SOLVING "PROBLEMS NO ONE HAS SOLVED": COURTS, CAUSAL INFERENCE, AND THE RIGHT TO EDUCATION

*Christopher S. Elmendorf**
*Darien Shanske***

For several decades now, liberal public-interest litigators have argued that insufficiently generous subsidies for the education of disadvantaged children violate the education or equal protection clauses of state constitutions. Their opponents responded that the evidence that more money would substantially improve student outcomes was too speculative to warrant judicial intervention. More recently, conservative public-interest litigators have started attacking teacher tenure and seniority protections on the same constitutional grounds. In response, liberals are parroting the evidentiary and causation arguments that conservatives made in school-finance cases. Both factions in this back-and-forth have overlooked a critically important fact: The state's own choices substantially determine whether researchers—and hence litigators—can produce credible evidence concerning the causal effect of state laws and funding arrangements on the outcomes that ground the education right. States exercise this control through the architecture of administrative data systems; through the rules for assigning students, programs, and funding to schools; through the manner in which educational reforms are rolled out; and through the terms on which the state provides access to administrative data.

Recognizing that the information needed to enforce the education right is endogenous to law, we make the case for a new, information-oriented, education-rights jurisprudence in which courts would intervene not simply to resolve disputes about how to organize and fund the education of disadvantaged children, but to enable more credible tests of the competing predictions of warring education reformers. Our analysis directs attention to several issues that have been overlooked since education-rights litiga-

* Elmendorf is Martin Luther King Jr. Professor of Law, UC Davis School of Law.

** Shanske is Professor of Law, UC Davis School of Law. For helpful feedback on earlier drafts, we thank Ash Bhagwat, Derek Black, Kyra Caspary, Clayton Gillette, Joe Grodin, Angela Harris, David Horton, Orville Jackson, Kevin Johnson, Bill Koski, Carlton Larson, Bertrall Ross, Jim Ryan, Reuel Schiller, David Schleicher, Rich Schragger, Doug Spencer, Steve Sugarman, and Aaron Tang. We are also grateful to Paige Kowalski and Elizabeth Dabney of the Data Quality Campaign for sharing their experience and insights with us. The quotation in the title of this article is from the New Jersey Supreme Court's decision in *Abbott v. Burke*, 575 A.2d 359, 363 (N.J. 1990).

tion got underway in the 1970s and does so at a critical moment—as educational research undergoes a "scientific revolution" bearing on the very questions that must be answered to implement the education right.

TABLE OF CONTENTS

I. INTRODUCTION

Perhaps the most important feature of state constitutional law is the right to education. Nearly every state constitution provides for a system of free public schools, and many state courts treat their constitution as guaranteeing, through the education or equal protection clauses, a reasonably adequate or roughly equal education.[1] Animating these decisions is a vision of substantive economic and political opportunity for disadvantaged children. The resources owed to a school or school district vary with the needs of the student population, and the constitutional sufficiency of the educational system depends on the *quality* of education actually provided.[2] It is the duty of the state to prepare every child for a lifetime of gainful employment and meaningful democratic participation.

Standing between this noble vision and its realization is a not-so-minor epistemic problem: There is no social scientific (or political) consensus about what *changes* to the education system would most likely bring about substantial

[1] For a historical overview of these clauses, see EDUCATION IN THE 50 STATES: A DESKBOOK OF THE HISTORY OF STATE CONSTITUTIONS AND LAWS ABOUT EDUCATION (Institute for Educational Equity and Opportunity, July 2008).

[2] *Id.*

improvements in the adult outcomes of high-poverty, high-need student populations. Liberals want more money for public schools, smaller class sizes, universal preschool, and socioeconomic integration.[3] Conservatives (and some heterodox liberals) want to restructure the terms of teacher employment, and they want traditional public schools to face more competition from charter and private schools.[4]

From the 1970s to the present, liberals have tried to advance their program of reform through the courts, with mixed success.[5] Defendants and conservative intervenors responded that the evidence that more money would make a difference was too equivocal to warrant judicial intervention. More recently, conservatives have started bringing, or joining, cases to further their agenda, such as limiting teacher tenure. Liberals, in response, parrot the causation and evidentiary arguments that conservatives made against school-finance claims.

