICA'S PUBLIC UNIVERSITIES 79, 126 (2009) (explaining that "race/ethnicity is quite a strong predictor of SAT scores," and that after adjusting for SAT/ACT scores, HSGPA, family income, state residency, and university attended, "Black, Hispanic, and Asian men all earn [college] grades 5–6 percentile points lower than do comparable white men. Among women the white-Black gap in adjusted rank-in-class is still larger, at about 9 points, whereas the white-Hispanic gap is smaller, at about 4 points."); WILLIAM G. BOWEN & DEREK CURTIS BOK, THE SHAPE OF THE RIVER: LONG-TERM CONSEQUENCES OF CONSIDERING RACE IN COLLEGE AND UNIVERSITY ADMISSIONS 77, 383 (1998) (examining over 30,000 students from twenty-eight selective colleges and finding that "[B]lack students . . . equivalent to all students in their SAT scores, high school grades, socioeconomic status, and other characteristics included in the model (gender, selectivity of school attended, field of study, being athlete or not) still would have had a class rank that was, on average, roughly 16 percentile points lower than the class rank of apparently comparable classmates," and further finding that "[t]hese patterns hold whether considering the schools collectively or separately."); MASSEY ET AL., *supra* note 174, at 188 (analyzing data from over 3,400 college students and finding that Black and Latinx college students earn lower first-semester UGPA than white students, independent of academic preparation, socioeconomic status, and susceptibility to peer influence); CHARLES ET AL., *supra* note 45, at 63 (finding that Black and Latinx students receive lower grades than white students in their first two years of college after accounting for academic preparation, social preparation, college major, difficulty of courses, academic assistance, academic aspirations, family background, parental education, and socioeconomic status).

182 Thaxton, *supra* note 1, at 1013. The total race effect is a combination of the direct effect of race on bar passage and the indirect effect of race on bar passage operating through law school grades. *See supra* note 179 and accompanying text. Analysts can capture the total race effect by excluding all variables that are impacted by race (i.e., post-treatment variables), but this approach often ignores the racial dynamics that researchers wish to investigate. Sen & Wasow, *supra* note 129.

has on bar exam performance that operates through LGPA (i.e., mediation).[183] Richard Lempert and colleagues also raised this concern shortly after Sander published *Systemic Analysis*: "Others have identified a direct effect of race on first-semester law school grades, suggesting that index scores are not the whole story."[184] One can eliminate (or substantially mitigate) the aforementioned post-treatment bias and estimate an overall effect of race on bar passage by examining both the direct effect of race on bar performance (RACE → BAR) and the indirect effect of race on bar performance that operates through LGPA (RACE → LGPA × LGPA → BAR).[185] My reanalysis of the BPS data using this analytical framework discovered that there is a statistically significant total race effect on bar passage, such that racial minorities are less likely to pass the bar exam on the first attempt, even after controlling for UGPA, LSAT, undergraduate major, LGPA and law school difficulty/selectivity.[186] Consequently, the true impact of race on bar exam performance is improperly suppressed in any analysis that adjusts for LGPA, which is a post-treatment variable, without also accounting for the portion of the race effect on bar exam performance that is transmitted through LGPA.

Sander has used a similar analytical framework when examining the direct and indirect effects of tier location (i.e., school difficulty) on the likelihood of passing the bar exam when analyzing a sample of Black students, but he does not extend this analysis to the examination of racial differences.[187] Sander's

183 *See also* Clydesdale, *supra* note 168, at 747 (analyzing BPS data and discovering that racial gaps in the probability of passing the bar exam "are eliminated, however, once final law school GPA is controlled. This loss of statistical significance indicates that while race may have a direct effect on bar passage initially, its final effect is indirect—through law school GPA primarily").

184 Lempert et al., *supra* note 38, at 3.

185 Thaxton, *supra* note 1, at 1013; *see* Figure 2; *see also supra* note 129 and accompanying text.

186 Thaxton, *supra* note 1, at 1013–16. In *How Not to Lie About Affirmative Action*, I both describe and implement an analytical framework to "examine[] the overall effect of race/ethnicity with Sander's analy[tical] framework by calculating the direct effect of race/ethnicity on first-time bar passage and the indirect effect of race/ethnicity, operating through LGPA." *Id*. My findings strongly 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." *Id*. These results directly contradict "[Sander's] core conclusions about the (un)importance of race/ethnicity in explaining racial differences in students' performance in law school and on the bar exam." *Id*.

187 Thaxton, *supra* note 1, at 1015 (remarking that "it is unfortunate that Sander does not examine whether racial/ethnic differences exist with respect to an overall race/ethnicity effect using his mediation [i.e., direct and indirect effects] framework" because his framework "naturally lends itself to investigate other types of overall effects when the model posits both direct and indirect effects operating through an intervening variable"). The estimate of an overall effect of law school tier (i.e., law school difficulty) on bar passage is obtained by measuring

inclusion of LGPA in his models examining the BPS data underscores his understanding of the importance of analyzing LGPA when testing mismatch theory's utility in explaining both inter- and intraracial differences in bar exam performance. For example, he includes UGPA, LSAT, and LGPA as explanatory variables in his models predicting bar passage in *Systemic Analysis* to demonstrate that mismatch explains the observed racial disparities.[188] And in his response to Daniel Ho's critique of *Systemic Analysis*,[189] Sander employs two different analytical approaches that also include controls for UGPA, LSAT, and LGPA. One of those analyses is performed to show that mismatch can account for racial differences in bar passage.[190] The other is a mediation model applied only to Black students and expressly models LGPA as a function of UGPA, LSAT, and tier location to demonstrate the total effects of UGPA, LSAT, and tier location on bar passage and to disentangle those total tier effects into their direct and indirect effects (mediated by LGPA).[191] Sander's mediation model analysis discovered that the tier location has a positive direct effect on bar exam performance, and a negative indirect on bar exam performance that operated through LGPA. The magnitude of the negative indirect effect was larger than the positive direct effect, resulting in a negative total effect for tier location. Had Sander not examined the total effect of tier location, and only examined the direct effect, he would have reached the opposite conclusion about the impact of tier location on bar performance.[192] If Sander believes the relationship between the unstandardized UGPA/LSAT and standardized LGPA is meaningless when analyzing BPS data, then his own models would be invalid as well. Put differently, it seems strange for him to see the importance of including analyses of LGPA when the results purportedly support his claims about mismatch theory, yet see the analysis of LGPA as irrelevant when the results contradict his claims about mismatch theory.

---

both the direct effect of law school tier on bar performance (TIER → BAR) and the indirect effect of law school tier on bar performance that operates through LGPA (TIER → LGPA × LGPA → BAR). Sander, *Response to Ho*, *supra* note 123, at 2008; *see also infra* note 191 and accompanying text.

188 Sander, *Systemic Analysis*, *supra* note 147, at 444 tbl.6.1.

189 Sander, *Response to Ho*, *supra* note 123.

190 Richard Sander & Joseph Doherty, *Supplemental Notes on the Relative Effect of Law School Grades and Law School Prestige Upon Bar Passage* 4 (Apr. 27, 2005) (unpublished manuscript) (on file with the author); Thaxton, *supra* note 1, at 860.

191 Sander, *Response to Ho*, *supra* note 123, at 2008 fig.1; *see also* Thaxton, *supra* note 1, at 1008.

192 Sander's analysis of the impact of tier location on bar exam performance suffers from several significant methodological flaws. When those flaws are remedied, tier location is shown to have no statistically significant impact on bar exam performance. Thaxton, *supra* note 1, at 997–1008. Nonetheless, the general framework he employs to disentangle direct and indirect effects of tier location on bar exam performance is consistent with testing mismatch theory's proposed causal pathways.

It is also worth noting that the explanatory power of the models predicting LGPA with UGPA and LSAT increases when analyses are disaggregated by tier location compared with aggregated data for all tiers when analyzing the BPS data. This fact suggests that the lack of a school-specific standardization of UGPA and LSAT may be less of a concern when conducting tier-specific analyses of LGPA because students' distribution of entering credentials, relative to the middle student, is roughly similar across schools within tiers.[193] Sander's own analysis of a "cascade effect"[194] is informative on this issue when he shows that the Black-white gap in entering credentials, as well as the variability in entering credentials for all students, is similar for all law school tiers except for the tier consisting solely of seven historically Black colleges and universities (HBCUs).[195]

In their school-specific analysis, S&S report that, after adjusting for students' entering credentials and degree of mismatch, there are no statistically significant effects of being Black and Latinx on the likelihood of bar passage.[196] S&S conclude, "The mismatch effect in our models accounts for the large disparities in bar passage across racial lines. If our findings can be generalized, they largely account for the very serious disparity between the racial makeup of first-year law students and the practicing bar."[197] But these conclusions are not supported by their models because they fail to properly examine the influence of race and credential differences on learning mismatch, which the authors claim is the key mechanism through which mismatch impacts bar passage.[198]

193 *See also* Chambers et al., *supra* note 178, at 1878 n.85 (noting that prior research conducted by Clydesdale, *see supra* note 168, "sought to deal with the standardization problem by controlling for law school tier. This control should help because law schools tend to be homogenous within tiers (and different across tiers) on admissions credentials. Indeed, credential homogeneity was a factor Wightman [*see supra* note 17] used to sort schools into tiers"). *See supra* note 18 ("The BPS grouped law schools into six 'clusters' based on seven school characteristics: size, cost, selectivity, faculty-student ratio, percentage minority, median LSAT score, and median undergraduate GPA.").

194 A cascade effect occurs when middle- and low-tier schools duplicate the race-conscious admissions practices of the elite-tier schools. Sander, *Systemic Analysis*, *supra* note 147, at 372.

195 Sander, *Systemic Analysis*, *supra* note 147, at 416 tbl.3.2 ("It is hard to conclude from this data that the racial gap, or affirmative action, disappears at lower-tier schools. Except for the seven law schools that have historically served minorities—obviously a special case—the Black-white gap is nearly constant."). Sander reports the standard deviation for Index for white students only, but my analysis of the BPS data reveals that the standard deviation for Black students is similar across all of the non-HBCU tiers as well. *See also* Chambers et al., *supra* note 178, at 1878 n.85 ("Sander himself describes the standard deviation among whites and among African Americans in first-tier schools as 'strikingly small.' They are similarly small at most of the other tiers. Still, we cannot be confident how well the tier control does its job. Anthony and Liu's study [*see supra* note 172] does not have this problem, however, and its consistency with Clydesdale's findings [*see supra* note 168] is good reason to accept the latter's conclusions on this issue.").

196 Sander & Steinbuch, *supra* note 15, at 740.

197 *Id.*

198 According to mismatch theory, learning mismatch (a first-order effect) leads to second-

S&S's school-specific data avoid the alleged "lack of standardization" problem with UGPA and LSAT in the BPS data,[199] so it is puzzling that they do not report any results examining the effect of race on LGPA after accounting for entering credentials.[200]

In explaining their failure to analyze this key component of the data, S&S blur the distinction between causal *description* and causal *explanation*.[201] S&S argue that they do not need to control for law school grades because they can directly measure mismatch.[202] In making this argument, they misrepresent the very explanatory framework that mismatch posits because mismatch allegedly impacts "learning" in law school.[203] Whereas causal description "describes the

order effects, such as lower graduation rates, failing the bar examination, and lower grades for students experiencing mismatch. Sander, *Stylized Critique*, *supra* note 8, at 1643–45; *accord* Richard Sander & Aaron Danielson, *Thinking Hard About "Race-Neutral" Admissions*, 47 U. Mich. J.L. Reform 967, 984–88 (2014).

199 *But see supra* notes 193 & 195 and accompanying text (noting that the lack of standardization of UGPA and LSAT in the BPS data is unlikely to bias results when one accounts for tier location).

200 *See supra* note 47. Sander & Steinbuch, *supra* note 15, at 725 tbl.1 & 743–44 (reporting that their data have information on UGPA for some cohorts for UCLA and all cohorts from Davis & UALR).

Three cohorts were excluded from UCLA because transfers students could not be distinguished, but S&S do not indicate whether the excluded cohorts were the only cohorts that had data on UGPA or LGPA. Sander & Steinbuch, *supra* note 15, at 743. S&S did not exclude transfers in the analysis of bar-takers from UALR, claiming that the school "had very few such transfers, and there is no reason to think those transfers that did exist came from generally more elite or less elite schools." *Id.* at 725. While it is understandable why S&S would want to exclude incoming transfer students from some of their model specifications because their credential gap may have been different during their first year of law school at another institution, it is important to note that, according to UCLA's available annual public disclosures from 2011 to 2022, the number of *incoming* transfer students ranged from ten to forty-three (the average number of incoming transfers was thirty-two)—approximately 10% of the graduating class. Am. Bar Found., *509 Required Disclosures*, www.abarequireddisclosures.org/disclosure509.aspx. In other words, most of the bar-takers from UCLA (roughly 90%) were nontransfers. The value of an examination of the influence of UGPA, LSAT, and LGPA outweighs any modest bias from including a very small percentage of transfer students, especially as it pertains to mismatch theory's hypothesized mechanism—learning mismatch—that S&S claim to explain racial differences in bar passage (and differences in bar passage, more generally). It is worth noting that UALR has a part-time program, whereas UCLA and Davis do not. From 2011 to 2022, part-time students comprised between 25% and 30% of enrolled students at UALR. *Id.* S&S do not distinguish between full-time and part-time students in their analyses, although full- versus part-time status may impact learning for reasons other than mismatch.

201 *See generally* William R. Shadish et al., Experimental and Quasi-Experimental Designs for Generalized Causal Inference 9 (2002).

202 Sander & Steinbuch, *supra* note 15, at 737 ("But with the individual-level data in this paper, we do not need law school grades—we can directly measure and control for mismatch.").

203 *Compare* Sander, *Reply to Critics*, *supra* note 61, at 1966 (claiming that "those receiving racial preferences will struggle academically, receive low grades, and actually learn less in some important sense than they would have at another school where their credentials were closer

consequences attributable to . . . a treatment,"[204] causal explanation "clarif[ies] the mechanisms through which and the conditions under which that causal relationship holds."[205] Statistician Michael Sobel notes that "in many cases, researchers simply equate such a description with 'causal explanation' without even attempting to say what the terms 'causation' and 'explanation' mean."[206] Geographer Patrick Meyfroidt states that, "Causal analysis is too often plagued by the failure to recognize that establishing causality relies on two dimensions: causal effects and causal mechanisms . . . identifying causes is generally only an intermediate goal . . . understanding the causal mechanism is crucial . . . because the same cause can produce its effect through different mechanisms, requiring different interventions."[207] Similarly, sociologists Peter Hedström and Petri Ylikoski emphasize that "simply interpreting one's research findings in mechanism terms will not suffice [because] explanations should reflect the causal processes actually responsible for the observations."[208] S&S's use of a school-specific measure of credential difference does not alter this fact.[209]

The theory that animates the choice of a statistical model is critical to the interpretation of a measurement of a causal effect: "If a theory suggests why and how D brings about Y, then merely providing evidence of the amount that Y can be expected to change in response to an intervention on D does not then provide any support for the underlying theory."[210] For example, analysts

---

to the school median. The low grades will lower their graduation rates, bar passage rates, and prospects on the job market") *with* Rothstein & Yoon, *supra* note 166, at 659–60 (emphasizing that Sander makes a "claim about the causal effect of attending a highly selective school on a student's outcomes" when a student is admitted to a highly selective school where she is outmatched, according to her entering credentials, because of affirmative action, "she works hard but . . . by the end of the first year, she finds herself near the bottom of her class . . . [which] may demoralize her and lead her to doubt her own abilities . . . . she may miss key concepts in her struggle to keep up with classes and ultimately have trouble passing the bar . . . her transcript, while from a more prestigious school [i.e., one that is higher ranked], will not show a record of strong performance, potentially hurting her employment prospects.") (quote edited for clarity).

204 Shadish et al., *supra* note 201, at 9.

205 *Id.*

206 Michael E. Sobel, *Causal Inference in the Social and Behavioral Sciences*, *in* Handbook of Statistical Modeling for the Social and Behavioral Sciences 3 (Gerhard Arminger et al. eds., 1995).

207 Patrick Meyfroidt, *Approaches and Terminology for Causal Analysis in Land Systems Science*, 11 J. Land Use Sci. 501, 503 (2016).

208 Peter Hedström & Petri Ylikoski, *Causal Mechanisms in the Social Sciences*, 36 Ann. Rev. Soc. 49, 64 (2010).

209 S&S also undermine the granularity of their mismatch measure by top-coding and discretizing, thereby diminishing the utility of their analysis for causal description. *See* Part III.D.

210 Stephen L. Morgan & Christopher Winship, Counterfactuals and Causal Inference: Methods and Principles for Social Research 330 (2d ed. 2015); *accord* Eric Neumayer & Thomas Plümper, Robustness Tests for Quantitative Research 13 (2018) ("The causality of an 'identified' association is not understood unless the mechanism is understood . . . and

have consistently discovered that individuals with more money tend to report higher levels of happiness.[211] However, these same analysts disagree over the specific reasons why this relationship exists.[212] Some hypothesize that money increases happiness through fulfilling basic needs. Others claim that money contributes to happiness through conspicuous consumption and materialism. Still others claim that money improves happiness through social connections and interpersonal relationships. Simply demonstrating an association between money and happiness cannot reveal which theory is correct. It is only by examining the various hypothesized pathways through which money may lead to happiness are analyst able to determine which explanation (or explanations) are supported by the data, as well as their relative importance.[213]

Theories not only tell readers what they should expect to observe but also what they should *not* expect to observe.[214] Psychologist Andrew Hayes explains: "We know that we better understand some phenomenon when we can answer not only whether *X* affects *Y*, but also how *X* exerts its effect on *Y*, and when *X* affects *Y* and when it does not."[215] For example, scholars have suggested that money increases happiness up to point, then the relationship "flattens" and increased money has no effect on happiness.[216] Moreover, the relationship between money and happiness depends on the level on income inequality where a person lives: the relationship will be most pronounced in countries with greater income inequality.[217]

A key assumption of S&S's analysis is that race does not meaningfully impact learning mismatch, independent of credential distance. On their assumption, if race exerts an impact on LGPA after controlling for entering credentials and degree of mismatch, then Black and Latinx students are not learning less solely because of mismatch. It must be that something correlated with race, other than mismatch, is impacting Black and Latinx students' learning in law school.[218] Thus, a credible and scientifically defensible examination of

---

the identification of a causal effect and an unbiased estimate of its strength differs from understanding the causal mechanism—the chain of events [i.e., theory] that ultimately brings about the effect.").

211 Nicholas Buttrick & Shigehiro Oishi, *Money and Happiness: A Consideration of History and Psychological Mechanisms*, 120 Proc. Nat'l Acad. Sci. e2301893120 *1 (2023).

212 *Id*.

213 *Id*.

214 Shadish et al., *supra* note 201, at 25 (noting that causal explanation entails ruling out irrelevancies); *accord* Jasso, *supra* note 76, at 38.

215 Andrew F. Hayes, Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach 6 (2d ed. 2018).

216 Buttrick & Oishi, *supra* note 211.

217 *Id*.

218 Thaxton, *supra* note 1, at 1016; Rothstein & Yoon, *supra* note 166, at 670 n.81 (noting that the Black students have lower UGPAs and LGPAs than predicted by their entering

the cumulative impact of race must be measured both directly on bar passage and indirectly on bar passage, operating through LGPA (after controlling for mismatch-related variables).[219] This analytical approach to measuring racial discrimination has been endorsed by the National Research Council (NRC): "[M]easures 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."[220] Analogizing the pathway to becoming a lawyer to rafting down a forked river, one scholar further noted:

> For minority law students, the journey is never easy. Passing through the entry gate is often challenging, and the BPS analyses demonstrate that *all* minority law students must ride the white-water rapids side. And these truly are two different rivers, *as the analyses above hold constant "piloting" differences* [i.e., individual student quality differences]. The white-water rapids that *all* minority law students must navigate are not at all like the smooth currents what white law students enjoy.[221]

Consequently, even if S&S opt to omit LGPA from their models predicting bar passage because, in their view, the mismatch variable is an adequate proxy for grades,[222] the analysis of the relationship between race, entering credentials, and LGPA is important to understanding both (a) the amount of variation in

---

credentials, and that this finding strongly suggests that "other factors—greater difficulty in social adjustment, greater participation in part-time employment or work-study, outright discrimination—may serve to depress Black students' performance"); Ayres & Brooks, *supra* note 65, at 979 ("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.").

219 Thaxton, *supra* note 1, at 863 ("A more successful general approach [to mediation analysis] is to collect and use covariates that are predictive of both the intermediate potential outcomes [here, LGPA] and the final potential outcomes [here, bar passage].") (quoting 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)).

220 Nat'l Research Council, Measuring Racial Discrimination 226 (Rebecca M. Blank et al. eds., 2004); Thaxton, *supra* note 1, at 1013.

221 Clydesdale, *supra* note 168, at 752 (emphasis in original). "Piloting differences" include UGPA, LSAT, law school difficulty, hours of planned study, hours of planned employment, marital and children status, socioeconomic status, financial obligations, self-confidence, learning and physical impairments, prelaw preparation, number of family members who attended law school, and English proficiency. *Id*. at 752 n.34 & 748-49 tbls. 5A & 5B.

222 It is also important to emphasize that even if excluding law grades from the analysis is a defensible modeling choice—which it is not—S&S's robustness analysis focusing on the two California schools produces results that contradict their claim of mismatch accounting for racial differences in bar passage. *See* Part III.F .

LGPA explained by entering credentials and (b) whether race is predictive of LGPA, independent of entering credentials and law school difficulty.[223]

S&S are careful to note that "mismatch is not about race," but they claim that "the single most disturbing manifestation of mismatch" is the sizable racial gap in the likelihood of passing the bar exam because Black and Latinx students receive large admissions preferences that cause them to learn less than they would have in a less demanding and less competitive environment.[224] Nonetheless, S&S's analysis is, inarguably, constructed around testing competing hypotheses of racial disparities in bar exam performance. They expressly note that there are "three distinct theories about what might be happening [to cause] low minority bar passage rates . . . . The 'discrimination' hypotheses . . . the 'credential' hypothesis . . . [and] the mismatch hypothesis."[225] They further claim that their data allow them "to directly estimate mismatch effects and compare them with the absolute effects of credentials, the effect of race, and variations in school effectiveness in preparing students for the bar;"[226] and they claim that their "regression analysis provides a more rigorous test of mismatch . . . [because] 'controlling' for race in a regression would separate out the 'race' effect from the 'mismatch' effect."[227] But S&S's statistical analyses are incapable of doing what they advertise because properly separating these effects requires examining the causal pathways through which these effects influence bar passage. Political scientists Maya Sen and Omar Wasow emphasize that "race affects nearly every socioeconomic variable typically included in standard regression analyses, *including ones meant to detect mediating patterns*."[228] They add that the "existing practice of interpreting the residual impact of race is at best poorly conceptualized and at worst introduces serious bias."[229] S&S's omission of LGPA from their analysis not only undermines their ability to separate race effects from mismatch effects, but it negatively impacts their ability to make credible causal inferences about mismatch effects within racial groups.[230] "Thinking about mechanisms provides a descriptively adequate way of talking about science and scientific discovery [and] mechanisms are sought to explain

223 *See also supra* note 47 (noting that S&S do not examine the relationship between positive mismatch and LGPA, although such a test might be highly informative about the relationship between credential distance and the key mechanism identified by mismatch theory—i.e., learning mismatch).

224 Sander & Steinbuch, *supra* note 15, at 739.

225 *Id.* at 720.

226 *Id.* at 740.

227 *Id.* at 728–29.

228 Sen & Wasow, *supra* note 129, at 504–05 (emphasis added).

229 *Id.* at 505.

230 Lempert et al., *supra* note 38, at 17 & 23–25 (noting that mismatch is incapable of explaining gaps in LGPA or bar exam performance in any given school, within or between racial groups, because entering credentials explain only a modest fraction of the variability in LGPA).

how a phenomenon comes about or how some significant process works."[231] In fact, "it is widely recognized that a consistent estimate of a . . . causal effect of D on Y may not qualify as a sufficiently deep causal account of how D affects Y, based on the standards that prevail in a particular field of study."[232]

In order to adequately distinguish between alternative explanations of racial differences in bar exam performance, S&S must justify that their models support a causal interpretation (called "identification") by eliminating any distortion of the relationship between mismatch and bar passage because of omitted variables (i.e., confounders) and alternative causal pathways.[233] Otherwise, such distortion can lead to overestimation or underestimation of the true effects of the causal variable(s).[234] James Granato and colleagues strongly caution that the "reliance on statistically significant results means virtually nothing when a researcher makes very little attempt to identify the precise origin of the parameter in question. Absent this identification effort, it is not evident where the model is wrong. The dialogue between theory (model) and test is weakened."[235] Similarly, economist Charles Manski argues that "the study of [causal] identification [in empirical economics] logically comes first. Negative identification findings imply that statistical inference [i.e., significance tests and p values] is fruitless."[236]

In principle, three widely employed identification strategies for causal effects could shore up S&S's analysis if properly implemented: (1) exclusion criterion (identification by instrument), (2) front-door criterion (identification by mechanism), and (3) back-door criterion (identification by conditioning).[237] The exclusion criterion requires using a proxy variable that is associated with bar passage solely through its association with mismatch to estimate the causal impact of mismatch on bar passage.[238] The front-door criterion estimates the

231 Peter Machamer et al., *Thinking about Mechanisms*, 67 Philosophy of Science 1, 3 & 23 (2000).

232 Morgan & Winship, *supra* note 210, at 325.

233 *Id.* at 329.

234 *See* Ron Johnston et al., *Confounding and Collinearity in Regression Analysis: A Cautionary Tale and an Alternative Procedure, Illustrated by Studies of British Voting Behaviour*, 52 Quality & Quantity 1957, 1958 (2018) (describing the various ways confounding variables impact the relationship between the target and outcome variables).

235 Jim Granato et al., Empirical Implications of Theoretical Models in Political Science 14 (2021); *accord* James J. Heckman & Jeffrey A. Smith, *Assessing the Case for Social Experiments*, 9 J. Econ. Perspectives 85, 108 (1995) ("The end result of a research program based on [black box] experiments is just a list of programs that 'work' and 'don't work,' but no understanding of why they succeed or fail.").