The evidence-is-too-shaky arguments against judicial intervention have pervasively influenced the body of education-rights law that has developed over the last several decades.[6] Plaintiffs have lost several cases on causation. In other cases, courts used the evidentiary difficulties to justify deferential standards of review or nonjusticiability holdings. Conversely, when courts have intervened, they have often done so in ways that sidestep evidentiary disputes—for example, by reciting laundry lists of "problems" and telling the legislature to make things better (without identifying any intervention that *would* make things better), or by grounding the court's decision on quasi-procedural flaws in the legislature's implementation of the education clauses.

Yet despite the ubiquity and influence of arguments about the evidentiary problem, courts and litigators have largely overlooked an essential point about its nature: *The state's own choices substantially determine whether researchers—and hence litigators—can produce credible evidence concerning the causal effect of state laws and funding arrangements on the outcomes that ground the education right.*[7] States exercise this control through administrative systems for making and linking (or thwarting the linkage of) educational, tax, voting, welfare, criminal justice, and birth-and-death records; through the rules for assigning students, programs, and funding to schools; through the manner in which educational reforms are rolled out; and through the terms on which the state provides researchers with access to administrative data. In short, the information that courts and other state actors need to realize the constitutional objectives of the education system is endogenous to public law.

[3] *See, e.g.*, Jeff Bryant, *Revisiting A Progressive Education Agenda: What's Happened Since?*, EDUCATION OPPORTUNITY NETWORK (Mar. 26, 2015), http://www.commondreams.org/views/2015/03/26/revisiting-progressive-education-agenda-whats-happened.

[4] *See, e.g.*, ERIC A. HANUSHEK & ALFRED A. LINDSETH, SCHOOLHOUSES, COURTHOUSES, AND STATEHOUSES: SOLVING THE FUNDING-ACHIEVEMENT PUZZLE IN AMERICA'S PUBLIC SCHOOLS 76–77 (2009); Edwin J. Feulner, *Education at a Crossroads*, HERITAGE FOUND. (Sept. 30, 2013), https://www.heritage.org/education/commentary/education-crossroads. Regarding fissures among liberals, see, e.g., Lyndsey Layton, *Democrats Divided on Calif. Tenure Ruling*, WASH. POST, June 13, 2014, at A2 (describing responses to California trial court decision invalidating teacher tenure and seniority protections).

[5] For support for the claims in this paragraph, see *infra* Section II.B.

[6] For support for the claims in this paragraph, see *infra* Section II.B.

[7] *See infra* Part III.

The purpose of this Article is to explain this endogeneity—that is, how law affects the production of knowledge about the constitutional quality of educational systems—and to draw out the implications for education-rights jurisprudence. We argue that a clear-eyed appreciation of the information problem points to several new lines of attack for education-rights plaintiffs. In states whose courts have recognized what we term "framework duties" under the education clauses—such as a duty to promulgate educational standards and to establish systems of educational testing and school finance tied to those standards—it is straightforward to argue that the state must also adopt a reasonable plan to facilitate the production of knowledge about how to educate disadvantaged children more effectively. At a minimum, this plan must address the linkage of educational and other administrative databases and the terms on which the state provides access to administrative data. The plan should also identify actual or potential interventions that the state considers promising vis-à-vis the constitutional objectives of the education system and explain how the state will assess, or enable others to assess, the efficacy of those interventions.

Second, in the many states whose courts have adopted arbitrariness or rational-basis-plus standards of review for claims under the education clauses, plaintiffs should be able to attack educational or record-keeping arrangements that hinder the production of knowledge about what works. We illustrate this idea with an example from Oakland, California, where the school board, under union pressure, has resisted adopting a unified lottery for allocating spaces in charter and conventional public schools. Like many other school districts, Oakland presently runs one lottery for conventional public schools and separate lotteries for each charter school. The primary argument for a unified lottery—one in which parents could list any public or charter school—is that it is easier for parents to navigate. An unexpected side benefit is that unified lotteries enable researchers to produce much better estimates of the causal effect of innovative schools on disadvantaged students' outcomes. We argue that this side benefit affords a plausible basis for courts to compel unified lotteries.

Third, in certain cases, courts *should* order—or allow the defendant to elect—*randomized remedies*, pursuant to which plaintiff-sought reforms would be implemented on a temporary basis in a randomly selected subset of schools or school districts. Randomized remedies represent an appropriate resolution for certain hard cases, where plaintiffs can point to danger signs that suggest that the state has failed to exercise reasonable care in providing for the education of disadvantaged students, but where the state's control over policy and funding has made it impossible for the plaintiffs to introduce credible evidence about the causal effect of their proposed remedy.

We readily acknowledge that there are respectable separation-of-powers arguments against courts ordering the adoption of particular educational programs or funding arrangements, whether on a permanent/statewide or temporary/randomized basis. Indeed, some courts have held educational adequacy claims nonjusticiable on the ground that judges ought not supplant decisions of the people and their representatives about what educational outcomes to value and how to pursue them.[8] Importantly, however, the first two types of information-oriented claims we envision (planning requirements and attacks on policies that arbitrarily hinder the production of knowledge about what works) should remain available in these courts.[9] In vindicating such claims, the courts would bring the consequences of policy alternatives into public view, greatly enriching public discourse and democratic decision-making.

The Article proceeds as follows. Part II sets the stage by highlighting the lack of credible social scientific evidence about how to effect large-scale improvements in the constitutional quality of the education provided to disadvantaged children. (It is not the case that judges and litigants have simply missed a body of important social scientific research.) We explain that the lack of such evidence has influenced both individual case rulings and larger doctrinal developments, such as the recognition in some courts of a legislative duty to develop and adhere to a quasi-procedural framework for realizing the constitutional promise of a decent education for all. Part II closes by offering a preliminary normative defense of this "framework" approach, while questioning the courts' failure to require as part of the framework *some* effort by the state to figure out how actual or potential educational reforms affect the most constitutionally weighty student outcomes.

Part III addresses state control over the production of knowledge about how to realize the constitutional goals of the education system. This Part initially explains that for the research enterprise to reliably guide judicial and legislative implementation of the education right, three conditions must be satisfied. First, researchers must be able to observe constitutionally important outcomes and to associate those outcomes with the educational or policy "treatments" each student received. Second, researchers must have credible answers to what statistician Paul Holland famously dubbed "the fundamental problem of causal inference."[10] The fundamental problem is that causal effects are, by definition, differences between the outcomes that a student (or other object of study) would realize under different states of the world, yet counterfactual outcomes are, by definition, unobservable. Third, consumers of education research, such as the experts who advise courts and legislatures, must be able to make reasonable, transparent judgments about whether reported effect sizes and tests of statistical significance in a defined body of work are probably free from publication and related biases. Together, these three conditions comprise what we call the "useful causal research" chain. The main takeaway from Part III is that the strength of each link in the chain critically depends upon the state.

Law returns to the forefront in Part IV, which assesses the legal implications of state control over the production of knowledge. We explain our proposals for a knowledge-production planning requirement; for legal attacks on discrete components of the educational system that arbitrarily hinder the production of constitutionally significant knowledge; and for randomized remedies in certain hard cases. Part IV concludes with some reflections on pragmatic

[8] *See, e.g.*, Comm. for Educ. Rights v. Edgar, 672 N.E.2d 1178, 1191 (Ill. 1996).

[9] This is so because the claims would not displace the legislature's policy-making prerogatives.

[10] Paul W. Holland, *Statistics and Causal Inference*, 81 J. AM. STAT. ASS'N 945, 947 (1986).

grounds for reorienting education-rights jurisprudence toward the production of knowledge about how to educate disadvantaged children effectively.