236 Charles F. Manski, Identification Problems in the Social Sciences 6 (1995).

237 Morgan & Winship, *supra* note 210, at 26.

238 Identification by instrument (also called instrumental variable estimation) is commonly used to obtain an unbiased measurement of the impact of a variable on an outcome in the presence of confounding variables. Both Williams, *supra* note 67, and Rothstein & Yoon, *supra* note 166, employ an instrumental variables framework when assessing the impact of

impact of mismatch on bar passage through adjusting for mediating variables (e.g., LGPA).[239] The back-door criterion adjusts ("conditions") for all relevant omitted variables that are associated with mismatch and bar passage but are not a consequence of mismatch.[240] To be sure, each of these identification strategies has unique assumptions and limitations,[241] and depending on the particular problem, one strategy may be more plausible than others.[242]

Based on S&S's data and analytical approach, the front-door criterion is the most defensible identification strategy for their study.[243] Political scientists Adam Glynn and Kevin Quinn note that the front-door criterion rests on "causal assumptions that . . . might, in some situations, be more plausible than the standard [back-door criterion] assumption,"[244] and the front-door criterion "is most beneficial in situations where it is implausible to assume that treatment assignment is conditionally ignorable [i.e., identification by conditioning] . . . [and can] improve inference for total causal effects."[245] In his seminal text on causal modeling, sociologist Kenneth Bollen explains that the "underspecified equation [i.e., omitted variables] is less desirable than an equation that includes the intervening variables [i.e., front-door criterion] in

race on bar passage and other law-related outcomes; although as discussed *supra* notes 68–71, Williams's analysis has been deemed implausible. S&S do not adopt an instrumental variable framework, although it is the most common approach for causal inference in empirical economics. Joshua D. Angrist & Jörn-Steffen Pischke, Mostly Harmless Econometrics: An Empiricist's Companion 114 (2008) (noting that the most powerful weapon in economists' arsenal of statistical tools used to estimate causal effects is the method of instrumental variables); *see also* Thaxton, *supra* note 1, at 870–71 ("Most empirical scholarship in economics offers an extended discussion of the use of (or search for) valid instrumental variables, especially when key explanatory variables are endogenous such as LGPA, but Sander's work is silent on this issue.").

239 Judea Pearl, Causality: Models, Reasoning, and Inference 79 (2d ed. 2009) (defining the "front-door criterion" as the isolation of causal effects through adjusting for mediating variables); Thaxton, *supra* note 1, at 855 n.63 (discussing the assumptions of the front-door criterion).

240 Pearl, *supra* note 239, at 79 (defining the "back-door criterion" as controlling for pretreatment confounders that allow for spurious associations); Thaxton, *supra* note 1, at 853–55 (discussing the assumptions of the back-door criterion). The back-door criterion is also called conditional exogeneity/ignorability. *Id.* at 855.

241 Thaxton, *supra* note 1, at 853–56 (discussing necessary assumptions for causal claims).

242 *See also* Part III.F (noting that "the number and variety of possible robustness tests is large, and the relevance of these tests will depend on the specificities of model uncertainty, the intended inferences, and the data structure").

243 *See generally* Marc F. Bellemare et al., *The Paper of How: Estimating Treatment Effects Using the Front-Door Criterion* (Sept. 12, 2020) (unpublished manuscript, on file with the author) (describing assumptions regarding the front door criterion approach to identification).

244 Adam N. Glynn & Kevin M. Quinn, *Why Process Matters for Causal Inference*, 19 Political Analysis 273, 285 (2011).

245 *Id.*, *supra* note 244, at 273–74.

that the understanding of the process is far more complete when the mediating factors are explicitly incorporated."[246]

Rather than adopt the front-door criterion identification strategy, S&S's implicit identification strategy is the back-door criterion. However, their data are so limited that satisfying either the exclusion or back-door criteria is highly implausible.[247] S&S's most complex models only include race, LSAT/Index, credential distance, and school-level fixed effects.[248] Yet the mismatch effects that Sander reported using the BPS data disappear when analysts considered a richer set of confounding variables.[249] A credible implementation of the back-door criterion requires S&S to consider a wider range of confounding variables in order to support their claim about mismatch effects. Economists Mark Rosenzweig and Kenneth Wolpin explain that even under the best quasi-experimental designs, such as natural experiments, where assignment to the treatment and control groups can be considered as good as randomized (i.e., back-door criterion), strong assumptions about the mechanism through which the treatment influences the outcome are still needed to justify an analyst's causal interpretation of the estimates.[250] So even assuming, *arguendo*, that S&S were able to satisfy the back-door criterion, their analysis would, at best, capture

246 Bollen, *supra* note 121, at 47.

247 I have discussed the implausibility of Sander's causal identification assumptions when analyzing the BPS data, demonstrated Sander's violations of those assumptions, and advocated for identification through modeling the association of the disturbance terms (i.e., residuals) as an alternative to identification by instrument or conditioning when examining the impact of mismatch on bar exam performance. Thaxton, *supra* note 1, at 869–878. With respect to the final point, identification may be achieved via copula-based methods to causal inference—which adjust for unmeasured confounding (i.e., omitted variables) without identification by instrument or conditioning—by assuming that the effects of unmeasured confounding variables are manifested in the joint distribution of the residuals. Falkenström et al., *supra* note 91, at 303–305. Copula methods have been popular in the fields of business and economics for quite some time, and are increasingly being utilized in other social sciences, *see, e.g.*, *id.* (psychology); Florian M. Hollenbach et al., *Multiple Imputation Using Gaussian Copulas*, 50 Soc. Methods & Res. 1259 (2018) (political science); Mike Vuolo, *Copula Models for Sociology: Measures of Dependence and Probabilities for Joint Distributions*, 46 Soc. Methods & Res. 604 (2017) (sociology).

248 *See also supra* note 118–128 & *infra* notes 322–323 and accompanying text.

249 *See also* Thaxton, *supra* note 1, at 843 n.25 & 889 n.196 (noting that Sander considers fewer than ten variables when analyzing the impact of mismatch on bar passage with the BPS data, but takes into account more than two dozen variables in a study of nearly 4,000 attorneys across that nation who passed the bar in 2000); *id.* at 888 nn.191, 193–94 (citing three studies that consider a wider range of confounding variables and report no mismatch effects with the BPS data).

250 Mark R. Rosenzweig & Kenneth I. Wolpin, *Natural "Natural Experiments" in Economics*, 38 J. Econ. Lit. 827 (2000); *accord* Heckman & Smith, *supra* note 235, at 95 & 108 (noting that randomized social experiments "provide little evidence on many questions of interest" and that "there is a sizeable divergence between the theoretical capabilities of evaluations based on random assignment and the practical results of such evaluations"); Leamer, *supra* note 92, at 45 ("Let's not add a 'randomization' button to our intellectual keyboards, to be pushed without hard reflection and thought.").

the total effect of credential difference on bar passage—a black-box approach to causality (see Figure 4)[251]—and not the effect of credential difference on bar exam outcomes resulting from learning mismatch (i.e., operating through LGPA).[252] This is highly relevant because Sander has explicitly stated that, in the law school context, learning mismatch causes students to fail the bar exam: "I [Sander] have written elsewhere about the importance of distinguishing different types of mismatch effects, but in the law school context these distinctions do not matter so much . . . . 'Systemic Analysis' contends that students tend to learn less when their credentials are much lower than those of their classmates, and this hurts their ability to pass the bar."[253] Therefore, in addition to concerns about causal identification, unpacking the relationship between race, credential distance, learning mismatch, and the likelihood of bar passage is central to understanding the complex dynamics that S&S claim they want to illuminate.[254]

It may be illuminating to consider S&S's decision to ignore mismatch theory's key causal mechanism, learning mismatch, when making causal claims (the "black box" approach) because the most popular texts addressing causal inference emphasize the importance of expressly studying mechanisms for causal analysis. This emphasis spans the gamut of behavioral and social sciences—e.g., anthropology,[255] economics,[256] epidemiology,[257] political science,[258] psychology,[259] sociology,[260] and statistics.[261] Political scientists Kosuke Imai, Luke Keele, Dustin Tingley, and Teppei Yamamoto caution that

251 Figure 4 accounts for students' absolute and relative (i.e., school-specific) differences in entering credentials by including law school UGPA and law school LSAT.

252 Morgan & Winship, *supra* note 210, at 114 n.11 (explaining that the goal of the back-door criterion is to estimate the total effect of the cause on the outcome and that analysts interested in estimating direct and indirect effects of the cause on the outcome must adjust for the mediating variable); *accord* Glynn & Quinn, *supra* note 244, at 273 (describing the black-box approach to causality, the so-called Neyman-Rubin model, as ignoring causal mechanisms and focusing exclusively on random assignment of the treatment and control conditions).

253 Sander, *Replication of Mismatch Research, supra* note 24, at 76.

254 *See generally* Neumayer & Plümper, *supra* note 210, at 11 ("Causal inference is much more than the mere identification of a cause-effect relationship which dominates the current debate in the social sciences. Social scientists do not only wish to establish the existence of a causal effect […] they aim at understanding the mechanisms by which causes take effect.").

255 McElreath, *supra* note 33.

256 Scott Cunningham, Causal Inference: The Mixtape (2021).

257 Tyler J. Vanderweele, Explanation in Causal Inference: Methods for Mediation and Interaction (2015).

258 Rethinking Social Inquiry (Henry E. Brady & David Collier eds., 2d ed. 2010).

259 Hayes, *supra* note 215.

260 Morgan & Winship, *supra* note 210.

261 Paul R. Rosenbaum, Observation and Experiment: An Introduction to Causal Inference (2017).

"the 'black box' approach to causality has been criticized across disciplines for being atheoretical and even unscientific . . . [and] for many researchers, estimating causal effects is insufficient and underlying mechanisms must be examined in order to test social science theories empirically."[262] Economists James Heckman and Jeffrey Smith caution that "simple black-box evaluations pose a serious threat to the accumulation of knowledge about the behavior of persons and institutions because they are not conducted within a behaviorally coherent framework of analysis."[263] Sociologists Stephen Morgan and Christopher Winship remark, "Social scientists have recognized for decades that the best explanations for how causes bring about their effects must specify in empirically verifiable ways the causal pathways between causes and outcomes."[264] Psychologists William Shadish, Thomas Cook, and Donald Campbell define causal explanation as "develop[ing] and test[ing] explanatory theories about the pattern of effects, causes, and the mediational processes that are essential to the transfer of a causal relationship."[265] With respect to mismatch theory, sociologist Timothy Clydesdale explains, "If one wants to understand race differences in bar passage, one must understand race differences in law school grades . . . . By contrast, analyses of law school grades show that understanding race differences in grades requires understanding far more than race differences in LSATs and undergraduate GPAs."[266] S&S's black-box approach stymies their efforts to prove their hypothesis about mismatch.

Even when S&S's overall analytical framework for separating race effects from mismatch effects is accepted on its own terms, there are other methodological problems that call into question their results—and several of these problems were previously identified in my critique of Sander's earlier work on mismatch. In the interest of space, my discussion of these issues will not be exhaustive.

B. Coefficient Comparisons

S&S use logistic regression to estimate the impact of various independent variables on the likelihood of passing the bar exam. At the core of their analysis is a comparison of the logistic regression coefficients (odds ratio) for the race

262 Kosuke Imai et al., *Unpacking the Black Box of Causality: Learning about Causal Mechanisms from Experimental and Observational Studies*, 105 Am. Pol. Sci. Rev. 765 (2011).

263 Heckman & Smith, *supra* note 235, at 108; *accord* Field Trials Toolbox, *supra* note 144 (emphasizing that understanding an intervention's impact on knowledge or attitudes may be a crucial initial step in determining whether the intervention is acting in the hypothesized manner when measuring behavioral changes); Hayes, *supra* note 215, at 7 ("The 'how' question [of causal association] relates to the underlying psychological, cognitive, or biological process that causally links X to Y.").

264 Morgan & Winship, *supra* note 210, at 325; *accord* Thaxton, *supra* note 1, at 863 ("Mediation analysis has gained considerable promise in the social sciences, and when properly conducted, assists in making credible causal claims.").

265 Shadish et al., *supra* note 201, at 25.

266 Clydesdale, *supra* note 168, at 747.

variables across "nested models" (i.e., models that overlap with respect to some but not all variables). The difference between coefficients reflects the degree to which the impact of race is mediated or confounded by mismatch. "Model comparisons of this general kind are ubiquitous in social research, as well as other areas, and they are especially prominent whenever regression techniques are applied for the analysis of nonexperimental data, such as survey data"[267] because "different theoretical explanations or 'causal structures' are represented in this fashion . . . [and] it is important to examine the *stability* of the coefficients associated with X when various modifications are made in the explanation or causal structure."[268]

S&S conduct an "eyeball" comparison of the coefficients for race, noting that the absolute value of the coefficient either decreases or becomes statistically insignificant in models that include mismatch-related variables.[269] However, it is well established that "comparisons of this general kind are invalid."[270] Eyeballing differences in coefficients has been criticized in the empirical literature,[271] so analysts are advised to take a more rigorous approach by computing standard errors, significance tests, and confidence intervals for the difference in coefficients.[272] Statisticians Andrew Gelman and Hal Stern state this point succinctly: "Statistical comparisons based on the statistical significance of one parameter and the non-significance of the other parameter is improper because these changes (significance versus non-significance) are often not themselves statistically significant."[273]

267 Clifford C. Clogg et al., *Statistical Methods for Comparing Regression Coefficients between Models*, 100 Am. J. Soc. 1261, 1263 (1995).

268 *Id.*, at 1288 (emphasis in original).

269 Sander & Steinbuch, *supra* note 15, at 734 ("As our models have gradually become complex (moving from Models 1 to Model 6), the independent effect of 'race' has become steadily less important.").

270 Clogg et al., *supra* note 267, at 1263 (explaining that invalid coefficient comparisons of the same variable across models take the form of whether (a) the coefficients are significantly different from zero in either model; (b) the level of significance of those coefficients differs between the models; or (c) the apparent size of those coefficients differs between the two models); *accord* Makin & Orban de Xivry, *supra* note 27, at 4 (noting an erroneous inference occurs when comparisons between two effects are made without a direct statistical test).

271 Clogg et al., *supra* note 267, at 1263.

272 *Id.*; *see also* Jerome Weesie, *Seemingly Unrelated Estimation and the Cluster Adjusted Sandwich Estimator*, 52 Stata Tech. Bull. 34, 35 (1999) (describing the implementation of the appropriate test for comparisons of coefficients with the *Stata* software that S&S report using for their analysis). The test statistic for the coefficient comparison is: $z = (B_r - B_f) \div [(V_{Br}) + (V_{Bf}) - (2 \times cov(B_r, B_f)]^{1/2}$; where B_r is the coefficient for the reduced model (i.e., the model excluding potential confounders/mediators), B_f is the coefficient for the full model (i.e., the model including potential confounders/mediators), V_{Br} and V_{Bf} are the variances of the coefficients, and $cov(B_r, B_f)$ is the covariance of the coefficients. These quantities are used to calculate the standard error, confidence interval, and significance test for the difference in coefficients. Clogg et al., *supra* note 267, at 1279.

273 Andrew Gelman & Hal Stern, *The Difference Between "Significant" and "Not Significant" is Not Itself*

The problem with eyeballing differences between coefficients across models is especially pronounced when using logistic regression, because "unlike linear models, the change in the coefficient of the variable of interest cannot be straightforwardly attributed to the inclusion of the confounding [or mediating] variable."[274] This is attributable to the fact that logistic regression coefficients are "are not identified separately from the residual variance . . . [so] when we compare the coefficients of different models fitted to the same data, it is inevitable that the models will differ in their residual [variance]."[275] In other words, coefficients for the same variable across nested logistic regression models are not measured on the same scale. Properly identifying confounding or mediating effects requires coefficients to be recalibrated before conducting comparisons.[276]

Alternatively, S&S could have calculated the marginal effect,[277] which both is easier to interpret than the logistic regression coefficient (odds ratio) and does not require recalibration to be compared across models.[278] In either case–

Statistically Significant, 60 Am. Stat. 328, 329 (2006). The focus of the comparison need not be between statistically significant and statistically insignificant coefficients, and the same logic applies to comparing two coefficients that are both statistically significant when the analyst is interested in whether any difference between the coefficients is substantively meaningful. Clogg et al., *supra* note 267, at 1263.

274 Kristian Bernt Karlson et al., *Comparing Regression Coefficients Between Same-Sample Nested Models Using Logit and Probit: A New Method*, 42 Sociol. Method. 286 (2012).

275 Richard Breen et al., *Interpreting and Understanding Logits, Probits, and Other Nonlinear Probability Models*, 44 Ann. Rev. of Sociol. 39, 41–42 (2018) (noting that the problems with comparing logit and probit coefficients across models have been identified for more than thirty years); *accord* J. Scott Long, Regression Models for Categorical and Limited Dependent Variables 70 (1997) ("Since the variance of [the latent dependent variable] changes when new variables are added to the model, the magnitudes of all [regression coefficients] will change even if the added variable is uncorrelated with the original variables. This makes it misleading to compare coefficients from different specifications of the independent variables."); Jun Yan et al., *Comparing Regression Coefficients Between Nested Linear Models for Clustered Data with Generalized Estimating Equations*, 38 J. Edu. & Behav. Stat. 172, 173 (2013) (noting that "the omitted variables in the reduced model lead to bigger variation . . . [and] the coefficient estimates of the existing covariates are attenuated towards zero").

276 Breen et al, *supra* note 275, at 48–49.

277 *See supra* note 65.

278 Trenton D. Mize et al., *A General Framework for Comparing Predictions and Marginal Effects across Models*, 49 Sociol. Method. 152, 162–63 (2019) ("[R]egression coefficients are not appropriate for comparing effects across logit and probit models because a change in the size of the coefficient across models can reflect both confounding and rescaling of the model. Marginal effects do not have this limitation and are thus appropriate for comparing effects across logit and probit models."); *see also* Alan Agresti & Claudia Tarantola, *Simple Ways to Interpret Effects in Modeling Ordinal Categorical Data*, 72 Statistica Neerlandica 210, 213 (2018) (noting that odds ratios tend to be unstable when including explanatory variables that are uncorrelated with the key variable of interest, but that average marginal effects do not experience this problem); *accord* Long, *supra* note 275, at 73–74 (explaining that marginal effects can be used to assess the relative magnitudes of two variables when holding the levels of all other explanatory variables constant).

(recalibrated) odds ratio or marginal effect—simply eyeballing differences should be avoided, and the proper statistical test of the difference in the race coefficients should be used.[279] The same criticisms apply to S&S's comparison of effect sizes of the mismatch variable when including school-specific fixed effects.[280]

An additional complication pertaining to S&S's assessment of the statistical significance of the effect of race on bar passage is their use of a two-sided test rather than a one-sided test.[281] A two-sided test, also called a two-tailed test or a nondirectional test, is appropriate when assessing whether a particular parameter is meaningfully different from a specified value—i.e., either larger (positive) or smaller (negative).[282] In contrast, a one-sided test, also known as a one-tailed test or directional test, is appropriate when the analyst is interested in determining whether a particular parameter is (strictly) larger or (strictly) smaller than a specified value.[283] Incorrectly employing a two-sided test may

279 Mize et al., *supra* note 278, at 165; *accord* Gelman & Stern, *supra* note 270, at 328 (explaining that eyeballing differences in the magnitudes and/or significance levels of coefficients of nested models is an error of interpretation). S&S do not indicate whether they account for the nonindependence of students within schools when calculating the standard errors of the coefficients in their models. S&S's inclusion of the school-specific fixed effects in their models captures unobserved heterogeneity between schools influencing the likelihood of passing the bar, but these fixed effects do not account for the nonindependence of students within those schools. Thaxton, *supra* note 1, at 1001 n.580. Robust-clustered standard errors should be calculated—either analytically or via resampling—to account for the nonindependence of students within schools. *Id.* at 1002 n.584. 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 school, and the direction of the bias will often depend on whether the within-cluster variance is larger/smaller than the between-cluster variance. *Id.* at 1001 n.579. A similar problem was identified in Sander's analysis of the BPS data, and when appropriate standard errors were used, one of his core findings supporting mismatch theory was contradicted. *Id.* at 1001 (noting that the effect of law school tier is statistically insignificant when accounting for the nonindependence of students within schools); *see also* A. Colin Cameron & Douglas L. Miller, *A Practitioner's Guide to Cluster-Robust Inference*, 50 J. Human Resources 317, 329 (2015) (noting that the inclusion of fixed effects generally does not control for all of the intracluster correlation of errors, so cluster-robust standard errors should be reported).

280 Sander & Steinbuch, *supra* note 15, at 734 ("Adding school controls makes the mismatch coefficients consistently lower in Model 6—i.e., the mismatch effect is more severe—while slightly weakening the LSAT (credential) effect.").

281 *Id.* at 731 tbl.3, 735 tbl.4 & 742 tbl.A.

282 Stephen B. Hulley et al., Designing Clinical Research 45 (4th ed. 2013) ("A two-sided alternative hypothesis states only that there is an association; it does not specify the direction. For example, 'Drinking tap water is associated with a different risk of peptic ulcer disease—either increased or decreased—than drinking bottled water'.").

283 Hulley et al., *supra* note 282, at 46 ("A one-sided alternative hypothesis specifies the direction of the association between the predictor and outcome variables. The hypothesis that drinking tap water increases the risk of peptic ulcer disease (compared with bottled water) is a one-sided hypothesis.").

result in type II error (also called "*beta* error")—i.e., a false negative.[284] In the context of S&S's analysis of the effect of race, the null hypothesis being tested is that race does not exert an independent impact on the likelihood of passing the bar, and a failure to reject the null hypothesis after accounting for students' degree of mismatch would provide support for mismatch theory.[285] S&S's null hypothesis should be examined with a one-sided test because the specific inquiry is whether the parameter of race—namely, being Black and Latinx—exerts a negative impact on bar performance, relative to being white, independent of mismatch.[286] Clearly, S&S specify a directional hypothesis, so a nondirectional test (i.e., a two-sided test) is unduly conservative and might lead to a failure to reject the null hypothesis of an absence of a true race effect.[287] Put simply, the one-sided test has more statistical power to reject the null hypothesis than the two-sided test. Further, the one-sided test is appropriate when, as here: (1) effects can exist in only one direction or (2) effects can exist in both directions, but the analyst is concerned about the effect in only one direction.[288] Because S&S's analytical framework for examining the effect of race would satisfy either of these conditions, a one-sided test is most appropriate (and not the two-sided test that they use). As Hulley and colleagues have remarked, "A

284 *Id.* ("A type I error (false-positive) occurs if an investigator rejects a null hypothesis that is actually true in the population; a type II error (false-negative) occurs if the investigator fails to reject a null hypothesis that is actually false in the population.").

285 *Id.* at 49 ("The null hypothesis has only one function: to act like a straw man. It is assumed to be true so that it can be rejected as false with a statistical test . . . . The key insight is to recognize that if the null hypothesis is true, and there really is no difference in the population, then the only way that the study could have found one in the sample is by chance.").

286 *Compare id.* at 45 ("A one-sided hypothesis may also be appropriate when there is very strong evidence from prior studies that an association is unlikely to occur in one of the two directions.") *with* Sander & Steinbuch, *supra* note 15, at 716 & 720 ("(Blacks) tended to have very low law school grades—an effect he found was entirely due to preferences—and were far more likely than other students to fail the bar after multiple attempts—another 'preference' effect rather than a 'race' effect . . . . It was well recognized by the 1980s that 'minority' students—particularly Black students—had bar passage rates well below those of whites.").

287 Technically speaking, "When a two-sided statistical test is used, the *p* value includes the probabilities of committing a type I error in each of the two directions, which is about twice as great as the probability in either direction alone . . . . A one-sided *p* value of 0.05, for example, is usually the same as a two-sided *p* value of 0.10." Hulley et al., *supra* note 282, at 49. Consequently, a one-sided hypothesis has the statistical advantage of having more power to detect statistical significance with a smaller sample size than a two-sided test. Herbert Hoijtink, Informative Hypotheses: Theory and Practice for Behavioral and Social Scientists 80 (2011) ("Compared to a noninformative [i.e., nondirectional] alternative, with an informative [i.e., directional] alternative, smaller samples are needed to obtain the same power.").

288 *See* Michael Borenstein, *The Shift from Significance Testing to Effect Size Estimation*, *in* 3 Comprehensive Clinical Psychology: Research and Methods 319 (Nina R. Schooler ed., 1998) ("A one-tailed test is a test that will be interpreted only if the effect meets the criterion for significance and falls in the expected direction [and] . . . is appropriate only if an effect in the unexpected direction would be functionally equivalent to no effect.").

good hypothesis must be based on a good research question. It should be simple, specific, and stated in advance."[289] S&S's justification of their use of a nondirectional test is especially important in the presence of collinearity because, as explained below,[290] the power to detect a significant race effect is also reduced by collinearity.[291] At minimum, S&S should have reported results for both one-sided and two-sided tests.[292]

To be clear, the question of changes in regression relationships across models (i.e., coefficient comparisons) is different from the question of incremental improvement in prediction across models (i.e., model fit), and methods suited for the former question need not be valid for the consideration of the latter question.[293] The issue of model fit is discussed next.

C. Model Fit: Overall and Incremental

Goodness of fit, also known as model fit, refers to how well the observed data correspond to the expected (i.e., fitted) values from a proposed model.[294]

289 Hulley et al., *supra* note 282, at 44.

290 *See* Part III.D.

291 Gareth James et al., An Introduction to Statistical Learning 101 (2d ed. 2021) ("In the presence of collinearity, we may fail to reject [the null hypothesis]. This means that the *power* of the hypothesis test—the probability of correctly detecting a *non-zero* coefficient—is reduced by collinearity.") (emphasis in original).

292 In addition to reporting both the one- and two-sided tests of statistical significance, which are based on the Wald test, it would have been advisable for S&S to examine the statistical significance of the parameter estimates (and construct confidence intervals) for the race variables using a likelihood ratio (LR) test. Although both the Wald and LR tests are asymptotically equivalent, the LR test is preferred because it does not assume normality of the parameter variance-covariance matrix and has better small sample properties than the Wald test. The LR-based confidence intervals are particularly useful in nonlinear models, such as the logistic regression. Patrick Royston, *Profile Likelihood for Estimation and Confidence Intervals*, 7 Stata J. 376 (2007).

293 Clogg et al., *supra* note 267, at 1262–63 ("Model comparisons generally involve two related but different questions. The first is whether the increment in prediction or explained variability obtained by adding Z is significant or important. . . . The second question is equally important when regression techniques are used to compare explanations. Are the coefficients that describe the relationship between Y and X in the first model different from the coefficients that describe the relationship between Y and X in the second model where Z has been added and has hence been 'statistically controlled'?").

294 Goodness of fit can be separated into two distinct but related tasks: *model selection* and *model diagnostics*. Long, *supra* note 275, at 85. Model selection, which is usually based on a scalar statistic (a single number), is most useful for comparing competing models to select a final model. *Id.* at 102. Models can be evaluated based on their accuracy in predicting new data or current data. Model diagnostics (sometimes referred to as model assessment) are concerned with evaluating the validity of the assumptions of the chosen model. The inquiry is based primarily on an examination of the observation-specific discrepancies between observed and predicted values in the form of model residuals. *Id.* at 98. Model diagnostics identify outliers (i.e., observations that are inconsistent with model predictions), permit the assessment of leverage (i.e., observations with unusual explanatory variable values) and of influence (i.e., the effect of an observation on the parameter estimates of the model), and assist in detecting

Although there are a variety of measures of model fit, some are poorly suited for certain types of analyses.[295] Sociologist Scott Long remarked, "I am unaware of convincing evidence that selecting a model that maximizes the value of a *given* measure of fit results in a model that is optimal in any [relevant] sense."[296] S&S report a single measure of model fit: a statistic called Somers' *D*. This statistic is a rank correlation between the predicted probability of an outcome and the observed binary outcome. It considers all pairs of observations that have different outcomes and estimates the probability that the predictions and outcomes are concordant—that is, the observations with the larger outcome values also have the higher set of model-fitted probabilities.[297]

It is well known, however, that Somers' *D* is a poor measure to use for imbalanced binary variables, such as bar exam success, because it inflates the fit (i.e., accuracy) of the model.[298] The bar passage rates for UCLA, Davis, and UALR, are respectively, 89%, 81%, and 78%, thus indicating that the binary

other systematic problems between the model and the data (e.g., incorrect functional form assumptions). *Id*. at 24 & 98–100.

My discussion of model fit in this section identifies problems with S&S's methodology pertaining to model selection. But even assuming, *arguendo*, that S&S's preferred model is the best among available alternatives, S&S do not offer a discussion of any diagnostic tests to inform the reader that their modeling assumptions should be trusted. For an accessible, although somewhat dated, discussion of goodness of fit for the type of models S&S examine (i.e., logistic regression), *see id*. at 85–113.

295 Long, *supra* note 275, at 102.

296 *Id.* (emphasis added).

297 *See* Ronald H. Somers, *A New Asymmetric Measure of Association for Ordinal Variables,* 27 Am. Soc. Rev. 799, 804 (1962). It should be emphasized that a model's explanatory capacity is distinct from a model's predictive capacity. "The type of uncertainty associated with explanation is of a different nature than that associated with prediction." Galit Shmueli, *To Explain or to Predict?*, 25 Stat. Sci. 289, 292 (2010). The goal of explanatory modeling is to minimize the discrepancy between the estimated and true model parameters (i.e., the focus is on the regression coefficients). In contrast, the goal of predictive modeling is to minimize the difference between the value of the expected outcome and the observed outcome (i.e., the focus is on the outcome variable). Although both approaches use the same regression modeling framework, "they differ in the questions they are asking from it, their focal point of success, and how they select the set of [independent] variables to address their respective concepts of bias." Cristobal Young, *The Difference Between Causal Analysis and Predictive Models: Response to "Comment on Young and Holsteen (2017),"* 48 Soc. Methods & Res. 431, 434 (2019). Consequently, higher predictive accuracy is not synonymous with greater explanatory validity. In fact, "a parsimonious but less true model can have a higher predictive validity than a truer but less parsimonious model." Shmueli, *supra* note 284, at 307. This seemingly counterintuitive result is attributable to the fact that "there are important reasons to exclude statistically significant [bad] control variables. . . . Indeed, intermediate outcomes will naturally be included in a predictive model due to their mechanical correlation with the main outcome but most clearly should *not* be included in [an explanatory] model." Young, *supra* note 284, at 437; *accord* Greiner, *supra note* 130, at 564.

298 Chambers et al., *supra* note 178, at 1871; *accord* Max Kuhn & Kjell Johnson, Applied Predictive Modeling 255 (2013) (explaining that the severely skewed class distribution present in imbalanced classification tasks may result in even more bias in the predicted probabilities, as they overfavor predicting the majority class).

variable is highly imbalanced.[299] Sander reported the Somers' *D* statistic as his sole measure of model fit in *Systemic Analysis*, which led David Chambers and colleagues to remark, "Somers'[] *D* presents a misleading portrait of how [Sander's] model does. It certainly should not have been the only regression diagnostic presented."[300]

The following example, adapted from Chambers et al.'s critique of Sander's use of Somers' *D*,[301] helps illustrate the problem. Suppose a model predicts that Student A has a 95% chance of passing the bar exam and Student B has a 90% chance of passing the bar exam, and that Student A passes the bar, while Student B fails the bar. Similarly, suppose this same model predicts that Student Y has a 10% chance of passing the bar exam and Student Z has a 5% chance of passing the bar exam, and Student Y passes the bar, while Student Z fails the bar. The model assigns a student who failed the bar exam a 90% predicted probability of passing (Student B), and assigns a student who passed the bar a 90% predicted probability of failing (Student Y). In both situations, the Somers' *D* measure will treat the pairs as properly classified because it is a rank-order statistic.[302] But it should be clear to any casual observer that the model does not perform well at correctly classifying the students.[303] Given the high initial probability that any given student will pass the bar exam, it is likely that both students in many of the concordant pairs will have estimated bar passage probabilities above 50%, thereby leading to a high Somers' *D* if the model distinguishes between those who have a greater and lesser chance of passing. Yet the same model does not accurately identify as failures most students who actually failed. The model is considered *poorly calibrated* because there is considerable disjunction between model outputs and observed data.

Somers' *D* is a measure of a model's discriminative ability (i.e., ability to differentiate between those with and without the outcome), but it is "insensitive to systematic errors in calibration" because it does not account for the lack of agreement between observed outcomes and predictions.[304] "If the purpose of the models is only to rank-order subjects, calibration is not an issue. Otherwise, a model having poor calibration can be dismissed outright. Given that the

299 *See, infra*, Table 2.

300 Chambers et al., *supra* note 178, at 1871.

301 *Id.* at 1871 n.57.

302 A rank-order statistic transform data with numerical or ordinal values to their rank when the data are sorted and is unconcerned with the original values of the data. Herbert A. David & Haikady N. Nagaraja, Order Statistics 3 (3d. ed. 2004).

303 Chambers et al., *supra* note 178, at 1871 n.57.

304 Ewout W. Steyerberg et al., *Assessing the Performance of Prediction Models: A Framework for Some Traditional and Novel Measures*, 21 Epidemiology 128, 132 (2010); *accord* Frank E. Harrell, Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis 258 (2015) (noting that if rank-ordering of subjects is the analyst's sole purpose, "rank-based indexes have the advantage of being insensitive to the prevalence of positive responses").

two models have similar calibration, discrimination should be examined critically."[305]

The problem with Sander's use of this measure was mentioned by several critics of mismatch, including me, and several alternatives have been suggested.[306] Amy Farley and colleagues analyzed in-state first-time bar exam performance of law school graduates and explained that "predictive models should seek to maximize accuracy of at-risk identification [i.e., percentage of correctly predicted failed attempts] and minimize false positives [i.e., percentage of incorrectly predicted failed attempts]."[307] Similarly, my prior criticism of Sander's use of Somers' *D* as a model fit statistic notes that "given the skewed nature of bar passage rates because nearly everyone eventually

305 Harrell, *supra* note 304, at 92 (2015); *accord* David W. Hosmer et al., Applied Logistic Regression 185–86 (3d ed. 2013) ("We always recommend the performing of a goodness of fit analysis. If the fitted model is to be used to discriminate between the two outcome groups, then and only then, do we recommend looking at . . . summary measures of discrimination . . . . If the study's objective is to estimate the Pr(y = 1) [i.e., fitted probabilities] then we need to assess calibration and discrimination.").

306 Thaxton, *supra* note 1, at 892 ("A more useful statistic to assess model fit in binary regression models is Tjur's *D*. The statistic simultaneously considers correctly predicted 'passes' and 'fails,' and is preferable when the binary variable either has a lot of 1's or 0's."); Chambers et al., *supra* note 178, at 1870 n.50 (noting that appropriate model diagnostic measures for Sander's models in *Systemic Analysis*, such as the Nagelkerke R^2, "would have alerted the knowledgeable reader to the likelihood that [Sander's results] leave[] much of what leads to bar passage unexplained"). *See generally* Klaas Sijtsma, Never Waste a Good Crisis: Lessons Learned from Data Fraud and Questionable Research Practices 44 (2023) ("There is little if any excuse for using obsolete, sub-optimal, or inadequate statistical methods or using a method irresponsibly, and inquiring minds find out that better options exist even when they lack the skills to apply them.").

307 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 Leg. Stud. 605, 618 (2019).

When describing the various approaches to assess the calibration of binary classification models, such as logistic regression, statistician Joseph Hilbe remarked, "Analysts are not only interested in correct prediction, but also in such issues as what percentage of time does the predictor incorrectly classify the outcome." Joseph M. Hilbe, Practical Guide to Logistic Regression 82 (2015). When the binary outcome is highly skewed, it is advisable to raise the threshold required to classify a predictive value—especially when there may be different costs associated with false positive and false negative misclassification. Commonly used techniques to assess the performance of a binary classification model at various thresholds are confusion matrices, receiver operator characteristic curves (ROC), and sensitivity-specificity (S-S) plots. *Id*. at 81–88. Each of these tests is based on a threshold that determines the optimal probability value to separate estimated probabilities to successes or failures. *Id*. at 82; Farley et al., *supra*, set their threshold at 75%. A prominent data scientist has explained, "The bottom line is that when studying problems with imbalanced data, using the classifiers produced by standard [statistical] algorithms without adjusting the output threshold may well be a critical mistake (depending on your research question)." Foster Provost, *Machine Learning from Imbalanced Data Sets 101*," *in* Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI-00) 3 (Rob Holte et al., eds., 2000).

passes, greater attention needs to be given to the model's ability to correctly predict who fails the bar exam."[308]

A related problem is that because Somers' *D* utilizes ranking information only, it cannot distinguish between models that yield the same orderings but provide different predictions.[309] At minimum, S&S should have presented several model diagnostic measures mentioned in prior critiques of Sander's work given the shortcomings of the Somers' *D* statistic.[310] Furthermore, when reporting rank-order statistics, like Somers' *D*, analysts are strongly advised to report a confidence interval (i.e., level of uncertainty) for the statistic.[311] So rather than presenting a single number, S&S should have presented the Somers' *D* for *each* model as a range of possible numbers (also see discussion of Somers' *D* in Part III.F). The calculation of confidence intervals is useful for

308 Thaxton, *supra* note 1, at 892.

309 Agresti & Tarantola, supra note 278, at 218 ("Because it [Somers' *D*] utilizes ranking information only, however, it cannot distinguish between different link functions or linear predictors that yield the same stochastic orderings."); *accord* Harrell, *supra* note 304, at 93 ("Rank measures [e.g., Somers' *D*] only measure how well predicted values can rank-order responses. . . . [A]lthough fine for describing a given model, [they] may not be very sensitive in choosing between two models.").

310 *See supra* notes 296 & 298; *accord* James S. Krueger & Michael Lewis-Beck, *Goodness-of-Fit: R-Squared, SEE and "Best Practice,"* 15 Pol. Methodologist 2, 4 (2007) (encouraging the reporting of multiple measures of model fit because "[a]fter all, the production and publication of these statistics is essentially cost-free").

S&S note that they used the statistical software *Stata* to conduct their analyses. Sander & Steinbuch, *supra* note 15, at 744. As a default, *Stata* provides the pseudo-R^2 statistic (and not Somers' *D*) as the measure of model fit for the logistic regression models that S&S analyze, so it seems odd that S&S do not report the pseudo-R^2. Generally speaking, the pseudo-R^2 statistic is a better measure of model fit than Somers' *D* because it is better able to capture differences in predictive ability between models. Harrell, *supra* note 304, at 93; *accord* Timothy M. Hagle & Glenn E. Mitchell, *Goodness-of-Fit Measures for Probit and Logit*, 36 Am. J. Pol. Sci. 762, 764–65 (1992) (encouraging analysts to use the pseudo-R^2 to evaluate logistic regression models).

S&S's decision to not present alternative (and more appropriate) model fit statistics, such as the pseudo-R^2, is puzzling because Sander reports these measures in his response to Ayres and Brooks, and these measures indicate very poor model fit—i.e., the explained variance ranging from 1% to 11%. Sander, *Replication of Mismatch Research, supra* note 24, at 81–82, tbls.5 & 7. Timothy Clydesdale's analysis of bar exam performance using the BPS data, which was published around the same time as Sander's *Systemic Analysis*, reported pseudo-R^2 measures for his models. Clydesdale, *supra* note 168, at 748–49 tbls.5A & 5B. Amy Farley and colleagues analyze bar exam performance in Ohio and note that they "compared Nagelkerke pseudo *R*-squared values for each model . . . they represent more generally an empirical estimate of how well a given model explains the data when compared to similar models . . . and may be more helpful when selecting among candidate models during model building." Farley et al., *supra* note 307, at 618.

311 Alan Agresti, Analysis of Ordinal Categorical Data 203–204 (2d ed. 2010) (discussing confidence intervals for Somers' *D*); *accord* Roger Newson, *Confidence Intervals for Rank Statistics: Somers' D and Extensions*, 6 Stata J. 309, 331 (2006) (explaining so-called nonparametric methods are often based on population parameters estimated from sample data, such as logistic regression analysis, to account for confounding variables, so confidence limits should be reported to account for estimation uncertainty).

both determining whether the association between the observed and predicted orderings is due to chance and comparing differences in Somers' *D* values between alternative model specifications.[312]

An appropriate model achieves a balance between goodness of fit and complexity.[313] The Somers' *D* statistic does not incorporate a penalty for model complexity, so the inclusion of additional variables may appear to improve model fit simply by capitalizing on chance, even when those variables do not appreciably improve the explanatory power of the model.[314] Moreover, "a model that is more complex than a working model need not provide much more explanatory power, regardless of whether its extra terms are statistically significant."[315] Statistical significance tests are not an indicator of the validity or stability of a model because these tests are based upon the underlying assumption that the model is correct.[316] In fact, it is not uncommon for a statistically significant parameter or parameters to fail to appreciably explain much of the variability in the dependent variable.[317] As I have explained previously, "Particularly telling is that Sander exclusively focuses on tests of statistical significance rather than the more important issue of the incremental

312 Newson, *supra* note 311, at 313 & 328 (noting that one cannot claim that one model is superior to another when the confidence interval for the difference between two Somers' *D* statistics includes a zero value).

313 Kenneth P. Burnham & David R. Anderson, Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach 29 (2d ed. 2002) (noting the "conflict between increasing model fit and increasing numbers of parameters to be estimated from the data").

314 Somers' *D* does not account for varying levels of complexity of the nested regression models used to calculate predicted probabilities. Newson, *supra* note 311; *see also* John K. Kruschke, Doing Bayesian Data Analysis: A Tutorial with R, Jags, and Stan 290 (2d ed. 2015) (noting that "data 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."); *accord* Krueger & Lewis-Beck, *supra* note 310, at 4 (discouraging the use of model fit statistics that do not impose a penalty for model complexity because such "measures [are] inflated, capitalizing on a chance fit through exhaustion of degrees of freedom"); Hagle & Mitchell, *supra* note 310, at 768 (discussing the importance of imposing complexity penalties for certain goodness-of-fit measures for logistic regression models).

315 Agresti & Tarantola, *supra* note 278, at 217.

316 John Fox, Applied Regression Analysis and Generalized Linear Models 670 (3d ed. 2016) (advocating for analysts to employ a criterion other than statistical significance to decide questions of model selection); *see also* Thaxton, *supra* note 1, at 893 ("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 [mismatch] variable to the model.").

317 To be clear, goodness of fit (i.e., degree of correspondence between expected and observed conditional means) is a separate question from statistical significance (i.e., whether the relationship between the explanatory variable(s) and the dependent variable observed in the data is due to chance).

improvement in model fit."[318] Given the additional parameters (variables) estimated across the different model specifications, adding a penalty for model complexity is highly advisable. So, for example, S&S's Table 3–Model 2 (see below) includes variables for race (Black and Latinx) and LSAT, and it has a Somers' *D* of 0.35. Table 3–Model 3 adds the mismatch variable, and the Somers' *D* minimally increases to 0.36. The fully specified model, which estimates nine additional parameters, has a Somers' *D* of 0.39. These trivial variations in Somers' *D* statistics strongly suggest that including mismatch variables in the model (either linearly or nonlinearly) does not appreciably improve model fit. Using a model fit measure that penalizes model complexity would have resulted in a decrease in model fit when estimating those additional parameters.

The lack of meaningful incremental improvement in S&S's models is consistent with my analysis of the BPS data, which reveals that a model that does not include mismatch when predicting first-time bar passage fits the data better than a model that does include mismatch when taking model complexity into account.[319] S&S's results from Table 4–Models 7–10 (which exclude UCLA) do not present a trimmed model (like Table 3–Model 2) that includes controls only for race (Black and Latinx) and entering credentials (LSAT/Index). This omission is unfortunate because it would have allowed for the assessment of whether adding the mismatch variable appreciably improved model fit. Nonetheless, the Somers' *D* for Model 7 (0.36) which includes LSAT-based mismatch as a variable, is nearly identical to that in Model 2 (0.35), which does

318 Thaxton, *supra* note 1, at 893.

319 *Id.* at 1005 ("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."); *see also* Smyth & McArdle, *supra* note 45, at 367 tbl.2 (discovering that the inclusion of a school selectivity variable in the model predicting graduation in STEM fields does not improve model fit).

Sander's reporting of misleading model fit statistics is not limited to his use of Somers' *D*. When conducting a mediation analysis on the BPS data to determine the direct and indirect effect (via LGPA) of mismatch on bar passage, Sander reported a single model fit statistic, the Bentler-Bonett Normed Fit Index (NFI), without including a discussion of the statistic and its limitations. Sander, *Response to Ho*, *supra* note 123, at 2008. Researchers strongly recommend against using the NFI because it is both overly sensitive to sample size (i.e., the NFI will indicate adequate model fit for data with more than 200 observations) and does not account for model complexity. *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). If the NFI is used, then multiple measures of model fit should also be reported. *Id.*; *accord* 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). I reanalyzed Sander's mediation analysis and reported two highly endorsed model fit statistics that are not unduly sensitive to sample size and impose a penalty for model complexity—the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). Thaxton, *supra* note 1, at 1004. Both measures, as well as their variants, the Consistent AIC (CAIC) and Adjusted BIC (ABIC), revealed that the inclusion of the mismatch variable (proxied by the tier location variable in the BPS data) did not improve model fit. *Id.* at 1005.

not include the LSAT-based mismatch variable. And as in Models 5 and 6, the estimation of seven more parameters in, respectively, Models 9 and 10 do not appreciably improve model fit (0.38 and 0.39).[320] The poor predictive ability of Sander's statistical models analyzing bar passage with the BPS data stems, in part, from the very limited number of explanatory variables included in those models. It is highly unlikely that only a handful of variables capture the key differences between students that impact law school outcomes. The BPS data contain a rich set of relevant variables that have been shown to improve the prediction of bar passage, so it would have been advisable for S&S to explore a wider range of explanatory variables to increase the predictive accuracy of their models.[321]

Another consequence of failing to account for a wider range of relevant variables that might be correlated with other variables in the model (i.e., confounders) is that the exclusion of these relevant variables may bias the finding of a mismatch effect.[322] Indeed, several reexaminations of Sander's analysis of the BPS data revealed that the mismatch effect (proxied by law school tier) vanished after controlling for a wider range of relevant explanatory variables.[323]

D. Model Dependence

Model dependence is a function of the distance from the counterfactual to the observed data.[324] "Some counterfactuals are more amenable to empirical analysis than others. . . in particular, some counterfactuals are more strained, farther from the data, or otherwise unrealistic."[325] Statistical inference becomes highly dependent on seemingly trivial, and often indefensible, modeling assumptions when counterfactual questions are posed too far from the data.

320 By comparison, the Somers' *D* is 0.51 for a similar model estimated on the BPS data, which is roughly equivalent to a pseudo-R^2 indicating that 14% of the overall variance in first-time bar passage is explained by the model.

321 Thaxton, *supra* note 1, at 863 ("The poor predictive ability of [Sander's] statistical models examining the likelihood of passing the bar [with the BPS data] underscores this point [i.e., Sander fails to consider relevant explanatory variables].")

322 *Id.*, at 886–88 & 974 n.461 (advising analysts to directly assess sensitivity of the estimate of the mismatch effect to the unconfoundedness assumption); *accord* Ho, *supra* note 28, 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").

323 *See* Daniel E. Ho, *Evaluating Affirmative Action in American Law Schools: Does Attending a Better Law School Cause Black Students to Fail the Bar?* (Mar. 9, 2005) (unpublished manuscript) (on file with the author); Jesse Rothstein & Albert Yoon, *Mismatch in Law School* 21–23 (Nat'l Bureau of Econ. Rsch., Working Paper No. 14275, 2008); Gregory Camilli & Darrell D. Jackson, *The Mismatch Hypothesis in Law School Admissions*, 2 Widener J.L. Econ. & Race 165, 199–203 (2011).

324 Thaxton, *supra* note 1, at 907.

325 Gary King & Langche Zeng, *When Can History Be Our Guide? The Pitfalls of Counterfactual Inference*, 51 Int'l Stud. Q. 183, 184 (2007).

"A testable counterfactual statement is one in which its probability can be inferred from the data, so when counterfactual questions invoke hypothetical questions that are incompatible with what is actually known or observed, testing counterfactuals is fraught with conceptual and practical difficulties."[326] S&S's counterfactual (i.e., "what if") question is what would happen to the racial gap in bar passage if Black, Latinx, and non-Black/non-Latinx students were evenly matched, academically speaking, in law school? To properly assess this counterfactual, the racial groups in the school-specific data "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."[327] Absent these "close enough" comparisons, "estimation of the model heavily relies on more dissimilar cases (more divergent counterfactuals) to measure the hypothesized treatment [i.e., race] effect . . . . 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 was similar along all other relevant dimensions."[328] In other words, in the absence of appropriate counterfactuals in the data, the "model basically pretends that there are data values that do not actually appear in the sample (or are extremely sparse in the sample)."[329] When examining racial differences in bar passage, sparse data for valid counterfactual comparisons may severely degrade the estimation of the race effect because (1) the observations exhibit insufficient variation across the range of included explanatory variables in the model to properly isolate the effect of race (i.e., collinearity) and (2) the limited number of observations available for comparison are unlikely to accurately represent the counterfactuals of the larger population (i.e., overfitting).[330]

326 Thaxton, *supra* note 1, at 894.

327 *Id.*

328 *Id.* at 896–98.

329 *Id.* at 906.

330 *Id.* at 898. Data problems impacting counterfactual inference also impact the functioning of the algorithms used to estimate the models. Generally speaking, analysts employing logistic regression must be mindful of three related issues: zero cell count, complete separation, and collinearity. All of these issues create numerical problems when attempting to estimate a model and produce unreliably large standard errors and effect sizes. In brief, these issues result from the structure of the data and negatively impact the computation of parameter estimates. The most likely culprit is "thin" data—i.e., not enough outcomes for the dependent variable and/or small frequencies of observations for a categorical covariate. Zero-cell-count problems result from spreading the data over too many levels of a categorical variable, with no observations for a particular level of the explanatory variable. Complete separation occurs when a collection of the covariates completely separates the outcome groups (e.g., pass/fail). Consequently, knowing the value of one variable in the model would perfectly predict the value of the outcome variable because there is no overlap in the distribution of the covariates between the outcome groups. Collinearity (also known as multicollinearity and ill conditioning) occurs when one or more explanatory variables in the model are approximately determined by a linear combination of other explanatory variables in the model. *See* Hosmer & Lemeshow, *supra* note 305, at 145–50. In practice, analysts often encounter near-zero cell count, near-complete separation (i.e., "quasi-complete separation"),

These two problems are closely related, leading one scholar to state, "In fact, at a theoretical level, multicollinearity, few observations, and small variances on the independent variables are essentially all the same problem."[331] Below, I address to these issues—collinearity and overfitting—in reverse order.

With respect to the second issue (i.e., nonrepresentative or nonexistent counterfactuals), S&S mention that there is substantial overlap in entering credentials across schools, so relevant comparisons could be made. Table 2 displays overlap for LSAT, but not UGPA (or Index);[332] however, Table 2 does not display cross-racial overlap, which is puzzling considering S&S's focus is on whether controlling for mismatch eliminates the impact of race on the

---

and near collinearity (i.e., "non-exact collinearity"), such that the model can be estimated on the data, but the parameters are unstable. Micah Altman et al., Numerical Issues in Statistical Computing for the Social Scientist 242 (2004) ("Of the two conditions, complete and quasi-complete separations, the latter is far more common.").

331 Christopher H. Achen, Interpreting and Using Regression 82 (1982); *accord* Guoping Zeng & Emily Zeng, *On the Relationship Between Collinearity and Separation in Logistic Regression*, 50 Comm. Stat. 1989 (2018) (noting that highly correlated covariates, covariates with very strong effects, rare outcomes, rare exposures/treatments, small samples, and sparse data result in the same type of biases).

332 According to S&S, Table 2 "makes a very powerful case that the mismatch effect not only exists, but is quite large." Sander & Steinbuch, *supra* note 15, at 726. Their claim must be viewed with caution because they exclude three cohorts from Davis in Table 2, stating that these three "cohorts had a median LSAT close to 162" and they "would not expect these cohorts to provide any 'contrast' with [UCLA]." *Id.* at 726 n.24. Consequently, 1411 Davis bar-takers (60% of the Davis bar-takers) are *not* included in the first-time bar passage rates reported in Table 2.

S&S's explanation for removing these bar-takers is unconvincing for at least four reasons. First, in a prior draft of their paper, they report data from cohorts from UCLA from different years (2001, 2005 and 2006), and they include all available cohorts from Davis. Richard Sander & Robert Steinbuch, *Mismatch and Bar Passage: A School-Specific Analysis (Oct. 2017 Draft)*, 11 tbl.4 (Oct. 2017) [hereinafter Sander & Steinbuch, *2017 Draft*]. Second, S&S could have included all cohorts by subtracting each student's LSAT from their cohort median LSAT and examined the impact of credential deficits on first-time bar passage. S&S adopted this approach in a prior draft of their paper when comparing Black and non-Black bar-takers. Richard Sander & Robert Steinbuch, *Mismatch and Bar Passage: A School-Specific Analysis (Aug. 2018 Draft)*, 14 tbl.6 (Aug. 2018) [hereinafter Sander & Steinbuch, *2018 Draft*]. In fact, S&S calculate credential deficits, rather than model cohort median LSAT/Index, in their logistic regression models. Sander & Steinbuch, *supra* note 15, at 733. Third, S&S exclude 121 UCLA bar-takers from Table 2 (16% of all UCLA in-state bar-takers in the years studied) without explanation. Perhaps these bar-takers had missing data on their LSAT score or bar results, but S&S do not offer an explanation. S&S note that the final two cohorts of Davis data were impacted by sizable missing data on bar outcomes, but they did not indicate UCLA's data were similarly hampered by missing values. Finally, S&S do not account for fluctuations in bar difficulty within each jurisdiction. For the years examined (1997–2011), the bar passage rate in California ranged from 48% to 63% for the July bar exam, and from 33% to 42% for the February exam. Students with identical absolute and relative credentials may perform differently on the bar exam simply because that particular administration of the bar was substantially more (or less) difficult compared to other years. Comm. Bar. Exam. of State Cal., *General Bar Examination Pass Rate Summary* (2023), http://www.calbar.ca.gov/Portals/0/documents/admissions/Examinations/GBX-Passrate-Summary.pdf.

likelihood of bar passage.[333] S&S acknowledge that "the racial composition

333 *See, e.g.*, Rothstein & Yoon, *supra* note 166, at 667–69 (describing racially disaggregated distributions of UGPA, LSAT, and Index for law school applicants, 1990–1991). S&S's exclusion of racially disaggregated distributions is puzzling because, in an earlier draft, they include a table with these numbers. Sander & Steinbuch, *2018 Draft*, supra note 332, at 14 tbl.6. This table (Table 6) is reproduced in the appendix of this article, also as Table 6, and it reveals three issues with their analysis. First, S&S separate students into two groups—"non-Black bar-takers" and "Black bar-takers only"—but the bulk of their analyses disaggregate students into three racial groups: Black, Latinx, and "mainly Anglos and Asian-Americans." *Id*. at 10, 13–14. Assuming that S&S included Latinx students in the "non-Black bar-takers" group, it seems odd that S&S would group non-Black and Latinx students together given that S&S posit that "African-Americans and Hispanics [are] substantially less likely to pass (and more likely to fail) the bar than . . . Anglos and Asian-Americans," *id*. at 10, and given that S&S's core analyses separate Black and Latinx students. Combining Black and Latinx students would have been a sensible choice, not combining Latinx with Caucasians and Asian-Americans. If S&S excluded Latinx students from the "non-Black bar-takers," that is not obvious from their paper.