II. "PROBLEMS NO ONE HAS SOLVED"

Inspired by *Brown v. Board of Education* and the Warren Court's rights revolution, legal scholars in the 1960s began theorizing about the possibility of a judicially discoverable right to education.[11] From the outset, they recognized that social scientific uncertainties about how to measure school quality—and how to improve it—might end up dashing the entire project.[12] A widely celebrated book by Coons, Clune, and Sugarman argued that courts could navigate these tricky shoals by undertaking to equalize school districts' ability to raise revenue, while treating arguments about school quality beyond the judicial ken.[13]

Yet, most courts that went beyond rational basis review of education claims put school quality front-and-center. Some of the decisions were grounded on the equal protection clause of the state constitution;[14] more commonly, courts construed the constitution's education clauses as requiring the state to provide schooling of an "adequate" level of quality. Though "adequacy" and "equality" (a/k/a "equity") represent distinct theories of the education right,[15] many scholars have observed that adequacy and equity cases converge in practice.[16] Regardless of the nominal governing theory, courts stress the importance of education for students' future participation in economic and political life.[17] And in

[11] 347 U.S. 483 (1954). This history is well summarized in William S. Koski, *Of Fuzzy Standards and Institutional Constraints: A Re-Examination of the Jurisprudential History of Educational Finance Reform Litigation*, 43 SANTA CLARA L. REV. 1185, 1188–91, 1260–61 (2003) [hereinafter Koski, *Fuzzy Standards*].

[12] Koski, *Fuzzy Standards*, *supra* note 11, at 1189–91.

[13] JOHN E. COONS, WILLIAM CLUNE & STEPHEN SUGARMAN, PRIVATE WEALTH AND PUBLIC EDUCATION 339–93 (1970).

[14] In California, for example, judges in equal protection cases are to assess the "actual quality of [a school or a district's] program, viewed as a whole," and compare it to "prevailing statewide standards." Butt v. California, 842 P.2d 1240, 1252 (Cal. 1992) (emphasis added).

[15] For an influential argument in favor of adequacy over equity, see Peter Enrich, *Leaving Equality Behind: New Directions in School Finance Reform*, 48 VAND. L. REV. 101, 130 (1995). For a powerful response, see William S. Koski & Rob Reich, *When "Adequate" Isn't: The Retreat from Equity in Educational Law and Policy and Why It Matters*, 56 EMORY L.J. 545, 589 (2006). Note that at least six state supreme courts that ruled against plaintiffs in equity cases later ruled for plaintiffs on adequacies theories. Koski, *Fuzzy Standards*, *supra* note 11, at 1276–77.

[16] *See, e.g.*, Richard Briffault, *Adding Adequacy to Equity*, *in* SCHOOL MONEY TRIALS: THE LEGAL PURSUIT OF EDUCATIONAL ADEQUACY 25, 45 (Martin R. West & Paul E. Peterson eds., 2007) [hereinafter SCHOOL MONEY TRIALS]; *see* Koski, *Fuzzy Standards*, *supra* note 11, at 1187–88; James E. Ryan, *Standards, Testing, and School Finance Litigation*, 86 TEX. L. REV. 1223, 1252 (2007); Julie K. Underwood, *School Finance Adequacy as Vertical Equity*, 28 U. MICH. J.L. REFORM 493, 516–17 (1995).

[17] For equity/equal protection cases characterizing "fundamental" status of the right to education on this basis, see, e.g., Horton v. Meskill, 376 A.2d 359, 372–73 (Conn. 1977); Serrano v. Priest, 487 P.2d 1241, 1255 (Cal. 1971). For adequacy cases making the same point, see, e.g., Campaign for Fiscal Equity, Inc. v. State of New York, 86 N.Y.2d 307, 316 (N.Y. 1995); Rose v. Council for Better Educ., Inc., 790 S.W.2d at 190, 212 (Ky. 1989). These courts generally reference *Brown v. Board of Education*, 347 U.S. 483, 493 (1954) (stressing importance of education "to succeed in life"); *see also* INSTITUTE FOR EDUCATIONAL AND OPPORTUNITY, EDUCATION IN THE 50 STATES: A DESKBOOK OF THE HISTORY OF STATE CONSTITUTIONS AND LAWS ABOUT EDUCATION (2009) (arguing that state constitutional education clauses are universally grounded on perceived importance of education for democracy and economic opportunity).

adequacy cases just as much as equity cases, courts focus on historically disadvantaged and underperforming groups, such as the poor, African Americans, Latinos, and English-language learners.[18] In recognition of these commonalities, we will refer to the cases collectively as "education quality" cases, for the crux of the issue is that certain schools are not *good enough*, either across the board or for certain groups of students.

Given that courts explained the fundamental status of education in terms of economic and political opportunity, one might expect litigation to have unfolded more or less as follows. Plaintiffs would introduce data showing that students in the plaintiffs' socioeconomic circumstances and assigned to the plaintiffs' school have lousy odds of achieving decent lifetime outcomes (employment, democratic participation, life-years not incarcerated, a stable family, etc.). The plaintiffs would also present social scientific studies showing concrete steps that the state could institute to improve the plaintiffs' schools and, concomitantly, plaintiffs' likely outcomes. Defendants might attack these studies or present contrary research findings. If the court were convinced by the plaintiffs' evidence, the court would then determine whether the state's failure to adopt the plaintiff-sought reforms was supported by sufficiently weighty reasons, in light of the constitutional values at stake. If the state's justification came up short, the court would find the state liable and order the state either to implement the plaintiff-sought reforms, or to allow the plaintiff-students to transfer to different, constitutionally compliant schools nearby.

But the law did not develop in this way—inevitably. As the New Jersey Supreme Court put it, the courts were being asked to solve educational problems that "[n]o one ha[d] solved," and to do so in the absence of any generally accepted measure of school quality.[19]

The purpose of this Part is to briefly sketch the underlying problem, *i.e.*, the lack of information about how to measure and improve the constitutional quality of the schools, and then explain what the courts have done in the absence of this information. This will set up our analysis, in Part III, of (state-controlled) barriers to answering the empirical questions on which education-quality cases *ought* to turn, given the conception of the right as one entitling children to an

[18] Though state courts have generally said that the education clauses guarantee only an adequate opportunity to be educated, not outcomes, persistently bad educational outcomes for historically disadvantaged groups have often been treated as evidence of a constitutional violation. *See, e.g.*, Morath v. Texas Taxpayer & Student Fairness Coal., 490 S.W.3d 826, 850 (Tex. 2016) (stating that adequacy standard is "result-oriented"); Abbeville Cty. Sch. Dist. v. State, 767 S.E.2d 157, 167–69 (S.C. 2014) (finding constitutional violation based on persistently poor outcomes for students in the plaintiff districts, which served a disproportionate number of poor, African American students, notwithstanding that "[t]he instrumentalities of learning—funding, curriculum, teachers, and programs—are present and appear at the very least minimally adequate"); Gannon v. State, 319 P.3d 1196, 1236–37 (Kan. 2014) (holding that the constitution requires an educational system "reasonably calculated to have *all* Kansas public education students meet or exceed [certain educational] standards") (emphasis added); Montoy v. State, 112 P.3d 923, 939 (Kan. 2005) ("outputs are necessary elements of a constitutionally adequate education . . . ."); Hoke Cty. Bd. of Educ. v. State, 599 S.E.2d 365, 380 (N.C. 2004) (endorsing opportunity-not-outcomes principle but upholding trial court finding of unconstitutionality based largely on outcomes).

[19] Abbott v. Burke, 575 A.2d 359, 363 (N.J. 1990); *see also* Robinson v. Cahill, 303 A.2d 273, 295 (N.J. 1973) ("We deal with the problem [on the basis of discrepancies in dollar input per pupil] because dollar input is plainly relevant and because we have been shown no other viable criterion for measuring compliance with the constitutional mandate.").

education that appropriately prepares them for future participation in economic, civic, and political life.