Robert Steinbuch published a prior co-authored study that analyzed the same data from UALR and reported racially disaggregated statistics for the 723 bar exam-takers: 624 white students, 56 Black students, 11 Latinx students, 11 Native American students, 9 Asian-American students, 7 multiracial students, and five students of undeclared background. Robert Steinbuch & Kim Love, *Color-Blind Spot: The Intersection of Freedom of Information and Affirmative Action in Law School Admissions*, 20 Tex. Rev. L. & Pol. 181, 215–16 (2016). The bar passage rate was 79.8% (498) for white students, 58.9% (33) for Black students, 81.8% (9) for Latinx students, 45.5% (5) for Native American students, 100% (9) for Asian-American students, 71.4% (5) for multiracial students, and 60% for the undeclared students. *Id*. at 216 tbl.1.

Second, Table 6 in S&S's 2018 draft reveals very few comparable Black students at the various levels of mismatch. The table reports numbers for the two schools with information on students' Index scores (Davis & UALR) and reveals that there are only 120 Black students in the entire sample. In the "no deficit" group (i.e., the non-mismatched group), there are only 13 Black students (compared with 1,495 non-Black students). As the degree of mismatch increases, the number of Black students in each cluster is, respectively, eleven, twenty-three, thirty, twenty, eighteen, and five. Sander & Steinbuch, *2018 Draft*, *supra*, at *14 tbl.6. Steinbuch's prior analysis revealed that fifty-six of those Black students graduated from UALR, Steinbuch & Love, *supra*, at 215, so the remaining sixty-four Black students in the dataset were from Davis.

Third, 56% of the 120 Black students in Davis and UALR passed the bar exam on the first attempt, compared with 81% of the 2,885 non-Black students—a difference of twenty-five percentage points. It is quite revealing that there are substantial sample-based differences in the likelihood of passing the bar between non-Black students and Black students at several of the mismatch levels, although S&S hypothesize that that such differences should not exist. So, for example, at mismatch level 2 (involving a 3-to-6-point deficit in a student's Index score relative to the median student in the same school-specific cohort), non-Black students pass the bar at a rate nineteen percentage points *higher* than Black students (71% v. 52%). In contrast, at mismatch level 4 (with a 9-to-12-point deficit), non-Black students pass the bar at a rate twenty-one percentage points *lower* than Black students (39% v. 60%). For "no deficit" and mismatch levels 1, 3, 5, and 6, the percentage point differences are much smaller (or nonexistent)—respectively, 3%, 8%, 7%, 4%, and 0%.

While the majority of the mismatch levels reveal modest racial differences in bar performance, Black students at mismatch levels 2 and 4 comprise, collectively, over one-third (35.8%) of the Black students in the sample. The cross-racial comparisons of the other mismatch levels should also be interpreted with caution because the bar passage rate for Black students would change significantly if a very small number of students were to pass

of students with LSATs of 150 to 155 varies considerably across our three schools"[334] and opine that " 'controlling' for race in a regression would separate out the 'race' effect from the 'mismatch' effect"[335] if "most of these students at UCLA were Black, while most of these students at Davis were white."[336] But to accurately estimate the race effect (or absence thereof) there must be sufficient numbers of both groups of students for every level of credential difference that is examined.[337] With multiple control variables (e.g., LSAT/Index), there must be sufficient numbers of both racial groups with the same combination of values of the control variables.[338] When no cross-comparisons with similar values on the control variables exist in the data (or are extremely sparse in the sample), the expected outcome value (and hence the treatment effect) is solely a function of the parametric assumptions of the model.[339] Gary King and Langche Zeng underscore that there can be "vast" differences between regression coefficients based on linear and nonlinear parametric assumptions when counterfactuals are not data driven; they caution that they "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."[340]

---

or fail the bar at the various mismatch levels; but the same would not hold for non-Black students because of their large numbers in each grouping.

334 Sander & Steinbuch, *supra* note 15, at 729.

335 *Id.*

336 *Id. But see infra* Part III.E (explaining that S&S's inclusion of school-specific fixed effects removes between-school variability and provides misleading answers to questions about the effects of the student-level characteristics).

337 Makin & Orban de Xivry, *supra* note 27, at 9 (noting that designs with small sample sizes are also more susceptible to missing an effect that exists in the data and that, "if the researchers wish to interpret a non-significant result as supporting evidence against the hypothesis, they need to demonstrate that this evidence is meaningful. The *p*-value in itself is insufficient for this purpose.").

338 Thaxton, *supra* note 1, at 895.

339 *Id.* at 898 n.231 (noting that, when there are no or very few counterfactuals that exhibit sufficient variation in treatment, the model will, essentially, guess the treatment effect for that counterfactual). As one economist explained,

> Empiricists are likely to compensate for [a] shortage of observations, and even for the absence of any observations, with certain combinations of variables, by interpolations—but this requires some assumptions about underlying distributions. . . . To be sure, extrapolations and interpolations are reasonable when [the functional form assumptions are warranted]. But with zig-zags [i.e., nonlinearities] or other omitted variables that completely change the underlying distribution on one side of the line to be filled by interpolation, the process is misleading or simply impossible.

Levmore, *supra* note 80, at 623.

340 Gary King & Langche Zeng, *The Dangers of Extreme Counterfactuals*, 14 Pol. Anal. 131, 135 (2006).

S&S offer a thin discussion of their construction of the mismatch variable.[341] For the LSAT-based mismatch variable, students' LSATs are subtracted from the school-specific median LSAT for their cohort scores.[342] Students with LSAT scores at or above the median LSAT are recoded as "zero" (also known as "top-coding" or "censoring to bounds") and the students below the median LSAT have mismatch scores reflecting the distance of their LSAT from the cohort-specific median LSAT.[343] S&S employ a similar approach for their Index-based mismatch variable.[344] While the literature on censored variables is concerned primarily with *dependent* variables and has developed methods to address top-coding without excessive risk of biasing results, censoring *independent* variables discards important information, alters the mean and variance of the variables, and can bias results.[345] Top-coding results in "measurement error [that] is not random, but a direct function of the independent variable that is subject to a ceiling effect."[346]

341 *See* Makin & Orban de Xivry, *supra* note 27, at 7 (noting that researchers should justify the independence between the adopted model specification and the effect of interest).

342 Sander & Steinbuch, *supra* note 15, at 729 ("[W]e measure 'mismatch' as the LSAT 'deficit' between a student's LSAT and the median LSAT of her classmates.").

343 *Id.* at 729 ("If a student's LSAT is [at or above the median LSAT for her cohort], then the student is not mismatched, and has a mismatch value of zero.").

344 *Id.* at 735 ("A common way to combine information on LSAT and UGPA is by creating an 'academic index' that weighs the two credentials in a way that roughly maximizes their joint ability to predict performance in law school. We used the data from Davis and UALR to create such an index, and we scaled it from 0 to 100 [sic] so that it would be comparable to the scale used for our LSAT-only measures . . . . We also recalculated each student's mismatch level, examining each of our six remaining cohorts separately and determining whether, and how far, each student's index put her below the median index of her cohort."). *But see infra* note 502 (noting that S&S's do not examine the sensitivity of their results to different weights assigned to UGPA and LSAT for the creation of the Index).

345 Roberto Rigobon & Thomas M. Stoker, *Bias from Censored Regressors*, 27 J. Bus. & Econ. Stat. 340 (2009). ("Coefficients also will be biased if the values of regressors are censored; however, there has been very little study of this phenomenon in the literature."). Top- and bottom-coding (e.g., S&S's coding of all nonmismatched students as "zero") are instances of nonignorable data coarsenings. *Id.* Data are "coarse" when they are an imperfectly observed version of the underlying ideal data, and the coarsening is "nonignorable" when the degree of coarseness depends on the true value of the ideal data. Put simply, S&S's censoring of the mismatch variable likely induces statistical dependence between the mismatch variable and the error term in the model (i.e., endogeneity). *Id.*

It is also important to recognize the relationship between censored regressors and correlated regressors (i.e., collinearity) as it pertains to bias. Rigobon & Stoker note that "[b]ias from censoring one variable can contaminate the estimates of coefficients of other variables, a situation we refer to as bias transmission," *id.* at 346, and the "transmission of bias with correlated regressors can result in enormous bias in estimation," *id.* at 341. Bias transmission is a serious concern irrespective of whether the censored variable is of primary interest. *Id.* at 346; *see infra* notes 380–399 and accompanying text (discussing how collinearity exacerbates small biases from model misspecification).

346 Peter C. Austin & Jeffrey S. Hoch, *Estimating Linear Regression Models in the Presence of a Censored Independent Variable*, 23 Stat. Med. 411, 413 (2004); *see also* Thaxton, *supra* note 1, at 949 (noting

S&S's decision to top-code their mismatch variable is not the only source of measurement error. They also compound this measurement error by dividing the mismatched students into smaller groupings in an ad hoc manner in an effort to capture any potential nonlinearity in the effect of mismatch on bar exam success. For their linear specification, the effect of mismatch on the odds of bar passage is assumed to be a constant factor change across the range of the censored mismatch variable.[347] For the nonlinear specification, S&S "break 'mismatched' students into a total of seven categories, with 'MM1' comprising students only slightly below the median, and 'MM7' comprising those students furthest from the median."[348] The effect of mismatch on the odds of bar passage for the nonlinear specification is the factor change relative to the non-mismatched reference group.[349] Yet, it is unclear how S&S divided the data into categories, which itself may affect results. That is, S&S do not offer a detailed discussion of how the division of the variable was conducted or of the sensitivity of their results to different coding choices.[350] Recoding a continuous variable into categories—called "discretization"—may have a strong influence on the estimation of the effect of the variable,[351] as well as the generalizability

that "measurement error bias is of particular concern when the error in measuring a variable of interest is correlated with the true value of that variable").

347 Long, *supra* note 275, at 80–81. Sometimes it is more intuitive to compute the percentage change in the odds rather than a factor change. *Id*.

348 Sander & Steinbuch, *supra* note 15, at 733 ("[W]e separately test the effect of having an LSAT just slightly (e.g., 1 or 2 points) below the median, of having an LSAT a little further (e.g., 3 or 4 points) from the median, and so on."). In an earlier draft, S&S construct the mismatch levels (using Index) using three-point increments, but they neither offer an explanation as to why those particular cut-points were selected nor do they discuss any sensitivity analyses conducted on those cut-points. Sander & Steinbuch, *2018 Draft*, *supra* note 332, at *13–14 & tbl.6. In another draft of their paper, S&S use slightly different cut-points. Sander & Steinbuch, *2017 Draft*, *supra* note 332, at *15 tbl.7.

349 Long, *supra* note 275, at 80.

350 *See* Raj Echambadi et al., *Encouraging Best Practice in Quantitative Management Research: An Incomplete List of Opportunities*, 43 J. Mgm't Stud. 1801, 1807 (2006) ("[S]plitting continuous variables into categorical variables leads to reduced power and deleterious consequences. If the variable must indeed be [discretized], then researchers must conduct robustness tests to show that the results do not change when the dividing line is specified differently[.]"). *See generally* Angrist & Pischke, *supra* note 93, at 16 (noting that transparent discussions of research design have been crucial to keeping the "con" out of econometrics); *accord* Michael C. Lovell, *Data Mining*, 65 Rev. Econ. & Stat. 1, 10–11 (1983) (criticizing analysts' failures to report and justify results from model specifications that adopt different coding conventions for variables); Thaxton, *supra* note 1, at 969 (noting that "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.").

351 Harrell, *supra* note 304, at 8 (cautioning that categorizing continuous predictors allows the analyst to manipulate the results); Makin & Orban de Xivry, *supra* note 27, at 7 (stating that an unjustified or poorly justified division of data/variables in subgroups (i.e., "binning") is often an indication of an attempt to manipulate results); Patrick Royston & Willi Sauerbrei, Multivariable Model-Building: A Pragmatic Approach to Regression

of the model.[352] According to biostatistician Frank Harrell, "It is a common belief among practitioners who do not study bias and efficiency in depth that the presence of non-linearity should be dealt with by chopping continuous variables into intervals. Nothing could be more disastrous."[353] Rather than top-coding and modeling the nonlinear effect of mismatch with a discretized variable, S&S should have used nonparametric smoothing to allow the data to reveal any potential nonlinearity (called functional form) in the relationship between mismatch and bar exam performance. This is the preferred choice to estimating nonlinear relationships when there is no *a priori* reason to settle on a particular nonlinear relationship between the explanatory and outcome variable.[354] S&S's recoding of the raw mismatch variable seems counterintuitive because they expressly claim that the value of the school-specific data over the BPS data is the granularity of the school-specific mismatch measure. In fact, S&S criticize LSAC for "engineer[ing] [the BPS to be] a second-rate research tool that "was unable to provide direct answers to many of the most interesting questions [the LSAC] had been intended to study."[355] But S&S's coding of the mismatch variable does the same: It discards important information about students' degree of mismatch, and S&S are not transparent about the sensitivity of their results to different coding conventions.[356]

In addition to potential bias from top-coding and discretizing their mismatch variable, a third concern with S&S's mismatch variable is the presence of random measurement error (as distinguished from the

Analysis Based on Fractional Polynomials for Modelling Continuous Variables 151 (2008) ("For continuous predictors, categorization is inadvisable because the results depend on the chosen cut-point(s)."); Melissa A. Hardy, Regression with Dummy Variables 84 (1993) (noting that the "proper use and interpretation of dummy variables does involve a number of complexities"); *see also infra* notes 381–386 (identifying additional concerns with S&S's construction of their mismatch variable).

352 Thaxton, *supra* note 1, at 906 n.261; *see also infra* note 409 and accompanying text.

353 Harrell, *supra* note 304, at 19.

354 *Id.* at 21 ("A better approach [than discretizing] that maximizes power and that only assumes a smooth relationship is to use a regression spline [i.e., nonparametric regression] for predictors that are not known to predict linearly."); *accord* Luke Keele, Semiparametric Regression for the Social Sciences 8 (2007) ("Instead of assuming that we know the functional form for a regression model, a better alternative is to estimate the appropriate functional form from the data. In the absence of strong theory for the functional form, this is often the best way to proceed."). With nonparametric regression, "an estimate such as a mean or regression [coefficient] is estimated between Y and X for some restricted range of X." *Id.* at 9. Discretizing a continuous variable may inadvertently contribute to problems of (near-)zero cell count, (quasi-)complete separation, and (near-exact) collinearity by artificially limiting the observed variation in the explanatory variables. *See supra* note 330. S&S mention that they explored nonlinear effects for their mismatch variable using polynomials, Sander & Steinbuch, *supra* note 15, at 734 n.33, but modeling nonlinearity via conventional (i.e., low-rank) polynomials is generally discouraged because of their limited range of nonlinear forms, *see* Royston & Sauerbrei, *supra* note 351, at 2.

355 Sander & Steinbuch, *supra* note 15, at 723; *see supra* note 137.

356 *See supra* note 351.

aforementioned nonrandom error due to top-coding). An individual's true LSAT score lies within an interval which reflects uncertainty associated with each administration of the test to the same test-taker.[357] LSAC, which administers the LSAT, reports a general measure of reliability for the LSAT (a score band) that is plus/minus 2.6 points of the obtained score.[358] The actual magnitude of this uncertainty differs at various points along the score scale, with the maximum uncertainty (at the middle of the score range) often more than twice the minimum (at the extremes of the score range).[359] However, S&S's analysis treats a respondent's LSAT score as if it accurately captures their underlying ability without accounting for uncertainty in this measure.[360] By discretizing a variable (LSAT) that is known to be measured with error, S&S's analyses potentially distort the relationship between mismatch and bar passage by assigning a level of mismatch to a student that does not reflect the true credential difference relative to other students.

This problem of uncertainty in the measurement of students' true LSAT score is especially salient here. "Measurement error bias is of particular relevance in causal inference because using error-prone measurements will either misidentify the treatment and control groups, misjudge the treatment dosage, or fail to balance the treatment and control groups with respect to the confounding variables,"[361] thereby negatively impacting parameter estimation, hypothesis testing, prediction, and model selection.[362] S&S's analyses are susceptible to this problem, and their decision to top-code and discretize the mismatch variable adds more noise to an already unreliable measure.[363]

S&S note that, for the discretized mismatch variable, the mismatch coefficients reveal contrasts relative to the baseline category (indicating "no mismatch"), so adjacent categories may not be meaningfully different.[364]

357 Thaxton, *supra* note 1, at 953 (discussing measurement problems associated with the LSAT and the potential impact of these measurement issues on examinations of mismatch theory).

358 *Id.*

359 *Id.*

360 *Id.* (identifying techniques to incorporate the level of imprecision in the measurement of a variable if the reliability of the measurement is known in advance, as in the case of the LSAT).

361 *Id.* at 929.

362 *Id.* at 929 n.302.

363 *Id. see also* Loken & Gelman, *supra* note 137, at 584–85 (explaining that effect sizes, in the presence of measurement error, may be biased upward or downward because researchers can engage in data mining to find the desired effect and neglect to report unsupportive effects).

364 Sander & Steinbuch, *supra* note 15, at 734. To be clear, splitting an explanatory variable into categories may be desirable when the goal of the analyst(s) is to "present results in a form that is understandable to a general audience [in order to] minimize the gap between the numerical results and the substantive conclusions." Andrew Gelman & David K. Park, *Splitting a Predictor at the Upper Quarter or Third and the Lower Quarter or Third*, 62 Am. Stat. 1 (2008). Nonetheless, discretization is not generally recommended because of loss

Standard errors for the coefficients for each of the mismatch categories are not reported, but it is likely that the contrasts between several (if not all) of the adjacent categories are statistically indistinguishable given the similarity in effect sizes (relative to the baseline category). Table 3–Models 5 and 6 reveal that the first two mismatch categories (i.e., the categories representing the two lowest levels of mismatch) are statistically indistinguishable from the baseline category (and each other). Across all of S&S's models, the first (or first two) mismatch categories is (are) statistically indistinguishable from the baseline category. S&S do not report or mention the results from any tests of comparisons of the discretized coefficients, but that would have been informative as to whether seven categories were, in fact, needed.[365] This is relevant because the extra parameters, if redundant, are likely to result in overfitting the data.[366] Collapsing some of these adjacent categories into a smaller number of categories (sometimes called dimensionality reduction) would result in more sizable groups of cross-racial comparables that are not too divergent.[367]

More pointedly, it is crucial to know how many Black and Latinx students in each school would have the same LSAT/Index as white students with a similar degree of mismatch. This would reveal whether the comparisons are

---

of statistical power and sensitivity to the choice of cut-points. *Id*. Moreover, the growing trend in empirical legal studies is for analysts to avoid presenting results in tabular form (e.g., tables with regression coefficients) and instead utilize effective data visualization (e.g., graphs and figures). *See* Epstein et al., *supra* note 65, at 804–05 ("[R]esearchers should almost always graph their data and results [because] a well-designed graph is a superior choice to a table. To put it another way, with only limited exceptions, we interpret the phrase 'effective communication' in our title to mean 'effective graphical presentations.'"). Consequently, in most situations, discretizing a continuous variable to simplify the communication of results will be unnecessary.

365 Harrell, *supra* note 304, at 186 (describing the Wald test that is used for testing the (dis) similarity of regression coefficients for levels of a categorical variable); *see also* Fox, *supra* note 316, at 320 (discussing the *F*-test for assessing the appropriate number of categories for a discretized continuous variable).

366 Recall that, as the number of parameters (variables) increases, data become sparser, which results in overfitting. Overfitting undermines the generalizability of the model. *See supra* note 314 and accompanying text; *see also* Makin & Orban de Xivry, *supra* note 27, at 7 (noting that data mining "works by recruiting noise [in the data] to inflate the desired effect," and that "the most straightforward solution is to use a different part of the dataset for specifying parameters for selecting sub-groups and for examining differences across sub-groups") (quote edited for clarity).

367 Thaxton, *supra* note 1, at 902 (explaining that, in controlling for race-related covariates, it becomes difficult to find similarly situated cross-race comparisons because certain traits are highly clustered within certain groups (i.e., there is a lack of "common support" for the distribution of data values for the different racial groups), so cross-racial comparisons are typically heavily reliant on model assumptions rather than the actual data). *See* Rothstein & Yoon, *supra* note 323, at 3, 19 (noting very little overlap in credentials between Black and white students in the lower-tier law schools); Sen & Wasow, *supra* note 129, at 505 (explaining that common support problems occur because certain traits are highly clustered within certain groups, such that it becomes difficult to find cross-racial comparisons).

data driven as opposed to merely model driven. Other scholars explain why: "The kind of high:low estimates that result from categorizing a continuous variable are not scientific quantities [and] their interpretation depends on the entire sample distribution of continuous measurements within the chosen interval."[368] When there are "sample size limitations in the very low and very high range of the variable, the outer intervals (e.g., outer quintiles) will be wide, resulting in significant heterogeneity of subjects within those intervals, and residual confounding."[369] Prior research on mismatch using the BPS data underscored that there are very few Black-white comparables in terms of their degree of mismatch—especially within schools located in the high and low tiers.[370] S&S's failure to include data on the number of Black and Latinx students in each school with the same LSAT/Index as white students with a similar degree of mismatch makes it impossible to know whether their results are data-driven or model-driven.

As I have previously noted, in examining the impact of race on law-related outcomes and using a parametric model to adjust for potential confounding variables, bias arises from controlling for confounders with the wrong functional form when the distribution of the covariates for the treatment group (e.g., Black/Latinx students) differs from that for the control group (e.g., white students).[371] And this bias arises even if the form of the relationship between the treatment variable and the outcome of interest is correct.[372] If the parametric assumption is incorrect, then the statistical adjustment will fail to remove the bias from the included confounding variable(s).[373] Put simply, the coefficient of the race effect will be biased (either upward or downward) if the functional form of the effect of mismatch is incorrect.[374]

A comparison of Table 3–Models 4 and 6 underscores the sensitivity of the alleged race effect based on the parametric assumption of the effect of mismatch. Specifically, the race effect for Black students is statistically significant in Model 4 with a linear effect for mismatch, but the race effect is statistically insignificant in Model 6 with the nonlinear mismatch effect.[375] The Somers' *D*

368 Harrell, *supra* note 304, at 20–21.

369 *Id.* at 19.

370 Thaxton, *supra* note 1, at 908–11; *see supra* note 19.

371 *Id.*, at 897.

372 *Id.*

373 *Id.*; *accord* Andrea Benedetti & Michal Abrahamowicz, *Using Generalized Additive Models to Reduce Residual Confounding*, 23 Stat. Med. 3781, 3782 (2004) (noting that a proper nonparametric representation of the effect of an explanatory variable will result in less bias than a parametric effect based on a categorization of the variable).

374 Long, *supra* note 275, at 24 (noting that, in addition to omitting a confounding variable, model misspecification arises from measurement error, incorrect functional form, and unmodeled interaction effects).

375 Sander & Steinbuch, *supra* note 15, at 731 tbl.3.

statistics reported for both models are identical (0.39), suggesting that both specifications provide the same degree of model fit and there is no principled reason to choose one model over the other. Eliminating or significantly mitigating the aforementioned issues with S&S's coding and modeling of the mismatch variable would entail: (1) using an uncensored version of the mismatch variable (i.e., no top-coding);[376] (2) accounting for random measurement error;[377] and (3) modeling the nonlinear effect of mismatch nonparametrically.[378] Moreover, because S&S do not directly examine learning mismatch (proxied by LGPA), their models are not comparing students with the same degree of learning (relative to the median student) because, as explained earlier, the relationship between entering credentials and LGPA differs across races, *and* entering credentials explain only a modest amount of variation in LGPA (i.e., entering credentials do a poor job of predicting LGPA). Relatedly, the relationship between entering credentials, LGPA, and bar passage may vary across schools.[379]

With respect to the first issue (i.e., collinearity), the inconsistency between Table 3–Models 4 and 6 may also reveal problems with S&S's construction and analysis of their mismatch variable.[380] To avoid exact collinearity between the individual LSAT variable and the mismatch variable, S&S censor the mismatch variable at "zero" for all nonmismatched students.[381] S&S also state

376 *See supra* notes 345–346 and accompanying text.

377 *See* Thaxton, *supra* note 1, at 953 ("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 from the data (if there are multiple measurements of the same underlying phenomenon).").

378 *See supra* note 354 and accompanying text.

379 *See, e.g.*, Klein & Bolus, *supra* note 176, at 12.

380 *See also supra* notes 350–352 and accompanying & *infra* note 508 (describing the impact of collinearity on the estimation of the effect of race).

381 *Compare* Sander & Steinbuch, *supra* note 15, at 729 (noting that "we use LSAT scores **both** to measure student credentials and to measure a student's degree of mismatch. That means that the 'credential' variable and the 'mismatch' variable will be correlated—perhaps highly correlated.") (emphasis in original) *with* William D. Berry, Understanding Regression Assumptions 25 (1993) ("In practice, perfect multicollinearity will occur in only three kinds of cases. One is when an analyst mistakenly includes a set of independent variables that have a "built-in" linear relationship among them.").