A. *Knowns and Known Unknowns: Schools and Socioeconomic Mobility*

Because state courts have explained the fundamental status of education in terms of economic mobility and political participation, it is worth situating our discussion of the education literature within the larger body of research on intergenerational mobility.[20] The best available evidence suggests that the vision of economic opportunity that informs the education right is realized today in some parts of the United States, and not in others. Using administrative data from the Internal Revenue Service, leading economist Raj Chetty and co-authors have shown that the United States is not a uniform land of opportunity, but rather "a collection of societies" in which children "escape from poverty" at very different rates.[21] "Some [localities] have . . . [intergenerational socioeconomic] mobility comparable to the highest mobility countries in the world, such as Canada and Denmark, while others have lower levels of mobility than any developed country for which data are available."[22]

Chetty and Hendren observe that this geographic heterogeneity can be decomposed into *sorting* and *causal* components—a critically important distinction and one to which we will return in Part III.[23] *Sorting* occurs when parents who make relatively large investments in their children, or whose children are otherwise likely to climb the economic ladder for reasons independent of location, clump together in the same geographies. *Causality* in this context means that something about the society in a geographic area—perhaps public policies, institutions, or culture—results in children in that area achieving better outcomes than they would have achieved if they were brought up elsewhere.

"Sorting" and "causality" are difficult to disentangle without controlled experiments,[24] but by comparing siblings who moved to high-mobility zones at different ages and by isolating people who were displaced from a commuting zone by economic shocks, Chetty and Hendren estimated that one-half to two-thirds of the geographic variation is probably causal.[25] Children who had the misfortune of growing up in Baltimore, Maryland (one of the worst places for

[20] There has been much less work on intergenerational *political* mobility, but, intriguingly, several recent, well-designed studies on educational interventions in the early childhood of poor children found a payoff in terms of future political participation. *See* Rachel Milstein Sondheimer & Donald P. Green, *Using Experiments to Estimate Effects of Education on Voter Turnout*, 54 AM. J. POL. SCI. 174, 185 (2010); John B. Holbein, Making Good Citizens: Policy Approaches to Increasing Civic Participation (2016) (unpublished Ph.D. Dissertation, Dept. of Pub. Pol'y, Duke University).

[21] Raj Chetty et al., *Where is the Land of Opportunity? The Geography of Intergenerational Mobility in the United States*, 129 Q.J. ECON. 1553, 1554 (2014) [hereinafter Chetty et al., *Where is the Land of Opportunity?*]. The geographic localities in this study are census-defined "communizing zones." *Id.* at 1555–56.

[22] *Id.* at 1553, 1556 (citing *Community Zone and Labor Market Areas,* DEP'T OF AGRIC.–ECON. RESEARCH SERV., https://catalog.data.gov/dataset/commuting-zones-and-labor-market-areas (last updated June 20, 2017)).

[23] Raj Chetty & Nathaniel Hendren, *The Impacts of Neighborhoods on Intergenerational Mobility: Childhood Exposure Effects and County-Level Estimates* (Harv. Univ. and NBER, Working Paper, May 2015), http://scholar.harvard.edu/files/hendren/files/nbhds_paper.pdf.

[24] *See infra* Section III.B.

[25] Chetty & Hendren, *supra* note 23, at 2–4.

socioeconomic mobility) would have done much better if their parents had relocated to DuPage County, Illinois (one of the best).[26]

Chetty and Hendren's estimates also shed some light on the conjecture that school quality matters for intergenerational mobility. An income-adjusted measure of student performance on standardized tests is highly correlated with Chetty and Hendren's estimate of the causal effect of "the commuting zone" on socioeconomic mobility.[27] What this means is that counties where public-school students outperform the national average *of students with the same family income* are also counties in which a poor child is likely to realize better socioeconomic outcomes than he probably would somewhere else.

But for courts or other policy-makers to use this result to implement the education right, they need to know more. Specifically: Are the school "effects" causal? Are the school effects substantively, as well as statistically, significant? What exactly is it that good schools do differently than bad schools? And what policy levers can be used to transform or replace the bad schools?

Do "Good" Schools Cause Socioeconomic Mobility? It seems plausible that schools are an important part of the mechanism by which counties that have a big effect on intergenerational mobility achieve this effect. But it might equally be the case that something outside of the schools—say, a county-level neonatal program or community institutions like churches—causes *both* the impressive test score results and the high level of socioeconomic mobility. If schools are not the mobility mechanism, it is doubtful the state should be faulted for failing to maintain "adequate" or "equal" schools, particularly if the state is doing other things to enable mobility.