S&S include students' credentials to capture a "credential effect" on bar exam performance, in addition to the "mismatch effect." Sander & Steinbuch, *supra* note 15, at 729. Assuming, *arguendo*, that controlling for students' credentials is a defensible modeling choice, *but see infra* note 386, S&S's top-coding and discretizing of the mismatch variable but not the credential variable distorts the interrelationships (i.e., both bivariate and partial correlations) among the variables in the model. In fact, S&S alert readers to interpret the relative effects of LSAT and mismatch with caution because "LSAT is being measured on a broader scale, with a broader range of possible values, than is mismatch." Sander & Steinbuch, *supra* note 15, at 733. A common approach to comparing parameter estimates of variables measured on different scales is to report standardized coefficients, but because

that the mismatch measure varies across the different cohorts of students, depending on the median LSAT of each cohort.[382] In other words, students who attend the same school with identical individual LSAT scores but enter law school in different cohorts will have different mismatch scores. Based on this description, it is unlikely that the cohort-specific mismatch measure would appreciably reduce the within-school collinearity between LSAT and mismatch, given that the median LSAT for each cohort within a school would be rather consistent over the time frame studied, especially considering that not all cohorts are distinguishable.[383]

According to S&S, their choices in coding were driven by differences in the specificity of data provided by the schools included in their study. S&S report that only the data for UCLA distinguish between all three cohorts. For Davis, five of the fifteen cohorts can be distinguished, and no distinguishable cohorts exist for UALR.[384] This implies that near-exact collinearity is primarily avoided by collapsing (i.e., top-coding) roughly half of the students (i.e., matched and positively mismatched) into a single category. S&S's approach underscores how one decides to code the mismatch variable may significantly alter the results and why S&S need to be more transparent about their coding preferences.[385]

---

the truncation of the scale of the mismatch variable is a consequence of S&S's coding decisions, the problem cannot be remedied by standardized coefficients. *See generally* Long, *supra* note 275, at 16 (explaining that semi-standardized coefficients represent a unit change in the dependent variable for a standard deviation increase in the explanatory variable, and that fully standardized coefficients represent a standard deviation change in the dependent variable for standard deviation change in the explanatory variable).

382 Sander & Steinbuch, *supra* note 15, at 729.

383 S&S provide information about the median LSAT for each school across all cohorts but do not provide cohort-specific median LSAT data for each school. Sander & Steinbuch, *supra* note 15, at 725 tbl.1. The American Bar Association (ABA) provides annual data via its website on, inter alia, the LSAT and UGPA for entering classes at all ABA-accredited law schools. Am. Bar Found., *supra* note 200. Although the available data on the website dates back only to 2011, the past decade of disclosures for UCLA, Davis, and UALR reveal that the median LSAT and UGPA for each cohort within the law school was quite stable. *Id*. This is not surprising considering that admissions committees typically operate with a "target" median LSAT (and median UGPA) for each entering class. *See also infra* note 386 (noting how the consistency of LSAT scores across cohorts for UALR implicitly "controls" for median LSAT when examining data solely for UALR).

384 *See supra* note 19 and, *infra*, "Appendix: Table A."

385 *See infra* notes 350–369 and accompanying text (describing potential problems caused by discretizing a continuous variable). For example, it is common for researchers to discretize a continuous variable into three groups: one standard deviation above the mean, one standard deviation below the mean, and one in between those two groups. When the continuous variable is normally distributed, the middle group will contain 68% of the students, and the high and low groups will each contain about 16% of the students. If the median (fiftieth percentile) is preferred, then one group is above the seventy-fifth percentile, one group is below the twenty-fifth percentile, and one group is between the twenty-fifth and seventy-fifth percentiles (called the interquartile range).

It also would have been helpful for S&S to report the correlations and (generalized) variance inflation factors of the parameter estimates in the models in the overall sample and separately for each school to identify the actual magnitude of parameter collinearity and its impact on the parameter uncertainty.[386] According to political scientists Micah Altman, Jeff Gill, and

386 *See* David A. Belsley et al., Regression Diagnostics: Identifying Influential Data and Sources of Collinearity 91 (2005) (noting that diagnosing collinearity is important for regression analysis and consists of two related elements: detecting the presence of collinear relationships in the data and assessing the degree to which those relationships degrade estimated parameters). John Fox & Georges Monette, *Generalized Collinearity Diagnostics*, 87 J. Am. Stat. Assoc. 178, 180 (1992) (developing a test for assessing parameter uncertainty in the presence of collinearity for models that include sets of dummy variable regressors). When analyzing collinear data, parameter estimates "can be highly unstable with respect to very small amounts of noise in the data." Altman et al., *supra* note 330, at 46. This feature underscores the importance of model cross-validation and model averaging to assess the degree to which noise in the data impacts the instability of parameter estimates. Thomas Lindner et al., *Beyond Addressing Multicollinearity: Robust Quantitative Analysis and Machine Learning in International Business Research*, 53 J. Int'l Bus. Stud. 1307, 1311–12 (2022). *See also infra* notes 406–410 and accompanying text.

When conducting their robustness analyses, S&S also should have examined models that included a noncensored mismatch variable and excluded individual LSAT/Index. *See supra* note 381. S&S justify their inclusion of individual LSAT/Index because it is a strong predictor of performance on the bar exam *and* may explain up to half of the Black-white bar passage gap. Sander & Steinbuch, *supra* note 15, at 729 (positing a so-called "credential effect" that exists independent of degree of mismatch). S&S reference Steinbuch and Love's analysis of identical data for UALR to support their claim of a credential effect. *Id*. at 729 n.28. However, Steinbuch & Love's analysis does not distinguish between these two types of effects. Steinbuch and Love's core model predicting bar exam passage includes LGPA, LSAT, and UGPA. Steinbuch & Love, *supra* note 333, at 221 tbls.4 & 5. The seven cohorts in their analysis cannot be distinguished, so Steinbuch and Love are unable to measure any between-cohort variability in the median student. Consequently, their model assumes, implicitly, that median LSAT and UGPA do not vary across cohorts (that is, these variables are "held constant"). This is a reasonable assumption because, even if Steinbuch and Love were able to distinguish the cohorts, the median LSAT does not change considerably in the years studied. *See supra* note 383. For example, between 2012 and 2022 for UALR, the median LSAT varied between 150 and 152. Am. Bar Found., *supra* note 200. By including LGPA—which is a proxy for learning mismatch—as a control variable in their models, the impact of LSAT and UGPA may capture some of the influence of entering credentials on bar exam performance that operates independently of LGPA; however, mismatch theory posits that LGPA is a consequence—not merely a correlate—of UGPA and LSAT. *See supra* notes 45–47 and accompanying text. Steinbuch and Love's analysis does not properly measure the direct and indirect effects (operating through LGPA) of entering credentials on bar passage. *See* Sander, *Response to Ho*, *supra* note 123, at 2007 (positing that mismatch theory is amenable to "a type of analysis specifically developed to deal with situations in which one is concerned about independent variables indirectly affecting one another . . . [and] allows us to directly measure those indirect effects"). *See generally* Kidder & Lempert, *White Paper*, *supra* note 14, at 8–9 (explaining that first-time bar passage potentially reflects much more than what has been learned in law school, and "for these reasons one hears stories of high ranking graduates of top law schools who failed the bar on their first try").

Assuming, *arguendo*, that S&S's inclusion of individual LSAT/Index is a reasonable decision when assessing the overall predictive ability of their models, including those variables is unnecessary when examining whether mismatch is able to account for racial differences in bar performance. In other words, if the core claim of mismatch theory is that

Michael McDonald, "Far and away the most common way of recovering from computational problems resulting from forms of collinearity is [model] respecification. Virtually every basic and intermediate textbook on linear and nonlinear regression techniques gives this advice."[387] S&S's failure to report the correlations and variance inflation factors of the parameter estimates in the models makes it impossible to identify the actual magnitude of collinearity or understand how it may negatively impact on their findings.

The well-known harm from collinear variables is that they do not provide information different from that which is already inherent in other variables, such that many different model specifications will exist that will all generate virtually identical predictions with quite different coefficients for the individual variables.[388] Consequently, it becomes very difficult—perhaps impossible—to make credible inferences about the separate influence of certain explanatory variables on the outcome variable.[389] Collinearity is usually characterized as

racial differences in bar performance can be explained by mismatch because of credential distance from the middle student, then individual LSAT/Index is important in explaining this gap only insofar as it is used to determine mismatch distance. If a Black student and a white student have the same degree of mismatch from the median student, then mismatch theory posits that they experience roughly the same level of learning mismatch and will perform similarly on the bar exam. Indeed, Sander has stated that it is not the absolute ability of a student but the student's ability relative to her peers that determines academic success. Sander, *Systemic Analysis*, *supra* note 46, at 452.

S&S's analysis of the impact of entering credentials and mismatch also ignores the fact that these variables are themselves impacted by race. Specifically, UGPA and LSAT (as well as LGPA) are impacted by (and not merely correlated with) race. Thus, controlling for these variables without properly measuring the indirect effect of race on bar passage that operates through UGPA, LSAT, and LGPA biases the estimates of the race effect by incorrectly removing some of the effect of race (i.e., post-treatment bias). *See supra* note 181 and accompanying text. Steinbuch and Love's analysis of UALR repeats this error. *See* Steinbuch & Love, *supra* note 333. Because entering credentials and the mismatch variable are both based on the LSAT (and UGPA for Index), including both of those variables in the model will likely exacerbate post-treatment bias when measuring the impact of race.

It is also worth reemphasizing that the first-time bar passage rates for UCLA, Davis, and UALR are, respectively, 89%, 81%, and 78%. The median LSAT scores for these three schools are, respectively, 164 (ninetieth percentile), 160 (eightieth percentile), and 152 (fiftieth percentile). *See* Table 1. These statistics suggests that substantial differences in the schools' median LSAT scores do not result in sizable differences in bar passage rates—especially when comparing Davis and UALR. In other words, at the school level, there is not much evidence of a "credential effect." And this is consistent with the findings that entering credentials explain only a small proportion of the variance in LGPA, and that entering credentials and LGPA explain only a modest proportion of the variance in graduation rates and bar passage rates. *See supra* notes 174–178 and accompanying text.

387 Altman et al., *supra* note 330, at 192; *accord* Fox & Monette, *supra* note 386, at 183 ("The identification of specific sources of imprecision in estimation may, in certain instances, suggest how estimates can be improved . . . . [T]he discovery of collinearity may motivate respecification of a statistical model or reorientation of the goals of a study.").

388 Belsley et al., *supra* note 386, at 86.

389 *Id.*

a "weak or uninformative data" problem, not a statistical problem.[390] This view is only partly correct because collinearity also increases the sensitivity of parameter estimates to small errors in model misspecification,[391] and this sensitivity will be present even in large samples.[392] The reason for the increased parameter sensitivity is that the influence of model misspecification on parameter estimates is magnified when independent variables are highly correlated with one another, even when the independent variables are only slightly correlated with the error term (i.e., endogeneity).[393]

S&S neither report any of the applicable sensitivity analyses that would have been helpful in identifying potential misspecification bias,[394] nor do they explore how this bias may be exacerbated by collinearity.[395] But both model misspecification and collinearity are likely to impact S&S's analysis for several reasons: (1) the limited number of relevant explanatory variables,[396] (2) the construction of the mismatch variable as a near linear combination of other variables in the model (i.e., UGPA, LSAT, and school-specific LSAT/Index),[397] (3) the very few Black students in the data,[398] and (4) the disproportionate concentration of Black students in the lower end of the entering credential distribution relative to non-Black students.[399]

These errors have impacted Sander's prior analysis of mismatch effects. Chambers and colleagues describe how multicollinearity negatively undermined Sander's analysis of race effects in *Systemic Analysis*, noting that "law school GPA, LSAT scores and UGPA . . . are better predictors of whether someone is black or white than they are, along with race, gender, and law school tier [i.e., mismatch], of bar passage. Hence it is not surprising that

390 Christopher Winship & Bruce Western, *Multicollinearity and Model Misspecification*, 3 Sociol. Sci. 627, 628 (2016) ("Within the traditional frequentist perspective, collinearity has been analyzed as a problem of imprecise estimates or equivalently large standard errors resulting from having weak or too little data.").

391 *Id.*, at 628.

392 *Id.*

393 *Id.*; *see also supra* note 345 (noting that a censored independent variable magnifies bias from model misspecification when correlated with other independent variables in the model (called "bias transmission")).

394 *See supra* notes 294 & 295.

395 *See generally* Carl F. Mela & Praveen K. Kopalle, *The Impact of Collinearity on Regression Analysis: The Asymmetric Effect of Negative and Positive Correlations*, 34 Appl. Econ. 667, 671 (2002) (describing the effect of collinearity on model misspecification); *accord* Winship & Western, *supra* note 390, at 634 (explaining that bias from model misspecification increases nonlinearly with increasing collinearity, such that small errors in model misspecification can lead to large biases in parameter estimates).

396 *See supra* notes 321–323 and accompanying text.

397 *See supra* notes 381–385 and accompanying text.

398 *See supra* note 333 and accompanying text.

399 *See supra* note 106 and accompanying text.

when race is included in this model it has no significant effects."[400] S&S do not include a table of descriptive/summary statistics for their variables (e.g., means, standard deviations, ranges) or correlations, which is a standard feature of empirical economics papers.[401] Their failure to provide this standard information prevents the reader from quickly ascertaining some of the aforementioned problems with their data.[402]

As corrective proposals to avoid the problems of collinearity or model dependence, in prior work, I advocated utilizing data preprocessing (i.e., covariate balancing), model averaging, and model cross-validation to explore the degree of model dependence of analyses of the BPS data.[403] Data preprocessing involves "select[ing] a sample where the treatment and control groups [i.e., racial groups] are more balanced than in the original full sample [so] inferences are more robust and credible."[404] Balancing makes it plain whether or not comparable untreated observations are available for comparison with the treated observation in the data.[405] 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."[406] Cross-validation

400 Chambers et al., *supra* note 178, at 1872 n.58.

401 *See, e.g.*, Paul Dudenhefer, *Designing Tables for Economic Papers: A Handout from the EcoTeach Center*, Duke University 1 (2007) ("In economics papers, tables may be used for any number of purposes, but three are more common than others. In empirical papers there is usually a table of descriptive statistics. Also called 'summary statistics,' descriptive statistics usually give a sociodemographic profile of a sample population.").

402 For example, S&S remove a table describing racially disaggregated data from the published version of their paper, but that information was included in an earlier draft and revealed very few Black students in their sample. *See supra* Sander & Steinbuch, *2018 Draft, supra* note 333. Steinbuch and Love's article analyzing the UALR data reported racially disaggregated summary statistics for bar-takers, UGPA, LSAT, and LGPA. Steinbuch & Love, *supra* note 332, at 216 tbl.1, 218 tbl.2 & 227–28 tbls.8 & 9. Of the 723 in-state bar-takers from UALR, fifty-six (7.7%) were Black and eleven (1.5%) were Latinx, while 624 (86.3%) were white. The bar exam success rate for white students (79.8%) was slightly lower than for Latinx students (81.8%) and significantly higher than for Black students (58.9%).

403 Thaxton, *supra* note 1, at 901–02 & 973–76.

404 *Id.* at 902. I explain in prior work that "[t]he importance of employing nonparametric matching methods [i.e., balancing] 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]."). *Id.* at 905.

405 *Id.* at 908.

406 *Id.* at 973. Model averaging can be used to obtain "both weighted and unweighted average causal effects, with the weights being determined by predictive accuracy of the model (in other words, the model fit)." *Id.* 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." *Id.* at 973 n. 457.

"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 . . . thereby providing a measure of the quality of the chosen model."[407] Whereas model averaging "is best used as a robustness check to show that our inferences are not overly sensitive to plausible variations in model specifications,"[408] cross-validation is best used "to examine the extent to which the predictive capacity of the model differs across the subsets [of the data]"[409]—i.e., an assessment of the degree of overfitting. In a nutshell, covariate balancing, model averaging, and cross-validation are tools that can be helpful in identifying potential problems with a proposed statistical model.[410] If S&S had elected to deploy these tools and their results remained unchanged, we would have greater confidence in the accuracy and reliability of their conclusions.

Carefully diagnosing model dependence may be critically important when estimating fixed effects on small, sparse, and highly correlated data. S&S's use of school-specific fixed effects risks increasing model dependence through "the use of implausible counterfactuals . . . to characterize substantive effects"[411] because only within-school variation in the explanatory variables can be utilized to estimate their effects on the likelihood of bar passage. Consequently, changes in explanatory variables that are larger than any changes observed within schools may "extrapolate to regions of sparse (or nonexistent) data[,] rest on strong modeling assumptions, and can produce severely biased estimates of a variable's effect if those assumptions fail."[412] These complications are discussed in detail in the next section.

E. School-Specific Fixed Effects

Fixed effects analysis is a statistical methodology that is able to account for the effects of variables, both observed and unobserved, that do not vary within a larger grouping. "By using fixed effects, researchers make not only a methodological choice but a substantive one, narrowing the analysis to a particular dimension of the data."[413] The school-specific fixed effects (hereinafter "school fixed effects") remove the influence of any unobserved differences between schools on bar exam performance that may be correlated with the

407 *Id.* at 974–75.

408 *Id.* at 972.

409 *Id.* at 974. Three common causes of model validation failure are overfitting (i.e., where the model capitalizes on idiosyncratic features of the data rather than the underlying pattern), discretization of variables, and subject inclusion criteria. *Id.* at 906 n. 261.

410 *Id.* at 972.

411 *See generally* Jonathan Mummolo & Erik Peterson, *Improving the Interpretation of Fixed Effects Regression Results*, 6 Pol. Sci. Res. Methods 829 (2018).

412 Mummolo & Peterson, *supra* note 411, at 830.

413 *Id.* 829.

variables in the model. Accordingly, the effects of the explanatory variables in the model now relate solely to within-school variability in bar passage (as opposed to a weighted combination of the within- and between-school variability).[414] But even when the overall variation in the explanatory variables is substantial, the within-school variation used to estimate the effects of the variables may be much smaller.[415] Consequently, the within-school estimates deal only with a subsection of the variance in each variable that remains after accounting for school-level attributes, and the model is uninformative about the between-school effects or a general effect of the variables in the model.[416] "Attention to this variance reduction has the potential to greatly influence how researchers characterize the substantive significance of results."[417]

S&S claim that a "crucial strength of [their] data for the purposes of studying mismatch" is both substantial variation in students' median credentials across cohorts and schools and substantial overlap in the students' entering credentials across the different schools.[418] S&S posit that these features of the data allow them to "estimate how students at one of the schools might have performed had they attended one of the other schools, and vice versa."[419] As explained in Part III.A, S&S's analytical approach does not examine the impact of mismatch on LGPA (legal learning). Yet "for mismatch to result in worse outcomes, low credential students have to learn less than they would have at another [less competitive] school—not less than the middle-range student at their current institution."[420] If a student would have learned the same amount at either school, or if the increase in learning at the less competitive school would be insufficient to alter bar exam performance, then the student's performance on the bar exam cannot be attributed to mismatch.[421]

Assuming, *arguendo*, that mismatch dynamics can be adequately modeled without considering LGPA, S&S's school fixed effects approach utilizes only within-school variation in an explanatory variable's estimated effect. Therefore changes in the variable that are larger than any changes observed within the schools in the data, or changes that are extremely rare, lack any meaningful

414 Technically speaking, the school fixed effects parameters capture the amount by which the predictions of bar passage in the school must be adjusted upward or downward, given the explanatory variables. Andrew Gelman & Jennifer Hill, Data Analysis Using Regression and Multilevel/Hierarchical Models 226 (2007).

415 Mummolo & Peterson, *supra* note 411, at 829.

416 Andrew Bell & Kelvyn Jones, *Explaining Fixed Effects: Random Effects Modeling of Time-Series Cross-Sectional and Panel Data*, 3 Pol. Sci. Res. Methods 133, 139 (2015).

417 Mummolo & Peterson, *supra* note 411, at 830.

418 Sander & Steinbuch, *supra* note 15, at 726.

419 *Id.*

420 Barnes, *supra* note 46, at 792 n.4.

421 Rothstein & Yoon, *supra* note 166, at 659 & 678; Lempert et al., *supra* note 38, at 4 (same); Barnes, *supra* note 46, at 792 n.4 (same).

interpretation.[422] This problem also applies to one-unit shifts for categorical variables because a one-unit shift can present a change that is several times the size of a typical shift within schools.[423] Given the small number of Black students in the data at every level of mismatch and the heavily skewed distribution of Black students in the mismatched groupings relative to non-Black students,[424] the school fixed effects model specification further restricts the analysis to a subset of the variation of S&S's sparse data, thereby making inferences increasingly dependent on assumptions of their models rather than based on the available data. Bell and Jones explain that fixed effects "models that control out, rather than explicitly model, environment and heterogeneity offer overly simplistic and impoverished results that can lead to misleading interpretations."[425]

It is unclear why racial differences become statistically insignificant—as S&S report—when they include school-specific effects, given that their models are designed to adjust for degree of mismatch and LSAT/Index. That is, S&S report that the effect of race on bar exam performance disappears after school fixed effects are included (see Table 3-Model 6).[426] But according to mismatch theory, school environment should be irrelevant to the presence or absence of a race effect after accounting for school-specific mismatch effects (as well as entering credentials) for each student.[427] Recall S&S's claim that "'controlling' for race in a regression would separate out the 'race' effect from the 'mismatch' effect" when comparing different racial groups across schools

422 Mummolo & Peterson, *supra* note 411, at 830.

423 *Id.* at 833.

424 *See supra* note 333.

425 Bell et al., *supra* note 416, at 134. In prior work critiquing the use of the fixed-effect framework when examining environmental effects, I explain:

> By throwing away valuable information due to the assumptions of the fixed effects model, the investigation of case-level processes, as well as between-[cluster] variability, can be appreciably undermined. Fixed effects models require that a substantial portion of the cases within a [cluster] differ in their case-specific characteristics because the model focuses exclusively on within-jurisdiction variation. If most of the variation in case-level characteristics is between [clusters], then the fixed effects models give misleading answers to questions about the effects of the case-level characteristics. . . . Fixed effects models do not allow inferences to be made about the between-[cluster] variability in [the outcome], including whether or not the variability is substantively meaningful. All of these aforementioned problems originate from a common source: the inability of the fixed effects framework to simultaneously consider within- and between-[cluster] variability.
>
> Sherod Thaxton, *Un-*Gregg*-ulated: Capital Charging and the Missing Mandate of* Gregg v. Georgia, 11 Duke J. Const. L. & Pub. Pol'y 145, 167–68 (2016).

426 Sander & Steinbuch, *supra* note 15, at 731 tbl.3.

427 Thaxton, *supra* note 1, at 982; *accord* Rothstein & Yoon, *supra* note 166, at 679 (noting that mismatch theory assumes that schools do not matter—only students' entering credentials relative to those of their classmates).

with a similar degree of mismatch.[428] This claim about the model implies that it is unnecessary to control for the influence of unobserved school-level factors potentially correlated with race and degree of mismatch when comparing similarly mismatched students. Indeed, Sander has expressly stated that "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."[429] Yet S&S's results suggest that the ability of mismatch to account for the effect of race on bar passage depends, in part, on the school attended.

In other words, their results suggest that mismatch itself is not solely driving racial disparities in the likelihood of passing the bar exam; their analysis suggests that the disparities depend on school environment.[430] The impact of the explanatory variables in S&S's school fixed effects models relate solely to within-school dynamics and do not offer insights about the between-school effects or the general effects of the variables in the model.[431] Accordingly, in such studies, the mechanical application of fixed-effects designs without careful consideration of what is actually being measured is strongly discouraged.[432] In the words of renowned philosopher and education scholar John Dewey, "[N]eglect of context is the besetting fallacy of philosophical thought."[433] To that end, Andrew Bell and colleagues have emphasized that the grouping of individuals in particular settings "represents real and interesting societal configurations; they are not simply a technicality or consequence of survey methodology, as the population may itself be structured by social processes, distinctions, and inequalities. Structures are important in part because variables can be related at more than one level in a hierarchy, and the relationships at different levels are not necessarily equivalent."[434]

428 *See supra* note 336 and accompanying text.

429 Sander & Taylor, at *supra* note 22, at 60–61.

430 Thaxton, *supra* note 1, at 935 n.320 (describing research discovering that institutional environment has a strong influence on student outcomes, independent of entering credentials, in the fields of law, medicine, and STEM).

431 *See supra* note 416 and accompanying text.

432 *See generally* Bell et al., *supra* note 416, at 134; Mummolo & Peterson, *supra* note 411.

433 John Dewey, Context and Thought 219 (1931). According to Dewey, the philosophical fallacy is the erroneous assertion that whatever that is found true under certain conditions can be asserted universally or without limits and qualifications. *Id*.

434 Andrew Bell et al., *Fixed and Random Effects Models: Making an Informed Choice*, 53 Qual. Quant. 1051, 1053 (2019); *accord* Andrew Gelman, *Multilevel (Hierarchical) Modeling: What It Can and Cannot Do*, 48 Technometrics 432 (2006) ("One intriguing feature of multilevel models is their ability to separately estimate the predictive effects of an individual predictor and its group-level mean, which are sometimes interpreted as 'direct' and 'contextual' effects of the predictor."); *see also* Eduardo Bonilla-Silva, Racism Without Racists: Color-Blind

S&S acknowledge that their analysis "is not, of course, a randomized experiment, so student choices may produce student bodies at the various schools that are different in substantive ways we cannot ascertain."[435] But their statement blurs the important sociological distinction between properties of collectives and members.[436] Members belong to collectives, and various properties of both collectives and their members may be measured and analyzed at the same time.[437] Differential performance on the bar exam may be attributable to environmental factors that are unrelated to mismatch and irreducible to an aggregation of individual student attributes and choices.[438] Such environmental factors might include class size, faculty-to-student ratio, percentage of minority faculty, and percentage of underrepresented

Racism and the Persistence of Racial Inequality in the United States 9, 182 (2003) (defining racial structure as the "totality of the social relations and practices that reinforce white privilege," noting that the task of studying racial structures is to "uncover the particular social, economic, political, social control, and ideological mechanisms responsible for the reproduction of racial privilege in a society," and suggesting that the "race effect" tends to vary by the degree of closeness to "whiteness" of minority groups in question).

435 Sander & Steinbuch, *supra* note 15, at 718.

436 *See generally* Manski, *supra* note 7, at 30–32 (emphasizing the difference between environmental effects and correlated individual effects and explaining that these effects differ conceptually and in their policy implications but have often been confused"). An example of correlated individual effect occurs when "students in the same school tend to have similar family backgrounds and these background characteristics tend to affect achievement." *Id*. at 32.

437 Paul F. Lazarsfeld & Herbert Menzel, *On the Relation between Individual and Collective Properties*, *in* On Social Research and its Language 176 (Raymond Boudon ed., 1993).

The three core types of properties of collectives are *analytical* (i.e., aggregate information from the individual members, such as average LSAT or Index); *structural* (i.e., relational characteristics of the group, such as friendship density in a classroom); and *global* (i.e., characteristics of the collective not based on the properties of individuals, such as a school-specific policy). *Id*.; *accord* Manski, *supra* note 7, at 31.

There are four core types of properties of members: *absolute* (i.e., information that does not make reference to the collective or relationships between members, such as gender, level of education, and income), *relational* (i.e., morphological information about members, such as the number of friends an individual has at school), *comparative* (i.e., distributional information about a member, such as class rank or birth order), and *contextual* (i.e., information describing members by the collective to which they belong, such as enrollment in a school with a certain percentage of racial/ethnic minorities or residence in a densely populated census tract). Contextual properties are characteristics of collectives that are applied to members. Lazarsfeld & Menzel, *supra* note 422, at 177.

438 Thaxton, *supra* note 1, at 935; *supra* note 218; *see generally* Valerie E. Lee, *Using Hierarchical Linear Modeling to Study Social Contexts: The Case of School Effects*, 35 Educ. Psych. 125 (2000) (noting that school environment (e.g., structure, organization, and teachers' attitudes) influences student learning, achievement, aspirations, and engagement, and describing the appropriate methodologies for investigation school effects).

Ayres and Brooks challenge Sander's claim that students perform better on the bar exam at more selective schools, independent of entering credentials, because of hidden qualities that impact admission decisions at elite schools. The school-level effect can also be attributed to interactions with peers and teachers who challenge the students (i.e., structural) and to schools with greater resources and a stronger academic culture (i.e., global). Ayres & Brooks, *supra* note 65, at 1854.

minority students, just to name a few.[439] As Ian Ayres and Richard Brooks have noted, "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."[440] Relatedly, Timothy Clydesdale's analysis of the BPS data revealed that "law students who report overt race discrimination in law school have significantly lower [first-year] grades in addition to the general impact of being a minority."[441] His examination also revealed that law students exposed to greater faculty race diversity in the classroom have significantly higher cumulative grades.[442] Indeed, "Sander's own work [with the BPS data] provides indirect support for the proposition that law school racial context better explains the racial differential in law school grades than academic mismatch."[443] S&S state that they examine several types of interaction effects, including a race-mismatch interaction, and report that their results remain, essentially, the same.[444] However, S&S do not report analyzing any race-school fixed-effect interactions (i.e., a cross-level interaction),[445] although such an interaction might be highly informative about the direction, magnitude, and statistical significance of the effect of race depending on school environment.[446]

439 If the school-level factors were known, they could ostensibly be included as additional variables in the model. Gelman & Hill, *supra* note 414, at 227.

440 Ayres & Brooks, *supra* note 65, at 1829; *see also* D. James Greiner & Donald B. Rubin, *Causal Effects of Perceived Immutable Characteristics*, 93 Rev. Econ. & Stat. 775, 776 (2011) (emphasizing that the perceptions of race, and not race itself, are what really matter when examining the causal effect of race).

441 Clydesdale, *supra* note 168, at 737. The racial discrimination measure consists of a sum of eight law school arenas: general law school environment, the classroom, dealings with instructors outside class, dealings with law school administration, social interactions with classmates, academic activities with classmates outside the classroom, community where law school is located, and job recruiting. *Id*. at 722 tbl.1B. The models control for approximately two dozen relevant variables impacting first-year GPA, including prelaw school experience with race and gender discrimination, gender discrimination during law school, time spent studying, self-assessment of personal abilities, perception of quality of instruction, and socio-demographic factors.

442 *Id*., at 744. The faculty race-diversity variable captures the number of racial/ethnic minority faculty teaching the student during the first and second years of law school. *Id*. at 722 tbl.1B.

443 *See* Thaxton, *supra* note 1, at 979–80.

444 Sander & Steinbuch, *supra* note 15, at 734 n.33. S&S do not include the specific results from their analysis of interaction effects in their article, so it is unclear whether their claims are credible. On at least two prior occasions, Sander has incorrectly interpreted interaction effects and their associated measures of uncertainty, which lead to invalid conclusions about mismatch effects. *See infra* note 521.

445 *See also infra* note 526 and accompanying text.

446 *See infra* note 522 and accompanying text (explaining the effect size and associated standard error of an interaction varies across the range of the constituent variable). When individuals are clustered into groups, three types of nonmediated effects may be explored. First is a

S&S justify their inclusion of school fixed effects to account for each school's "unique learning environment." Specifically, they measure "what effect these school-wide effects have on bar passage" and adjust for the fact that students at UALR are taking a different bar exam (Arkansas) than students at UCLA and Davis (California).[447] Not only is this statement in conflict with their earlier claim in the paper that cross-race comparisons that controlled for degree of mismatch (and entering credentials) would isolate the effect of race from mismatch, but S&S's justification does not withstand scrutiny for at least four reasons that I discuss below.

First, the inclusion of school fixed effects does not remove the race effect when mismatch is modeled linearly *and* the inclusion of the nonlinear mismatch variable effect does not remove the race effect when school fixed effects are not included. The race effect becomes statistically insignificant only when both school fixed effects and the nonlinear mismatch variable are both included in the model. But, as noted earlier, the model fit (Somers' *D*) of the nonlinear specification with school fixed effects (Table 3-Model 6) is indistinguishable from the (a) linear specification with school fixed effects effects (Table 3-Model 4) and (b) the nonlinear specification without school fixed effects (Table 3-Model 5). Moreover, if S&S were to account for model complexity (i.e., adding more variables to the model), Table 3-Model 6 would likely provide the poorest model fit. It is also illuminating that the model with a linear specification and no school fixed effects (Table 3-Model 3) would likely fit the data equally well, if not better, when taking into model complexity and reporting a confidence band for Somers' *D*.[448] Political scientist Christopher Achen appropriately noted that "adding control variables to a policy outcome regression is most dangerous when these additional factors

lower-level direct effect, which examines the impact of a lower-level explanatory variable on a lower-level outcome (e.g., impact of race on bar exam performance). Second is a cross-level direct effect, which measures the impact of a higher-level variable on a lower-level outcome (e.g., impact of school environment on bar exam performance). Finally, a cross-level interaction measures whether the nature or strength of the relationship between two lower-level variables (e.g., race and bar exam performance) changes as a function of a higher-level variable (e.g., the impact of race on bar performance is stronger/weaker in some schools). Herman Aguinis et al., *Best-Practice Recommendations for Estimating Cross-Level Interaction Effects Using Multilevel Modeling*, 39 J. Mgm't 1490, 1493 (2013).

447 Sander & Steinbuch, *supra* note 15, at 733. S&S's description of their fixed-effects analysis implies the need for a more complex hierarchical model because of various levels of nested clustering—i.e., students clustered into schools and schools clustered into jurisdictions. Without explicitly modeling this multilevel nesting, the interpretation of S&S's school fixed effect is unclear in Tables 3 and 4 because there is a blurring of school environmental effects with jurisdictional effects. Generally speaking, the fixed-effects approach does not extend well to other clustering structures. Bell et al., *supra* note 434, at 1053. Table A largely avoids this problem by limiting the analysis to UCLA and Davis.

448 *See supra* note 311 and accompanying text.

increase the multicollinearity substantially without materially improving the specification."[449]

Second, the effect sizes of the coefficients for "Black" and "Latinx" across Table 3–Models 4, 5, and 6 are nearly identical, suggesting that the inclusion of both the nonlinear mismatch variable and school fixed effects has no impact on the magnitude of the effect of race, only the uncertainty (i.e., variance) of the race parameters.[450] This suggests that the alleged suppression of the race effect, via inclusion of the nonlinear mismatch variable and school fixed effects, is a consequence of insufficient within-school variation across the included explanatory variables (see previous discussion of model dependence in Part III.D) because school fixed effects remove all of the between-school variation. Thus, one may be unable to properly disentangle (isolate) the race effect from the effects of the other variables in the model.[451] This inability to islolate the effects of one variable over another in S&S's chosen model reinforces the importance of conducting the appropriate model diagnostics for sparse and collinear data.[452]

Third, S&S's robustness analysis,[453] which focuses exclusively on the two schools from California, inherently "controls" for bar exam difficulty and reveals that the effect of race remains statistically significant across all of their model specifications, including the model that contains the nonlinear mismatch variable and school fixed effect (Table A–Model 6).[454] The Somers' *D* statistics for Table A–Model 4 (linear effect and school fixed effect) and Table A–Model 6 (nonlinear effect and school fixed effect) are identical (0.33), and

449 Christopher H. Achen, The Statistical Analysis of Quasi-Experiments 28 (1986).

450 Whereas statistical significance examines whether the findings are likely due to chance, practical significance (i.e., effect size) relates to the magnitude of the differences between groups. Gail M. Sullivan & Richard Feinn, *Using Effect Size—or Why the P-value is not Enough*, 4 J. Grad Med Educ. 279 (2012) ("Statistical significance is the least interesting thing about the results. You should describe the results in terms of measures of magnitude—not just, does a treatment affect people, but how much does it affect them.") (internal citation omitted); *accord* Weisbach, *supra* note 8, at 113 ("A common mistake that authors make when interpreting results is focusing too much on statistical significance and not enough on the magnitudes of their estimates.").

As noted earlier (*see supra* Part III.B), eyeballing coefficient comparisons across nested logit models is discouraged in the empirical literature. I raise the issue of the consistency of the race effects across models to highlight the fact that, even when adopting S&S's approach to comparing coefficients across models, their results do not appear to support their core claim.

451 *See supra* note 386 and *infra* note 508 (explaining that high collinearity between the effects of race-related covariates and race may lead analysts to incorrectly conclude that no race effect exists).

452 *See supra* note 330 and accompanying text and *infra* note 532 and accompanying text. *See also* Chambers et al., *supra* note 178, at 1872 ("Numerous other statistical problems can be found in Sander's analysis. These include . . . models plagued by multicollinearity.").

453 *See infra* Part III.F.

454 Sander & Steinbuch, *supra* note 15, at 742 tbl.A.

goodness of fit would likely be lower for Table A-Model 6 if S&S accounted for model complexity. The Somers' *D* for Table A-Model 5 (nonlinear effect and no school fixed effect) reveals a better fitting model (0.37) than Table A-Model 6. It is also the case that the effect sizes for "Black" and "Latinx" are nearly identical across Table A-Models 3-6.[455] In other words, adjusting for LSAT, degree of mismatch, and school fixed effects does not appear to alter the relationship between race and bar passage for the two schools in California.

Finally, statistically significant school fixed effects across several model specifications potentially undermine S&S's claim that environment matters only to the extent that students are academically outmatched relative to the middle student at their school.[456] S&S acknowledge that students at UCLA (the most elite school) have the highest likelihood of passing the bar, independent of entering credentials and degree of mismatch,[457] and they suggest that this result may be attributable to unobservable differences in student quality.[458] While this is a plausible interpretation, it raises at least three concerns. First, the unobserved student qualities must be observable to the admissions committee members, and those committee members must systematically select on these features, independent of UGPA and LSAT. Second, a student who is rated higher on unobserved quality metrics than another student with an identical UGPA and LSAT may be more likely to *attend* a lower-ranked law school (e.g., scholarship awards may be based on additional information available to the admissions officer),[459] so it is not obvious that unobserved student quality differences at elite law schools account for the observed differences in bar passage. Third, as I have explained earlier, S&S's statement blurs the sociological distinction between properties of collectives and properties of members of collectives, [460] and "differential performance in bar passage may be attributable to environmental factors that are unrelated to mismatch and irreducible to an aggregation of individual student attributes and choices."[461] The finding of a positive school-level effect on bar exam performance, for

455 *But see supra* note 450.

456 This final critique of S&S's fixed-effects specifications is distinct from the prior concerns about the effect of race on bar exam performance "disappearing" only when school fixed effects are included in the model. As noted earlier, Sander argues that, after accounting for school-specific mismatch effects (as well as entering credentials) for each student, school environment is irrelevant to the presence or absence of a race effect.

457 Sander & Steinbuch, *supra* note 15, at 728.

458 *Id.*

459 Kidder & Lempert, *White Paper*, *supra* note 14; *accord* Dillon & Smith, *supra* note 110, at 46 & 63 (explaining that "undermatching" occurs when someone is a relatively strong student at one's college and "find[ing] evidence that suggests that financial constraints lead some students to undermatch").

460 *See supra* notes 436-439 and accompanying text.

461 *See supra* note 307 and accompanying text.

more competitive schools, after accounting for entering credentials, has been highlighted by scholars who examined the BPS data.[462]

In Table 3-Models 4 and 6, the coefficients for Davis and UALR reveal that students who are identical with respect to race, LSAT, and degree of mismatch (modeled either linearly or nonlinearly) are significantly less likely to pass the bar than students from UCLA. In Table 4-Model 9 and Model 10, which excludes UCLA and replaces LSAT with Index (for Model 10), but otherwise includes the same variables from Table 3-Models 4 and 6, students from Davis are less likely to pass the bar than students from UALR, albeit the effects are statistically insignificant. Finally, in Table A-Models 4 and 6, which compares the two California law schools (UCLA and Davis) and uses LSAT instead of Index, but otherwise includes the other variables, students from Davis are significantly less likely to pass the bar than students from UCLA. Focusing solely on the statistically significant effects from Table 3-Model 6 and Table A-Model 6, one would conclude that students who graduated from less selective law schools (i.e., law schools with lower relative median LSAT) are more likely to fail the bar exam, even when accounting for race, entering credentials (LSAT), and degree of mismatch.

These findings are inconsistent with mismatch theory *unless* the school-level fixed effect proxies an unobserved mismatch-related environmental factor that both (a) varied across these schools and (b) negatively impacted the performance of otherwise similarly situated students in less competitive schools.[463] Stated differently, one would not expect to observe a school-level effect when comparing students who are similar with respect to both their ranking in the overall distribution in entering credentials (i.e., LSAT/Index) and in the school-specific distribution of entering credentials (i.e., mismatch). The school fixed effects are statistically insignificant in Table 4-Models 9 and 10, which use Index rather than LSAT, and this might, theoretically, impact the results.[464] But as I explain below,[465] it is questionable whether an Index-based measure of mismatch is any better than a LSAT-based mismatch measure at identifying mismatched students. The analysis in Table 4-Models 9 and 10 should also be viewed with caution given that the comparisons are based on students taking the bar exam in different jurisdictions. A similar criticism applies to Table 3, which combines all three schools, so the comparisons between UCLA and Davis are more credible than those between UCLA and UALR (in Table 3).[466]

462 Thaxton, *supra* note 1, at 840.

463 *See supra* note 447 (discussing concerns over interpreting what S&S's fixed effects capture because of the complex clustering structure of their data).

464 The fixed-effect coefficient in Table 4 is in the same direction as the school fixed effects in Table 3 and Table A.

465 *See infra* Part III.F.

466 The California bar exam reports a lower first-time taker passage rate and a higher passing score requirement for the Multistate Bar Exam (MBE) than the Arkansas bar exam. For the

Sander has argued that any advantage of attending a more competitive school with respect to bar exam performance pales in comparison to the disadvantage resulting from obtaining a lower LGPA (because of mismatch) and, consequently, learning less in law school; however, Sander's contention has been challenged on both theoretical and methodological grounds.[467] But assuming, *arguendo*, that there is some merit to Sander's claim, S&S's decision to not model the total impact of the law school fixed effect—comprising a direct effect on bar passage and an indirect effect on bar passage that operates through LGPA—as Sander has done in prior analyses[468]—prevents S&S from providing supportive evidence of their claim with the school-specific data. Absent a total school-effect analysis, it may be informative to examine the relative impacts of school fixed effects and mismatch effects on bar exam performance. If the school fixed effects are similar in magnitude to the mismatch effects, or greater, then S&S's contention that observable credential differences are primarily responsible for the gap in bar passage must be called into question. Indeed, S&S's critique of race-conscious affirmative action is that Black students "receive (on average) the largest preferences"[469] and "minority students did worse [on the bar exam] because, on average, they entered law school with lower academic credentials."[470] When unobservable student qualities substantially influence admissions decisions, learning in law school, and the likelihood of bar passage—independent of UGPA and LSAT—it cannot be said that students are mismatched simply based on their entering numeric credentials.[471] It is more likely that the school fixed effects reflect environmental factors other than an aggregation of mismatch-related individual student attributes.[472]

July 2023 administration of the bar exam, the bar passage rates for first-time exam-takers in Arkansas and California were, respectively, 81% and 65%. The difference in passage rate for first-time exam-takers for the July 2022 exam was similar (75% vs. 62%). Nat'l Conf. Bar Exam., Bar Exam Results by Jurisdiction: July 2023 Bar Exam (Nov. 2023), www.ncbex.org/statistics-research/bar-exam-results-jurisdiction. A passing score on the MBE for California was 139, compared with 135 for Arkansas. California does not use the Universal Bar Exam (UBE), so only MBE comparisons are possible.

467 *See supra* note 66.

468 *See* Sander, *Mismeasuring Mismatch*, *supra* note 123, at 2007–2008 fig.1 (employing 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.").

469 Sander & Steinbuch, *supra* note 15, at 716.

470 *Id.* at 721.

471 *See supra* notes 108–113 and accompanying text (noting that Sander has acknowledged that admissions committees evaluate applicants based on factors other than UGPA and LSAT).

472 Thaxton, *supra* note 1, at 935 (noting that differential bar exam performance by students with similar Index scores who attended schools with similar median UGPA and LSAT is likely "attributable to lower tuition costs, smaller enrollments, far better student-faculty ratios, or some other factors unrelated to mismatch").

S&S's own data suggest that the law school environment exerts an impact on bar exam performance that is as strong as, if not stronger than, the first three levels of mismatch. Given the dichotomous nature of school fixed-effect variable and the discretized nonlinear mismatch variable, a direct comparison can be made (using the LSAT-based mismatch measure) in Table 3–Model 6 and Table A–Model 6.[473] S&S report odds ratios, and when examining odds ratios, values closer to one indicate a weaker effect.[474] If a value is greater than one, the effect of the variable on the outcome is positive (i.e., it increases the odds of observing the outcome). In contrast, if the value is less than one, the effect of the variable on the outcome is negative (i.e., it decreases the odds of observing the outcome). For Table 3–Model 6, which includes all three law schools, the school fixed effect for Davis (0.44) is stronger than the impact of the first three levels of mismatch (0.90, 0.74, 0.51). For UALR, the school effect (0.51) is stronger than the mismatch effect for the first two levels and identical to the third mismatch level. For Table A–Model 6, which includes only UCLA and Davis, the school fixed effect (0.41) is stronger than the mismatch effect for the first two levels (0.82, 0.58) and just slightly weaker than the third level (0.39). Again, these results strongly imply that law school environment exerts an impact that rivals the first three levels of mismatch.

In summary, S&S's analysis contradicts their core claim that attending lower-ranked schools improve the likelihood of higher bar performance for Black and Latinx students than attending higher-ranked schools where there is greater mismatch with their white peers. Recall that, according to mismatch theory, the less difficult school will better prepare students for the bar exam because their level of mismatch, compared with their classmates, will be less.[475] If the school fixed effect is either stronger than or the same as the mismatch effect for the first three mismatch levels, this implies that a student is better off attending UCLA than Davis or UALR if their degree of mismatch at Davis or UALR would *decrease* by an amount equal to or less than the first three levels.[476] But because these school fixed effects reveal that attending Davis

473 The coefficient comparisons are *within*-model, thereby avoiding the problem of comparing dissimilarly scaled coefficients that otherwise impact *between*-model comparisons for logit regression. But the problem associated with eyeballing *within*-model differences between coefficients remains. *See supra* Part III.B. I discuss coefficient comparisons of the mismatch and school fixed-effects coefficients to highlight that even when adopting S&S's approach to comparing coefficients across models, the results do not appear to support their core claim.

474 Sander & Steinbuch, *supra* note 15, at 730 ("A coefficient of '1' means that, in the given equation, a one-unit change in the independent variable has a neutral effect—it makes a positive outcome neither more nor less likely. A coefficient below '1' implies a negative effect, and a coefficient greater than '1' implies a positive effect."); *accord* Hosmer & Lemeshow, *supra* note 305, at 51–52 (an odds ratio of one "1" indicates no effect).

475 *See* Part I.B.

476 Lempert et al., *supra* note 38, at 4 (explaining that a mismatch effect "would happen only if the weaker school in some way better prepared the student to pass the bar, and the evidence does not indicate that this regularly occurs"); *accord* Barnes, *supra* note 46, at 792 n.4 (stating that for "mismatch to result in worse outcomes, low credential students have to learn less

and UALR negatively impacts students' bar exam performance, these results refute S&S's claim that attending a weaker school necessarily improves a student's likelihood of passing the bar exam. Also recall that the only credible counterfactuals entail comparing UCLA and Davis, given that these two sets of students take the same bar exam.[477] So this finding is especially relevant because the vast majority of mismatched students at UCLA would only modestly decrease their mismatch levels had they attended Davis.[478]

F. Robustness Analysis

The number and variety of possible robustness tests is large, and the relevance of these tests will depend the specificities of the type of model uncertainty, the intended inferences, and the data structure.[479] Analysts must justify their choice of robustness tests because it may be tempting for them to report tests "not because they really test the robustness of the estimates in the presence of model uncertainty, but simply because their baseline model proves to be robust to the carefully selected tests."[480] Moreover, "there remains widespread concern that even the robustness checks that are presented could be cherry-picked from a large set of possible alternatives."[481]

As I explain below, S&S's robustness analysis is of dubious value because they provide insufficient justification for including some tests and omitting others.[482] I noted earlier that carefully diagnosing model dependence (robustness checks) is critically important when analyzing the plausibility of

than they would have at another school—not less than the middle-range student at their current institution").

477 *See supra* note 101 and accompanying text.

478 *See* Tables 2 & 6. Also recall that credible counterfactuals from fixed-effect models must be based on within-school variation in mismatch effects, which is always smaller than the overall variation in the independent variable. *See supra* note 415 and accompanying text. Table 2 reveals that both the proportion of mismatched students and the concentration of students in particular mismatch levels varies considerably across schools. For example, only 18% of UCLA students have LSAT scores below the median LSAT category (162–164), whereas 34% of Davis students and 27% of UALR students have LSAT scores below the median LSAT category (respectively, 159–161 and 150–152). *Note*: The in-state exam-takers for UCLA in Table 2 sum to 631, but S&S report that there are 752 UCLA in-state bar-takers in Table 1. *See supra* note 332 and accompanying text.

479 Neumayer & Plümper, *supra* note 210, at 52.

480 *Id.*, at 26. S&S's lack of justification for their choice of robustness tests is present throughout their analysis. *See* Parts III.D & III.E.

481 Garret Christensen et al., Transparent and Reproducible Social Science Research: How to Do Open Science 120 (2019).

482 Neumayer & Plümper, *supra* note 210, at 30 (noting that it is "possible to distinguish between important and irrelevant robustness tests, [even though] it is not possible to conduct all relevant robustness tests"); *accord* Edward E. Leamer, *Let's Take the Con out of Econometrics*, 73 Am. Econ. Rev. 31, 43 (1983) ("If we are to make effective use of our scarce data resource, it is therefore important that we study fragility [i.e., lack of robustness] in a much more systematic way.").

model assumptions; therefore, covariate balancing, model averaging, cross-validation, and examination of model residuals are tools that can be helpful in identifying potential problems with a proposed statistical model.[483] S&S examine the (in)stability of their results when changing the measurement of one of their key variables and limiting their analysis to a subset of their data, which are commonly reported robustness tests;[484] however, S&S leave several key questions unaddressed when conducting these tests (and not others) that are central to their claim that the estimates from their core models are trustworthy.[485]

S&S's analysis of student performance on the California bar exam yields results that are at odds with some of their analysis that were presented earlier in their article. In Table A, S&S focus solely on the two California schools (UCLA and Davis) to account for possible differences based on jurisdiction (i.e., difficulty of the bar exam), given that UALR is located in a different jurisdiction.[486] S&S conclude that these results "parallel in all important respects the results in Table 3 . . . [and] are robust to a smaller, within-state analysis."[487] But upon closer inspection, these results contradict one of S&S's core assertions.[488] Specifically, they report that Black and Latinx students perform worse on the bar exam (relative to non-Hispanic white and Asian American students) even after controlling for LSAT and mismatch (both linearly and nonlinearly). In fact, the only model that eliminates the race effect for Latinx students is the model including race, LSAT, nonlinear mismatch, and a school fixed effect. Across all models, Black students are less likely to pass the bar. Equally important is that the magnitude of the coefficient for Black students is very similar across models that adjust for LSAT, irrespective of whether or not the model includes the mismatch variable. The same is true for Latinx students. In other words, mismatch neither fully accounts for nor substantially mitigates racial differences in bar passage rates.[489]

In these analyses, S&S's inclusion of the mismatch variable does not increase the predictive accuracy of the model. The nearly identical Somers' *D* statistics for models with and without a statistical control for mismatch underscore that fact (compare Table A–Model 2 with Table A–Models 3, 4, 5 and 6).

483 *See supra* notes 294 & 403–410 and accompanying text.

484 Neumayer & Plümper, *supra* note 210, at 26.

485 *See, e.g.*, *id.* at 230 ("Often . . . robustness tests are conducted with the intention of demonstrating the robustness of findings. The desired outcome—robustness—drives the robustness testing. The problem—model uncertainty—remains largely ignored.").

486 Sander & Steinbuch, *supra* note 15, at 742.

487 *Id.* at 734.

488 *See generally* Leamer, *supra* note 83, at 308 ("We must insist that all empirical studies offer convincing evidence of inferential sturdiness.").

489 *See supra* note 450 and accompanying text.

Considering the inconsistency of the results for the Black and Latinx students when school fixed effects are included or excluded across their sixteen different model specifications, it would have been helpful if S&S had conducted a disaggregated analysis on each school. Not only would this analysis have eliminated the need to include school fixed effects, but it would have also clearly illuminated in which school(s), if any, the race effect persists as well as shed light on the relative sizes of those race effects after taking into account LSAT/Index and level of mismatch.

S&S posit that the models including Index (i.e., a weighted composite of UGPA and LSAT), rather than LSAT, are more appropriate when accounting for racial differences. They claim that Index better captures unobservables relevant to learning in law school and bar success.[490] The explanation is plausible, but S&S do not provide any evidence that Black students would score systematically lower on these unobservables, holding level of mismatch constant, that would explain why Index (rather than LSAT) is the proper measure of entering credentials.[491] In other words, if I understand S&S correctly, the implication is that Black students, on balance, have lower UGPAs than white and Asian American students with the same LSAT and level of mismatch,[492] and therefore Blacks are actually *more* mismatched (relative to whites and Asian Americans) than they appear to be when considering only LSAT. Accordingly, controlling for Index (instead of LSAT alone) is the

490 Sander & Steinbuch, *supra* note 15, at 737–38, tbl.5. As noted earlier, *supra* note 107, Sander attributes Black students' enrollment into more elite law schools compared with similarly credentialed white students to racial preferences but attributes white students' enrollment in more elite law schools compared with similarly credentialed white students to unobserved quality differences between students. S&S's contention about the superiority of an Index-over-LSAT-based measure of mismatch is somewhat in tension with their claim that the "credential effect [is] driven particularly by the LSAT [and] may explain as much as half of the black-white [sic] bar passage gap." Sander & Steinbuch, *supra* note 15, at 729.

491 Rothstein and Yoon explain that, among Black and white students with the same observed credentials, one would expect Black students with better unobserved qualifications to earn higher grades in law school, but the opposite was true when analyzing the BPS data. Rothstein & Yoon, *supra* note 166, at 670 n.81 (citing Wightman, Beyond FYA, *supra* note 168). Furthermore, as Ho explains, "Unless the researcher gathers more data . . . we have no idea what impact unobservables have on the estimates," and the impact of unobservables on bar passage "depends on the correlation between unobserved and observed variables, and as such cannot be tested in the data"). Ho, *supra* note 28, at 2015.

492 This point merits elaboration. S&S provide a table that shows graduates from Davis have higher UGPAs, on average, than graduates from UALR, even when the graduates from both schools have the same LSAT scores. Sander & Steinbuch, *supra* note 15, at 737 tbl.5. S&S claim that UGPA is a proxy for unobserved credentials that are captured inadequately by LSAT. Thus, the measurement error of aptitude for legal learning will increase, among students who are similarly situated with respect to LSAT score, as UGPA increases. For Index to better account for Black students' lower bar passage rates as compared to white and Asian American students, a necessary, but insufficient, condition is that being Black and UGPA must be negatively (and meaningfully) correlated, holding LSAT and level of mismatch constant.

correct method to compare students to account for these racial differences in unobservables.[493] This claim raises six concerns that I discuss below.

First, the model fit statistics are nearly identical for the LSAT and Index models, increasing by a mere 0.01 for the models using Index.[494] S&S claim that this is a relevant increase in explanatory power, but as I explained earlier, Somers' *D* is not a useful measure to distinguish between models, *and* S&S should report a confidence interval for the statistic where, as here, they choose to use it to the exclusion of other statistics.[495] Consequently, the two statistics should be practically indistinguishable. Psychometricians Stephen Klein and Roger Bolus discovered that there is very minimal association between UGPA and LGPA, with UGPAs explaining about 4% of the variance in the students' LGPAs and only 1% of the variance in their bar exam scores.[496] Moreover, among a school's first-time bar-takers, the combination of LSAT scores and UGPAs is not much better than LSAT alone in predicting bar exam scores.[497] The lack of any meaningful improvement in model fit using Index instead of LSAT undermines S&S's claim that Index provides a better proxy for mismatch-relative unobservables.

Second, S&S's use of Index is peculiar: They are not attempting to predict the admissions decisions of law schools because the entire sample consists of admitted students. Rather, S&S are using UGPA and LSAT to predict the likelihood of success on the bar exam.[498] Given that the formula used to create Index is arbitrary for the purposes of S&S's ultimate inquiry (i.e., measuring the impact of unobservables), it would make more sense for S&S to use either a parametric or nonparametric interaction with UGPA and LSAT as component variables to best capture the manner in which these variables jointly influence bar passage rates.[499] The construction of Index, which "summarizes the

493 *But see* Ho, *supra* note 323, at 10 (controlling for 180 potentially confounding variables in the BPS data, including UGPA and LSAT, and concluding that "there is no evidence [of a mismatch effect] on bar passage rates").

494 Somers' *D* is 0.38 for the LSAT-based model (Table 4-Model 9) and 0.39 for the Index-based model (Table 4-Model 10).

495 *See supra* note 311 and accompanying text.

496 Klein & Bolus, *supra* note 176, at 15 n.1 (examining ten years of data from four ABA-accredited schools: two public and two private). In the BPS, UGPA explains 3.2% of the variance in students' LGPAs (for all schools) and ranges from 2.5% to 7.4% across the six tiers.

497 *Id.*

498 Sander, *Systemic Analysis*, *supra* note 46, at 392 (distinguishing between using Index for predicting admissions decisions (i.e., "How heavily are they, in reality, relied upon by admissions officers?") and predicting law-related outcomes, such as graduation, bar passage, and employment (i.e., "How useful are they? How important should they be in admissions?").

499 *See* Sherod Thaxton & Robert Agnew, *When Criminal Coping is Likely: An Examination of Conditioning Effects in General Strain Theory*, 34 J. Quant. Criminology 887, 903–05 (2018) (discussing parametric and nonparametric interaction effects in linear and nonlinear models); *see also supra* note 354 and accompanying text (noting that S&S should have used a nonparametric

academic 'numbers' of an applicant in a single number,"[500] is based on S&S's a priori weighting of UGPA and LSAT, and Sander has acknowledged that the weights chosen for UGPA and LSAT are based on an approximation of what law schools do.[501] The impact of Index on bar passage will depend, in part, on the weight given to each component, yet S&S do not report any sensitivity analysis of the race effect based on selecting different weights for Index.[502] Including UGPA and LSAT in the model, as well as testing for parametric or nonparametric interactions and selecting the model that best fits the data, would be most consistent with S&S's stated goal of capturing the influence of unobservables that are proxied by the joint effects of UGPA and LSAT.

Third, S&S do not mention the correlation between LSAT and Index in their paper but based on their description of the Index formula,[503] the correlation should be similar to what is reported in the BPS.[504] According to

---

smoother to avoid unverified functional form assumptions).

500 Sander, *Systemic Analysis*, *supra* note 46, at 393.

501 *Id.*, at 393 n.106.

502 *See* Thaxton, *supra* note 1, at 905 n.258 (noting Sander's inconsistency in functional form assumptions across his models analyzing mismatch and stating, "Perhaps these omissions are justified, but in most cases, Sander provides insufficient information to assist the reader in making this determination.").

According to S&S, "If LSAT scores run from 120 to 180 and UPGAs run from 0 to 4.0, then the 'academic index' we use below is calculated as (10×(LSAT-120)) + (100×UGPA), which scales student credentials from 0 to 1000." Sander & Steinbuch, *supra* note 15, at 719. Although it may not be obvious, this formula fixes the contributions from UGPA and LSAT to 60% for LSAT and to 40% for UGPA by allowing a maximum of 600 points for LSAT and 400 for UGPA. In prior work, Sander has defined Index in a manner that makes the weights assigned to LSAT and UGPA more transparent: "*Academic Index = 0.4× (UGPA) + 0.6×(LSAT)*, with both UGPA and LSAT normalized to a one-thousand-point scale." Sander, *Systemic Analysis*, *supra* note 46, at 393 (emphasis in original). The formula that both normalizes to 1000 points and accommodates different weighting schemes is: [((LSAT-120) ÷ (180-120))×1000]×[W_{LSAT}] + [((UGPA-0.0) ÷ (4.0-0.0))×1000]× [W_{UGPA}], where W is the assigned weight and (W_{LSAT}+ W_{UGPA} = 1). Consequently, the (dis)similarity between students based on Index can be magnified or minimized based on the values of W_{LSAT} *and* W_{UGPA} chosen by the analyst, although the students' UGPA and LSAT scores are constant. The importance of W becomes obvious with a simple demonstration using UGPA and LSAT data from UCLA's 2023 entering class: the seventy-fifth/fiftieth/twenty-fifth percentiles of UGPA = 3.98/3.92/3.72 and LSAT = 171/170/165. So fixing UGPA at the fiftieth percentile and moving LSAT from the seventy-fifth to the twenty-fifth percentile (called the interquartile range) yields a sixty-point drop in Index for the formula S&S adopt. Using an alternative weighting of 80% for LSAT and 20% for UGPA results in an eighty-point drop in Index.

503 *See supra* note 502. My comparisons of the relationship between LSAT and Index based on BPS and S&S's school-specific data assume that the identical formula was used–i.e., where LSAT and UPGA comprise, respectively, 60% and 40% of the Index.

504 The BPS data use a different scale for the LSAT score (10–48), but the LSAT normalized score range is from 0–600, as is the normalized range for the 120–180 LSAT scale. For both the BPS and the school specific-analysis, UGPA normalized score range is 0–400 and the total Index range is 0–1000. INDEX = [((LSAT-10) ÷ (48-10))×1000]×[W_{LSAT}] + [((UGPA-0.0) ÷ (4.0-0.0))×1000]× [W_{UGPA}].

the BPS data, the correlation between the two measures ranges between 0.79 and 0.90, depending on the law school tier, which means that from 65% to 81% of the variation in Index is shared with LSAT. In the NSLSP data, the correlation between LSAT and Index ranged from 0.78 to 0.87, depending on the specific school; this means that 62% to 76% of the variation in Index is shared with LSAT.[505] Sander and colleagues collected data from twenty-five public law schools from the 2005–2007 admissions cycles that contained, inter alia, information on UGPA and LSAT of law school applicants.[506] These data reveal that the correlation between LSAT and Index varied from 0.84 to 0.95 across schools, and therefore between 71% and 90% of the variation in Index is shared with LSAT. It seems implausible, then, that such a small residual difference between the two measures would have such a strong impact on the race coefficient (i.e., from a statistically significant negative effect to a statistically insignificant positive effect) in S&S's analysis.

In fact, S&S's results likely reveal a collinearity problem. A symptom of high collinearity is large confidence intervals that include unreasonable (i.e., nonsensical) parameter estimates.[507] For example, Sen and Wasow note that a high correlation between the effects of race-related covariates and race "may result in small changes having a large impact on the coefficient estimates—thus, standard errors may be large and lead researchers to assume no treatment [race] effects when treatment effects in fact exist."[508] In addition to inflated parameter variances (i.e., large confidence intervals), the negative correlation

505 Recall that UGPA and LSAT are standardized within law schools for the NSLSP data. *See supra* note 154. The formula used to calculate Index is similar to the formula used for the BPS—that is, 60% of the weight was attributed to LSAT and 40% attributed to UGPA.

506 Project SEAPHE, *Law School FOIA Dataset 1.1* (2007), http://www.seaphe.org/databases.php (lasted visited Mar. 20, 2022) (data available from author upon request). Sander was the principal investigator of Project SEAPHE, an acronym for Scale and Effect of Admissions Preferences in Higher Education, a consortium examining the effects of race-conscious affirmative action. As a part of that Project, Sander and colleagues submitted Freedom of Information Act (FOIA) requests to public law schools that were subject to state-level public records requirements.

507 Winship & Western, *supra* note 390, at 628; *accord* Mela & Kopalle, *supra* note 395, at 667 ("Various econometric references have indicated that collinearity increases estimates of parameter variance . . . and results in parameters with incorrect signs and implausible magnitudes."); Hosmer & Lemeshow, *supra* note 305, at 150 ("In general, the numerical problems of . . . collinearity are manifested by extraordinarily large estimated standard errors and sometimes by a large estimated coefficient as well.").

508 Sen & Wasow, *supra* note 129, at 505; *accord* Daniel E. Ho & Donald B. Rubin, *Credible Causal Inference for Empirical Legal Studies*, 7 Ann. Rev. L. & Soc. Sci. 17, 26 (2011) ("When groups differ sharply, regression may not credibly 'control' for confounding factors."); *see also* Johnston et al., *supra* note 234, at 1959 (explaining that high collinearity may impact the magnitude, direction, and standard error of regression coefficients); Harrell, *supra* note 304, at 193 (explaining that logistic regression estimates are especially susceptible to Type II Errors (i.e., false negatives) based on the Wald test—which S&S report—when effects in the model become very strong); *see supra note* 330 (defining high collinearity and its consequences on regression coefficients).

between race (for Black and Latinx students) and LSAT/Index may increase bias in parameter estimates because "negative correlations among independent variables have a much greater impact on omission bias [i.e., model misspecification] than equivalent positive correlations."[509] And as noted earlier, the fact that the Somers' *D* statistics for the LSAT and Index models are nearly identical underscores that the two measures are indistinguishable with respect to their ability to properly classify bar passage/failure.[510]

Fourth, S&S do not report any results for models that exclude student-specific Index to test the racial disparity hypothesis. But, as I noted earlier, entering credentials should be unnecessary to properly conduct the test.[511] Across racial groups, similarly mismatched students (based on deviations from the school-specific average Index) should be similarly disadvantaged, relative to the typical student, irrespective of their specific Index score, because the professor teaches to the "middle" student.[512] According to Sander, "[I]t was not the absolute ability of a student that determined [success], but rather the student's ability *relative to his or her peers*."[513] Again, the model fit statistics for the LSAT and Index models are indistinguishable. Therefore, it is unlikely that the Index-based mismatch variable captures race-correlated unobservables better than the LSAT-based mismatch variable. Indeed, limitations resulting from analyzing weak or uninformative data are more likely to distort the measurement of the influence of race.[514]

Fifth, S&S do not provide information about the relative effects of race on LGPA for models that include LSAT versus Index. If S&S are correct in their assertion that Index does a better job of capturing the impact of unobservables than LSAT, and as a result, Index is the more appropriate measure to account for racial disparities in learning mismatch, then one would expect the race effect (on LGPA) to be significantly smaller in the Index model than in the LSAT model (and by extension, the race variable would have less explanatory power in the Index model than in the LSAT model). The BPS data provide evidence to the contrary: The effect sizes for race (Black and Latinx) are

509 Mela & Kopalle, *supra* note 395, at 675 (demonstrating that there is an asymmetry in omitted variable bias based on the underlying correlation structure); *see also supra* notes 388–399 and accompanying text.

510 Sander & Steinbuch, *supra* note 15, at 735 tbl.4.

511 *See supra* note 386 (noting that, "if the core claim of mismatch theory is that racial differences in bar performance can be explained by mismatch because of credential 'distance' from the middle student, then individual LSAT/Index is important in explaining this gap only insofar as it is used to determine mismatch distance").

512 To be sure, S&S's assertion that law professors tailor their instruction to the "middle student" has not gone unchallenged. *See supra* note 46 (identifying three reasons to question the "middle student" assumption).

513 Sander, *Systemic Analysis*, *supra* note 46, at 452 (emphasis in original).

514 *See* Parts III.D & III.E; *see also* Thaxton, *supra* note 1, at 894 (explaining that the inclusion of more explanatory variables in a model can increase model dependence because stronger parametric assumptions are needed to properly estimate their effects).

nearly indistinguishable across the two models, suggesting that Index is no better at explaining racial differences in learning mismatch than LSAT.[515] The apparent inability of UGPA to serve as a useful proxy for performance-relevant unobservables not captured by LSAT may be attributable, at least in part, to both grade inflation and the incomparability of UGPA across undergraduate majors and institutions.[516] In fact, one study concluded that "evaluation has become so flawed that employers, graduate schools, and professional schools that try to use grades to identify outstanding prospects are likely often engaging in a futile exercise. Increasingly, they have to rely on standardized tests to make evaluations of student talent."[517]

Lastly, S&S mention that race-mismatch interaction effects were explored and did not impact their results (at least for the first set of analyses, Table 3–Models 1–6);[518] however, it is not clear how S&S examined these effects. Were the race-mismatch interactions examined for the Index-based measure of mismatch or only the LSAT-based measure? Were the race variables interacted with the linear or nonlinear effect of mismatch? Were interactions between race and other mismatch-relevant variables, such as LSAT, Index, and LGPA, examined (and, if so, linearly or nonlinearly)? These questions are important because a variable may have a nonlinear effect, a nonadditive effect, both, or

515 The effect size (β) and 95% confidence interval ([lower bound, upper bound]) for "Black" is β = -1.04 [-1.22, -0.85] in the LSAT model and β = -1.17 [-1.47, -0.87] in the Index model. The effect size for "Latinx" is β = -0.54 [-0.68, -0.39] in the LSAT model and β = -0.60 [-0.79, -0.42] in the Index model. These estimates are based on students attending a law school in the top two tiers (Tier 1 & Tier 2) and on matching students according to sex, undergraduate major, work status during law school, and law school tier. The results were very similar when disaggregating Tiers 1 & 2. My prior analysis of the BPS data revealed that:

> 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. This effect size was virtually identical when examining students in the next highest tier . . . . By focusing on the top two tiers, my analysis has the advantage of examining a small, and relatively homogenous, group of law schools. The top tier is composed of sixteen schools and the second tier is composed of fourteen schools, 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.

Thaxton, *supra* note 1, at 978.

516 *See generally* Stuart Rojstaczer & Christopher Healy, *Where "A" is Ordinary: The Evolution of American College and University Gradings*, 114 Tchr. Coll. Rec. 1 (2012) (noting that private colleges and universities, on average, award higher grades than similarly selective public institutions, whereas science and engineering-focused schools award lower grades than similarly selective schools).

517 Rojstaczer & Healy, *supra* note 516, at 18. *But see* Henderson, *supra* note 155 (noting that LSAT and UGPA have differential predictive ability depending on the type of assessment used and the year in law school) *and* Kidder, *supra* note 109 (discovering that the correlation between Index and LGPA decreases significantly each year, with the largest reduction between 1L and 2L years).

518 Sander & Steinbuch, *supra* note 15, at 734 n.33 (stating that "race-mismatch interaction effects . . . did not produce insights or results that depart in interesting ways from those shown in Table 3").

neither.[519] Without more information about their analysis of interaction effects, it is also uncertain as to whether S&S interpret these effects properly.[520]

These questions are especially relevant because, in prior work, Sander has incorrectly interpreted interaction effects and their associated measures of uncertainty, and that has resulted in him drawing invalid conclusions about mismatch effects.[521] A statistical interaction occurs when the effect (regression coefficient) of one explanatory variable on the outcome variable changes depending on the level of another explanatory variable. Therefore, when examining the effect of race across the range of the other explanatory variable comprising the interaction effect, both the effect size for race and the standard error of that effect size are conditional.[522] In other words, the race effect on bar passage and its associated statistical significance must be assessed at different levels of the other variable (e.g., mismatch). The race effect may be statistically significant at some levels of the other explanatory variable, but not others.[523] S&S use logistic (logit) regression and, as I have explained in prior work, "The interpretation of the effects of interaction terms is particularly complicated for nonlinear models, encompassing binary [logit] models . . . because of the distinction between the scale of interest (i.e., the natural scale) and the scale of estimation [logit]. . . . Consequently, the magnitude, sign, and significance of the interaction coefficient [on the logit scale] may not yield reliable information about the true effect of the variable of interest."[524] If the effect of race on bar passage is conditioned by mismatch, LSAT, Index, or LGPA, modeling the race effect as additive, rather than interactive, could

519 James Jaccard, et al., Interaction Effects in Multiple Regression 21 (1990); *see also* Thaxton & Agnew, *supra* note 499, at 902; Long, *supra* note 275, at 24 (emphasizing that endogeneity issues—i.e., correlation between the explanatory variable(s) and the residual term—arise, inter alia, when analysts fail to properly model nonadditive and nonlinear effects when such a relationship is present in the data).

520 *See infra* note 524 and accompanying text.

521 Thaxton, *supra* note 1, at 1009–10; *see also* Bjerk, *supra* note 73, at 4 (discussing Sander's incorrect interpretation of an interaction effect in his reply to Ayres and Brooks); Ayres & Brooks, *supra* note 73, at *10 (same).

S&S's analysis of race-mismatch interaction effects may be biased because of their top-coding and discretizing of the mismatch variable which appreciably reduces its variance and introduces nonignorable measurement error. *See supra* notes 345–346 & 376–378. The detection of race-mismatch interaction effects becomes particularly difficult in the presence of (1) multicollinearity, (2) measurement error, and (3) small sample sizes. *See generally* Thaxton & Agnew, *supra* note 499, at 891. These problems emerge, in large part, when analyzing weak, uninformative, or sparse data, *see* Achen *supra* note 331; Winship & Western, *supra* note 390, because highly correlated constituent variables (i.e., variables that comprise the interaction) result in low residual variance for the interaction term. When residual variance is low, the corresponding effect size for the interaction will be low as well. Thaxton & Agnew, *supra* note 499, at 891.

522 Jaccard et al., *supra* note 519, at 25–26.

523 *Id.* at 27.

524 Thaxton & Agnew, *supra* note 499, at 903–04.

potentially mask a significant race effect.[525] Any statistically significant race effect at a particular level of the other explanatory variable undermines S&S's claim about the spuriousness of the race effect.[526]

Conclusion

Nearly twenty years ago, Richard Lempert and colleagues remarked, "Professor Sander's rhetorical skills make untenable claims sound plausible, especially for those who are unfamiliar with data and methods of statistical analysis."[527] Other analysts added that Sander's "statistical misstatements and modeling errors in *Systemic Analysis* mean that the conclusions appear to have far more evidentiary support than they in fact do."[528] S&S's school-specific analysis of mismatch theory allegedly supports several of the prior assertions made by Sander and other proponents of mismatch when analyzing the BPS data. But S&S's analysis shares similar shortcomings of those prior studies that undermine the validity of S&S's results because of ineffective experimental design, inappropriate analyses, and flawed reasoning.[529] S&S's

525 If the effect of race on LGPA is conditioned by UGPA, LSAT, or mismatch, then modeling the effect of race as additive could also bias the results. *See, e.g.,* VanderWeele & Robinson, *supra* note 112, at 479 (explaining the importance of examining interaction effects with race and mediating variables when measuring discrimination). The combination of interactive and mediated effects is often referred to as a conditional indirect effect and "represent[s] mediation effects that vary in strength conditional on the value of at least one moderator variable." Kristopher J. Preacher, et al., *Addressing Moderated Mediation Hypotheses: Theory, Methods, and Prescriptions*, 42 Multivariate Behav. Res. 185, 195 (2007); *see also* VanderWeele, *supra* note 257, at 371 (describing the overall effect of an exposure variable as a combination of four components: the portion of the effect that is due to (a) neither mediation nor interaction (unmediated-noninteractive), (b) just interaction (unmediated-interactive), (c) just mediation (mediated-noninteractive), and (d) both mediation and interaction (mediated-interactive)); Dominique Muller et al., *When Moderation is Mediated and Mediation is Moderated*, 89 J. Personality & Soc. Psychol. 852, 856 (2005) (explaining that moderated mediation occurs when the interaction only impacts the indirect effect (i.e., mediated effect) of the treatment variable, whereas mediated moderation occurs when the interaction impacts both the direct (i.e., unmediated) and indirect effect of the treatment variable); *see supra* note 143 and accompanying text.

526 Hayes, *supra* note 215, at 7 (noting that causal inference entails, *inter alia*, understanding "under what circumstances, or for which types of people, does X exert an effect on Y and under what circumstances, or for which type of people, does X not exert an effect").

S&S do not indicate whether they examined race-school interactions (i.e., a cross-level interaction), but a significant interaction effect of this type would also contradict mismatch theory. *See supra* note 446. Researchers have emphasized the importance of analyzing interactions between race and institutional selectivity/difficulty when analyzing the impact of affirmative action policies on student outcomes. Smyth & McArdle, *supra* note 45, at 364 (noting that "college selectivity will be included in [] *slopes-as-outcomes* [i.e., cross-level interaction] models to assess whether this alters the fixed estimates [i.e., race effects] or accounts for any cross-college variation in [race] effects").

527 Lempert et al., *supra* note 38, at 7.

528 Chambers et al., *supra* note 178, at 1873.

529 *See supra* note 141. Economist Charles Manski explains that it is very "tempting for

models exploring the relationships between race, mismatch, and bar exam performance exhibit poor overall predictive ability, provide low incremental explanatory power, and appear to be quite sensitive to model (mis)specification. S&S's analyses concerning the confounding/mediating effect of mismatch on the relationship between race and bar exam performance fail to employ the appropriate statistical tests for credible statistical inference. And their analyses further neglect to separate confoundedness or mediation effects from scaling effects, so the inferences they draw from these models are unreliable.

S&S state that large racial disparities in bar passage are "a serious problem that must be addressed, and [the legal academy must] undertake inclusive, honest, and data-driven conversations."[530] On this point we agree,[531] which is why I

researchers to maintain assumptions far stronger than they can persuasively defend in order to draw strong conclusions" because the "scientific community rewards researchers who produce strong novel findings and the public, impatient for solutions to its pressing policy concerns, rewards researchers who offer simple analyses leading to unequivocal policy recommendations." Manski, *supra* note 7, at 2. Similarly, statistician Richard McElreath notes that "there are strong incentives for . . . statisticians to exaggerate the power and importance of their advice . . . [because] few want a statistician who admits the answers as desired are not in the data as collected. . . . Scientists desire results, and they will buy and attend to statisticians and statistical procedures that promise them. What we end up with is too often *horoscopic*: vague and optimistic, but still claiming critical importance." McElreath, *supra* note 33, at 565.

530 Sander & Steinbuch, *supra* note 15, at 741. *But see* Michele Landis Dauber, *The Big Muddy*, 57 Stan. L. Rev. 1899, 1910 (2005) (observing that Sander's claim that a "substantive, data-driven debate on law school affirmative action is beginning to emerge" is false, and that "the 'debate' is much more an effort at *post hoc* quality control by the scholarly community" given that "Sander has failed to establish that his analysis and methods are sufficiently valid to support his claims").

Honest, data-driven conversations must avoid exaggeration, especially in the face of inadequate data and fragile empirical results. *See generally* Phillip I. Good & James W. Hardin, Common Errors in Statistics (and How to Avoid Them) 165 (4th ed. 2012) ("We advocate interpreting [statistical] reports not merely with a grain of salt but with an entire shaker."). Sander's prior responses to critics routinely employ hyperbolic language when defending his analyses. For example, he has claimed: "The evidence points overwhelmingly toward a large law school mismatch problem," Sander & Taylor, at *supra* note 22, at 86; "[When] correctly done, [Sander's reanalysis of Ayres and Brooks's test] produced a stunning confirmation of [mismatch] theory," *id*. at 83; "These are stunning results," Sander, *Reply to Critics*, *supra* note 61, at 1974; "These results are even more striking," *id*. at 1975, "[Our] results are very powerful confirmation of mismatch," Sander, *Replication of Mismatch Research, supra* note 24, at 82; and "Once again, we have a strong finding that it is preferences, not race, which explain large racial disparities in educational outcomes," Sander, Brief in *SFFA*, *supra* note 13 at 29. Given that journal editors often select articles based on overclaiming and exaggeration, Sander's hyperbole increases the likelihood that his findings will be published and highlighted in the media, irrespective of their inaccuracy and unreliability; *see also* McElreath, *supra* note 33, at 567 (noting that journal editors often "select for hyperbole, since honestly admitting limitations of work only hurts a paper's chances").

531 *See infra* note 540 and accompanying text. *See generally* Max Weber, *Objectivity in Social Science and Social Policy*, *in* The Methodology of the Social Sciences 50 (Edward A. Shils & Henry A. Finch eds., 1949) (noting that the growth and specialization of the social sciences, along with the establishment of methodological principles, leads to a focus on the analysis of data

have advocated for greater transparency about research design limitations and underscored the importance of preprocessing data (e.g., covariate balancing) before testing mismatch theory with statistical analyses, together with conducting meaningful diagnostic checks afterward (e.g., model averaging, cross-validation, collinearity analysis, and assessment of model fit).[532] These approaches, when utilized correctly, can reduce unwarranted dependence on model assumptions and better facilitate data-driven conclusions.[533]

But to be clear, the concerns I have articulated with respect to S&S's comparisons of coefficients across nested models, assessment of model fit, model dependence, school fixed effects, and robustness analysis are secondary to the key issue of S&S's exclusion of law school grades from their analyses.[534] S&S's school-specific analysis neither demonstrates that mismatch in entering credentials explains learning mismatch independent of race, nor establishes that learning mismatch is primarily responsible for differences in bar exam performance. Both claims are core assumptions of their theoretical framework for explaining racial differences in bar exam performance.[535] Instead, S&S

---

as an end in itself).

532 Thaxton, *supra* note 1, at 940–41, 973–76, 984–88 & 1003–08 (transparency about data limitations, model averaging, cross-validation, covariate balancing, assessment of model fit); *see also* Cristobal Young, *supra* note 297 (model averaging); Harrell, *supra* note 304 (cross-validation); Johnston et al., *supra* note 234, at 1959 (collinearity analysis); Gary King & Langche Zeng, *supra* note 340, at 149 (covariate balancing); Long, *supra* note 275, at 98–100 (assessment of model fit).

533 *See supra* notes 330 & 403–410 and accompanying text. *See generally* Weisberg, *supra* note 130, at 279 ("Bias is an almost unavoidable fact of research life that must be taken into account in evaluating a study's methodology. Rather than an attempt to engineer perfection, the research process may be better regarded as attempting to solve a complex puzzle in light of ambiguous and sometimes misleading hints and clues."); Manski, *supra* note 7, at 2 ("The conclusions that one can draw from an empirical analysis are determined by the assumptions and the data that one brings to bear. In social science research, the available data are typically limited and the range of plausible assumptions wide; hence the generally accepted conclusions are necessarily weak."); *accord* Levmore, *supra* note 80, at 614 ("While empiricists like to say that 'some data is better than no data,' the seemingly obvious reliance or insistence on data-driven analysis is misplaced or even reckless because there is often reason to think that 'some data' often leads to misleading conclusions . . . [and] omitted variables are at the root of the problem.").

534 *See* Part III.A. The exclusion of the examination of race on LGPA is directly relevant to prior critiques of Sander's use of cross-racial comparisons when analyzing the BPS data. *See supra* notes 181–186 and accompanying text. A fundamental problem with the cross-racial comparisons in this earlier work is that the analyses do not take into account the impact race may exert on other mismatch-relevant variables (i.e., post-treatment bias). *See supra* note 129. In other words, Sander's analyses ignore the indirect effects of race on bar exam performance. *See supra* notes 156 & 181–182; *see also* Thaxton, *supra* note 1, at 864 (distinguishing total, direct, and indirect effects).

535 Sander & Steinbuch, *supra* note 15, at 716. S&S's have data on LGPA for all three schools to directly examine that relationship and disentangle the direct and indirect race effects. *Id.* at 743 (noting that their data contain information on UGPA and LGPA for UCLA for some, but not all, of the six cohorts; however, they do not identify which specific cohorts); *see also*

merely posit that credential distance affects bar passage through learning mismatch, and race does not (independent of credential distance), without providing sufficient justification.[536] By failing to empirically evaluate these assumptions, S&S's analysis falls short of causal explanation and, at best, constitutes mere causal description under the implausible assumption that S&S's models adjust for all relevant confounding variables.[537]

In *How Not to Lie About Affirmative Action*, I attempt to carefully distinguish between theoretical and empirical analysis.[538] Whereas theoretical analysis consists of studying propositions about particular processes or phenomena and deriving their consequences,[539] empirical analysis involves the assessment of the truthfulness or accuracy of the theory's assumptions, predictions, and plausible alternatives.[540] With respect to theory, I have emphasized that the "primary conceptual flaws present in [S&S's] analyses pertain to the posited theoretical relationships among the key variables forming the core of the mismatch hypothesis[.]"[541] With respect to empirics, I have demonstrated

supra note 47 & 200 and accompanying text.

536 *See* Part III.A; *see generally* Morton, *supra* note 77, at 145 (noting that analyses that evaluate nonverified assumptions are necessary to identify limits of theory and are crucial for understanding social phenomena); *accord* Granato et al., *supra* note 92, at 784 (cautioning that scientific discovery is severely undermined when analysts forgo the evaluation of causal mechanisms and instead rely on "hand-me-down" applied statistical techniques and the manipulation of standard errors and their associated significance tests); *see also supra* note 228 and accompanying text (noting that race may influence mediating variables, such that estimates of race effects may be seriously biased when this influence is not properly measured).

537 *See supra* notes 237–252 and accompanying text.

538 Thaxton, *supra* note 1, at 838 n.5 ("Conceptual [i.e., theoretical] and empirical analysis are distinct inquiries."); *accord* Robert K. Merton, Social Theory and Social Structure 141 (1968) ("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.") (emphasis in original).

539 Granato et al., *supra* note 235, at 12 (2021) (noting that theoretical analysis entails "comput[ing] 'the consequences' of the new law (predictions)"). Sociologist Guillermina Jasso claims that theoretical analysis is underutilized and often inadequate because "students typically learn an array of methods for empirical analysis but virtually no methods for theoretical analysis. Thus, the temptation to prematurely 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." Guillermina Jasso, *Principles of Theoretical Analysis*, 6 Socio. Theory 1, 11 (1988); *accord* Merton, *supra* note 538, at 141 ("It is my impression that current sociological training is more largely designed to make students understand the first [i.e., empirical analysis] than the second [i.e., conceptual analysis].").

540 Morton, *supra* note 77, at 101; Granato et al., *supra* note 235, at 12 (explaining that empirical analysis compares theoretical predictions with observations from experiments or experience).

541 Thaxton, *supra* note 1, at 838 n.5 & 968; *see, e.g.*, Parts II.B, III.E & III.F (discussing problems with S&S's propositions pertaining to the relationships between key variables in mismatch

that S&S "bridge data gaps with assumptions that consistently favor their theory and fail to conduct (or report) appropriate sensitivity analyses to examine the consequences of different assumptions and the robustness of their results,"[542] and that their "empirical analyses are plagued by several significant shortcomings that must be fully addressed before claiming any empirical support for mismatch theory."[543]

Science advances through the reasoned criticism of received wisdom,[544] so my goal in *How Not to Lie About Affirmative Action*, as well as my assessment here of S&S's *Mismatch and Bar Passage*, is to lay out both the challenges to mismatch theory and, potentially, fruitful avenues of future work.[545] I view my task as introducing or highlighting key methodological issues and sharpening our theoretical approach to understanding racial disparities in law school-related outcomes.[546] According to statistician Andrew Gelman, "Making errors is

---

theory).

542 Thaxton, *supra* note 1, at 969 (quote edited for clarity); *see, e.g.*, Parts II.B, III.B, III.C, III.D & III.E (describing deficiencies with S&S's quasi-experimental design, coding of variables, and use of specific statistical tests); *see also* Weisbach, *supra* note 8, at 103. ("[A] crucial element in a professional reporting of empirical results [in economics] is a serious discussion of their robustness . . . . Readers naturally wonder how these [model specification] choices affect the paper's conclusions.").

543 Thaxton, *supra* note 1, at 857 (quote edited for clarity). *Compare* Parts III.D & III.E (highlighting the interrelatedness of S&S's indefensible or highly implausible modeling choices that exacerbate bias) *with supra* note 89 (noting that Sander's empirical analyses, amicus briefs, and responses to critiques reveal a "lack of meaningful engagement with the relevant methodological literature establishing the rules and best practices of statistical inference to justify his experimental design, analyses, or reasoning").

544 *See generally* Karl Popper, The Logic of Scientific Discovery 278–79 (1959) ("Our method of research is not to defend [our theories] in order to prove how right we were. On the contrary, we try to overthrow them . . . [u]sing all the weapons of our logical, mathematical, and technical armory."); *accord* Weisberg, *supra* note 130, at 279 ("The scientist's job is to see through a fog of incomplete evidence and misdirection, and attempt to discern the underlying causal pattern. Toward this end, she must employ every tool of inference at her disposal and ultimately must make a convincing argument to support a credible, albeit often tentative, conclusion."); Epstein & King, *supra* note 34, at 9 ("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."); Burk, *supra* note 14, at 366 (noting that scientific inquiry is undermined when scholars adopt the legal profession's norms of adversarialism and advocacy).

545 Thaxton, *supra* note 1, at 984–93 (*see* "Properly Measuring Mismatch"). *Compare* Angrist & Pischke, *supra* note 238, at 115 (2008) (noting that credible "causal inference has always been the name of the game in applied econometrics") *and* Merton, *supra* note 538, at 122 (arguing that the discontinuity between theoretical work and empirical work retards the progress of social science).

546 Nine related research findings may be especially relevant for understanding the relationship between race, learning mismatch, and bar exam performance. First, Black students have lower UGPA than non-Black students, independent of AP courses taken, HSGPA, SAT/

inevitable; learning from them is not. To learn from errors, we want a system of science that facilitates and provides incentives for such learning[.]"[547] The

ACT scores, undergraduate major, difficulty of courses, academic aspirations, academic assistance, social preparation, family income, parental education, state of residency, and selectivity of college/university. Second, Black students perform worse on the LSAT than non-Black students, holding constant UGPA, college/university attended, college major, age, and year of graduation. Third, LSAT is a weak overall predictor of LGPA. Fourth, the relationship between LSAT and LGPA declines each subsequent year in law school. Fifth, LSAT is a weaker predictor of grades in writing and experiential learning courses than in exam-based courses. Sixth, the LSAT is a much weaker predictor of grades for Black students than non-Black students. Seventh, Black students may have the largest grade differentials between in-class test performance and other methods (even when accounting for the grading curve). Eighth, doctrinal courses explain only a modest amount of variation in bar passage. Ninth, an abundance of experiential learning courses taken in law school does not lower students' likelihood of passing the bar exam (and may even increase the likelihood). *See* Bowen et al., *supra* note 181; Charles et al., *supra* note 45; Curcio et al., *supra* note 155; Henderson, *supra* note 155; Kidder, *supra* note 109; Massey et al., *supra* note 181; Rothstein, *supra* note 181. Viewed collectively, these findings suggest that Black students are the most systematically disadvantaged by law schools' reliance on LSAT for reasons unrelated to aptitude for legal learning and eventual bar passage.

There is also *modest* evidence of a reverse mismatch effect—that is, affirmative action increases the number of Black lawyers or the quality of legal employment (and salary) for Black lawyers included in the BPS data. Ayres & Brooks, *supra* note 65, at 1809 & 1823 (noting that both white and Black students are more likely to pass the bar if they attended a more selective school, and the effect is even stronger for Black students); Ho, *supra* note 323, at 7 fig.3 (same); Rothstein & Yoon, *supra* note 323, at 21–23 (reporting that "Black students [attending more selective law schools] are much more likely to obtain good jobs than similarly-qualified white students, with a salary premium around 10–15 percent"); *see also* Bowen & Bok, *supra* note 181, at 59–65 (finding a reverse mismatch effect at the undergraduate level); Mary J. Fischer & Douglas S. Massey, *The Effects of Affirmative Action in Higher Education*, 36 Soc. Sci. Res. 531, 544 (2007) (same); Sigal Alon & Marta Tienda, *Assessing the "Mismatch" Hypothesis: Differences in College Graduation Rates by Institutional Selectivity*, 78 Socio. Edu. 294, 303 (2005) (same).

A reverse mismatch effect is akin to the sociological phenomenon of an "endogenous feedback effect" or "norm effect" in which the propensity of an individual to behave in some manner varies with the prevalence of that behavior in their social environment. Manski, *supra* note 7, at 31; *see also* Carnevale et al., *supra* note 5, at 1 (noting that students with modest entering credentials who enroll in the most competitive academic institutions rise to the challenge and perform comparably to their more highly credentialed peers).

547 Gelman, *supra* note 24, at 39; *accord* King, *supra* note 25, at 684 ("Although many mistakes are caught by perceptive colleagues, many more slip by. . . . If these . . . rules were followed—and adequate information were provided with which to assess quantitative analyses—many future mistakes could be avoided."); Ho, *supra* note 100, at 2004 ("Yet just as scholars have realized the potential for empirical techniques that have energized research frontiers in the social sciences, we must also become aware of the assumptions, limitations, and credibility of those techniques."). *But c.f.* Granato et al., *supra* note 539, at 14 ("While there has been improvement in quantitative social science, there is also cause for concern that social scientists are not absorbing the scientific lessons and emphasis of prior social scientists."); McElreath, *supra* note 33, at 567 ("[T]here is no system for linking published papers to later peer criticism, and there are few formal incentives to conduct it."); Levmore *supra* note 80, at 612 (suggesting that much of the work in empirical law and economics "should not be taken seriously" and a replication crisis in the field is afoot because seriously flawed empirical work

quality of scholarship in the field of empirical legal studies, on the whole, has drastically improved in recent years,[548] and this is attributable to both legal scholars' improved training in social scientific methodologies and the members of this intellectual community diligently (re)enforcing widely accepted norms and best practices of statistical inference.[549] Perhaps old habits die hard, but they do die.[550]

"spreads as more scholars copy the mistakes [of prior researchers] and engage in empirical work as a means of entry into the field.").

548 *See* Daniel S. Nagin & Robert J. Sampson, *The Real Gold Standard: Measuring Counterfactual Worlds That Matter Most to Social Science and Policy*, 2 Ann. Rev. Criminology 123, 124 (2019) ("The move to prioritize methodologies in terms of assessments about their capacity for providing convincing empirical demonstrations of causality has also changed the kinds of papers published–methods that were acceptable thirty years ago as a means for identifying causal relationships, such as a simple regression, typically no longer suffice.").

549 Lempert, *supra* note 81, at 1; *accord* Ho, *supra* note 28, at 2016 ("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."); Barnes, *supra* note 46, at 792 ("Empirical research, in particular, involves the often public debate regarding the appropriate methods, analysis, and conclusions to be drawn from data . . . . The academic process of replication, further investigation, and debate (like the methods of science more generally) is built to find flaws in current research in order to improve knowledge."); Gelman, *supra* note 65, at 11 ("And the afterlife of a published paper is not just with its readers but also with later researchers who apply its findings or methods. As such, post-publication criticism is an important part of the feedback mechanism of science, and we can see it as a form of collaboration occurring across time as well as space."); Brown et al., *supra* note 8, at 2567 ("Postpublication [sic] discussion platforms such as PubPeer, PubMed Commons, and journal comment sections have led to useful conversations that deepen readers' understanding of papers by bringing to the fore important disagreements in the field."); *see also supra* note 25 (noting that quantitative methodologists from a wide range of disciplines have both developed and enforced norms and best practices of statistical inference in their respective fields, titling their interventions *How Not to Lie*).

550 *See, e.g.*, Ruth Gunther McGrath, *Old Habits Die Hard, but They Do Die,* 95 Harv. Bus. Rev. 54 (2017).

Appendix: S&S's Tables 1–6 & Table A

Table 1: Characteristics of the Three Law School Datasets			
Data characteristic	UCLA	Davis	UALR
LSAT scores	Yes	Yes	Yes
UGPA	No	Yes	Yes
Index	No	Calculated	Calculated
Race	Yes	Yes	Yes
Law school grades	No	Yes	Yes
Incoming Transfers excluded?	Yes	Yes	No
In-state bar results?	Yes	Yes	Yes
Graduating years covered	2000, 2001, 2005	1997–2011	2005–2011
Number of distinguishable cohorts	3	5	1
All entering students included?	No; only eventual In-state bar-takers	Yes	Yes
Observations	752	3,290	899
Observations of in-state bar-takers	752	2,333	723
Source: Sander & Steinbuch, *supra* note 15, at 725, tbl.1.			

Table 2: First-time Bar Passage for Graduates Attempting the In-state Bar

LSAT Range	UCLA		Davis		UALR	
	Attempts	Passing %	Attempts	Passing %	Attempts	Passing %
143 or lower	n/a		1	0%	24	37%
144-146	n/a		8	25%	51	51%
147-149	n/a		16	44%	120	75%
150-152	9	22%	37	51%	149	79%
153-155	18	39%	76	71%	165	79%
156-158	27	67%	179	79%	99	86%
159-161	60	88%	305	85%	68	87%
162-164	193	92%	175	86%	29	97%
165-167	198	98%	80	95%	15	94%
168 or higher	126	97%	45	84%	4	100%
Median LSAT	164		160		152	
Total pass rate	89%		81%		78%	
Cohorts:	2001, 2002, 2005		1997-99, 2000-02		2005-2011	
Note: Median LSAT reported for the cohorts included in the data						
Source: Sander & Steinbuch, *supra* note 15, at 727, tbl.2.						

Table 3: Logistic Regression Models of First-time Bar Passage, 9 Cohorts at all Three Schools						
Independent variables	Model 1: Race alone	Model 2: Add LSAT	Model 3: Add Mismatch Deficit	Model 4: Add School Fixed Effects	Model 5: Categorical Mismatch	Model 6: Categor. Mismatch w/School Fixed Eff.
African Am.	.28***	.56**	.71*	.71*	.71*	.72
Hispanic	.41***	.53***	.76*	.79	.77*	.81
LSAT		1.10***	1.05***	1.03*	1.05***	1.04*
Mismatch Deficit			1.13***	1.16***		
Mm Lev 1					.94	.90
Mm Lev 2					.78	.74
Mm Lev 3					.55***	.51***
Mm Lev 4					.36***	.30***
Mm Lev 5					.42***	.37***
Mm Lev 6					.29***	.22***
Mm Lev 7					.15***	.11***
UALR				.48**		.51**
Davis				.45***		.44***
Constant	5.05***	.000***	.003**	.060	.001***	.016
Observations	3,656	3,656	3,656	3,656	3,597	3,597
Somers' *D*	.13	.35	.36	.39	.37	.39
Significance levels are: *p<.1; **p<.05; ***p<.005 (two-sided)						
Source: Sander & Steinbuch, *supra* note 15, at 731, tbl.3.						

Table 4: Revisiting Logistic Regression Models 3 and 6, from Table 3 Including Only Davis and UALR and Using, Alternately, LSAT and Index as Academic Measures

Independent variables	Model 7: Using LSAT	Model 8: Using Index	Model 9: Using LSAT, categorical mismatch, school FE	Model 10: Using Index, categorical mismatch, school FE
African Am.	.64**	1.11	.63**	1.02
Hispanic	.77	.96	.79	.93
LSAT/Index	1.03**	1.02**	1.04*	1.05**
Mismatch Deficit	1.15***	1.22***		
Mm Lev 1			.90	.85
Mm Lev 2			.77**	.54***
Mm Lev 3			.51**	.41***
Mm Lev 4			.33***	.38***
Mm Lev 5			.39**	.20***
Mm Lev 6			.24***	.29***
Mm Lev 7			.14***	.10***
Davis			.87	.73
Constant	.06	.14	.008	.002
Observations	3,005	3,005	3,005	3.005
Somers' *D*	.36	.37	.38	.39
Significance levels are: *p<.1; **p<.05; ***p<.005 (two-sided)				
Note: Forty-four students did not have UGPA data; we omitted these from all four models.				
Source: Sander & Steinbuch, *supra* note 15, at 735, tbl.4.				

Table 5: Average UGPA by LSAT Score, Davis & UALR		
For students with LSAT scores of…	…Average UGPAs (with # of students in parentheses) were:	
	at Davis	at UALR
146	3.69 (6)	3.29 (33)
150	3.50 (25)	3.26 (71)
154	3.54 (93)	3.21 (65)
158	3.61 (150)	3.23 (32)
Source: Sander & Steinbuch, *supra* note 15, at 737, tbl.5.		

Table 6: Index Deficit and Bar Passage Rates at Davis and UALR				
Index Deficit	Non-Black Bar-Takers		Black Bar-Takers Only	
	Number	% Passing	Number	% Passing
Below -15	7	0%	5	0%
-12 to -15	25	32%	18	28%
-9 to -12	51	39%	20	60%
-6 to -9	156	56%	30	63%
-3 to -6	355	71%	23	52%
-0.1 to -3	796	81%	11	73%
No deficit	1495	88%	13	85%
Total	*2885*	*81%*	*120*	*56%*
Source: Sander & Steinbuch, *2018 Draft*, *supra* note 332, at *14, tbl.6.				

Appendix Table A: Logistic Regression Models of First-time Bar Passage, California Schools Only						
Independent variables	Model 1: Race alone	Model 2: Add LSAT	Model 3: Add Contin. Mismatch	Model 4: Contin. Mismatch w/ School Fixed Effects	Model 5: Categorical Mismatch	Model 6: Categor. Mismatch w/School Fixed Eff.
Black	.25***	.52**	.58**	.60**	.58**	.59**
Hispanic	.37***	.67**	.74*	.76*	.75*	.77
LSAT		1.14***	1.06**	1.01	1.07**	1.01
Mismatch Deficit			1.11***	1.19***		
Mm Lev 1					1.03	.82
Mm Lev 2					.80	.58**
Mm Lev 3					.62*	.39***
Mm Lev 4					.42**	.22***
Mm Lev 5					.58	.30***
Mm Lev 6					.35**	.14***
Mm Lev 7					.18***	.07***
Davis				.43***		.41***
Constant	5.46***	.000***	.000**	2.05	.000**	1.81***
Observations	2,938	2,938	2,938	2,938	2,938	2,938
Somers' *D*	.13	.35	.36	.33	.37	.33
Significance levels are: *p<.1; **p<.05; ***p<.005 (two-sided)						
Source: Sander & Steinbuch, *supra* note 15, at 742, tbl.A.						

Figures: Causal Diagrams of S&S's Theoretical and Empirical Models

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

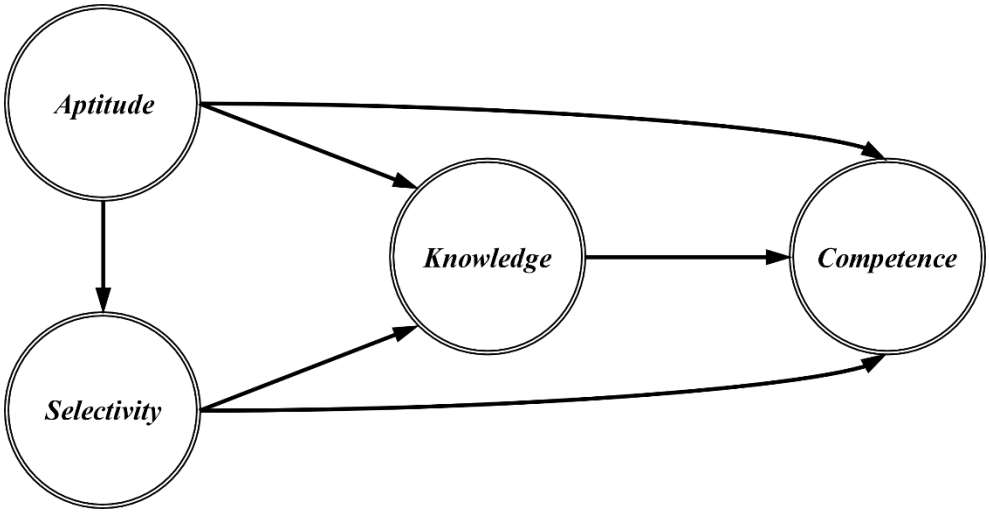

Figure 2: Sander's Theoretical Model (Manifest Indicators)

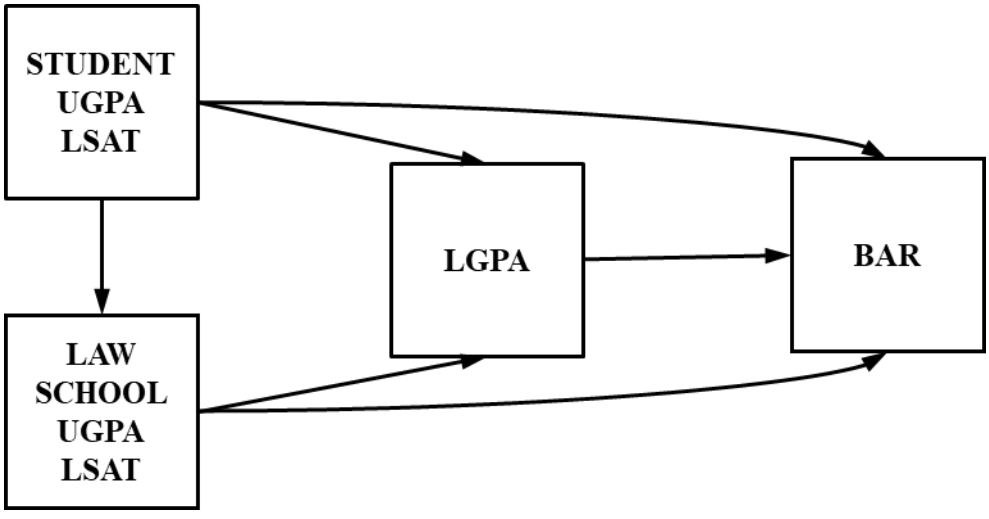

Figure 3: Sander & Steinbuch's "Black Box" Empirical Model
(Latent Constructs)

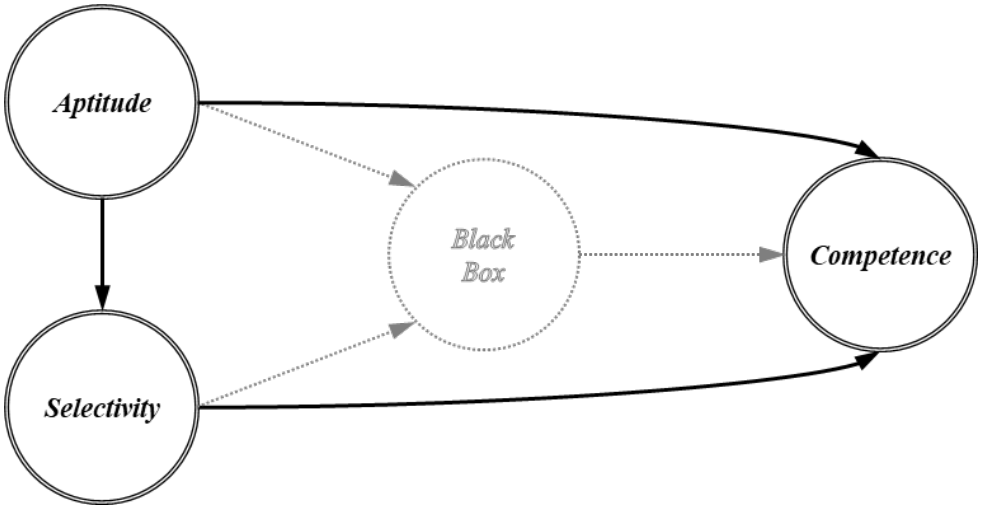

Figure 4: Sander & Steinbuch's "Black Box" Empirical Model
(Manifest Indicators)