Substantive vs. Statistical Significance. Chetty's studies of intergenerational mobility are based on an enormous dataset with more than 40 million records.[28] With big datasets, even tiny correlations are often statistically distinguishable from zero. But a tiny correlation between school quality and socioeconomic mobility probably would not justify big policy interventions to improve the schools. Chetty and Hendren estimated that for children in families with below-median income, "moving to a [locality] with a 1 standard deviation higher [school quality] causes . . . a 4.2% increase in incomes at age 26," if the child spends a full twenty years growing up in the better locale.[29] Is this effect big or small? To some extent, that judgment is in the eye of the beholder. But it also clearly depends on what it would cost to "buy" a one-standard-deviation improvement in school quality.

Making More Good Schools. Assuming the good schools really do make a difference, there remains the question of how to create more of them. For courts to say that the state has failed to provide disadvantaged students with an adequate education, it would seem necessary to identify features or attributes of a school or school system which, when present, result in disadvantaged students

[26] *Id.* at 6.

[27] *Id.* at 76–77.

[28] Chetty et al., *Where is the Land of Opportunity?*, *supra* note 21, at 1554.

[29] Chetty & Hendren, *supra* note 23, at 75.

realizing better outcomes than they otherwise would. It would also seem necessary to establish some practical means by which the state could ensure that those features are present. If no such means exist, the state cannot be said to have done anything wrong.

These questions bedevil researchers, educators, litigators, and courts. Grover Whitehurst, the first director of the Institute for Education Sciences ("IES") within the U.S. Department of Education, nicely described the lay of the land in the early 2000s, following the passage of the No Child Left Behind Act ("NCLB"), which required that schools undertake reforms based on "scientific[] research"[30]:

> A chief state school officer or a district superintendent would pull me aside and ask what scientific research had to say about professional development, or effective mathematics curriculum, or . . . (fill in the blank). "I'm desperate to meet the student proficiency requirements of NCLB. Just tell me what the research says works and I'll do it" was the gist of their position. For most such requests I had to say that I didn't know of any decent research that would be helpful.[31]

Thanks in part to the work of the IES, the picture is not quite so bleak today.[32] Researchers are making some headway on the question of what makes good schools good, particularly for disadvantaged children. But the unfortunate reality is that we still do not know very much about the *causal effects* of various educational and school reform interventions on the *adult outcomes* of disadvantaged students. Why this is so is the subject of Part III. For now, it is enough to convey the following stylized facts, which, together with the Chetty et al. results, comprise the empirical backdrop for education-quality litigation today:

- A number of educational interventions piloted on a small scale have improved disadvantaged students' educational attainment and graduation rates. Some of the more promising interventions include intensive pre-kindergarten programs,[33] class-size reduction,[34] extension of the school day or school calendar,[35] and certain pedagogical practices pioneered by urban

[30] SASHA ZUCKER, SCIENTIFICALLY BASED RESEARCH: NCLB AND ASSESSMENT, PEARSON EDUC. POLICY REPORT (Mar. 2004) ("A significant aspect of the No Child Left Behind Act of 2001 (NCLB) is the use of the phrase 'scientifically based research' well over 100 times throughout the text of the law.").

[31] Grover J. Whitehurst, *The Value of Experiments in Education*, 7 EDUC. FIN. & POL'Y 107, 109 (2012).

[32] The IES's "What Works Clearinghouse," is an excellent online archive of quality research about the causal effects of educational interventions. *Welcome to the What Works Clearinghouse*, INST. FOR EDUC. SCI., http://ies.ed.gov/ncee/wwc/ (last visited Jan. 16, 2018) [hereinafter *What Works Clearinghouse*].

[33] *See* SNEHA ELANGO ET AL., ECONOMICS OF MEANS-TESTED TRANSFER PROGRAMS IN THE UNITED STATES, VOLUME 2 56–58 (Robert A. Moffitt ed., 2016); W.S. Barnett, *Effectiveness of Early Educational Intervention*, SCI., Aug. 19, 2011, at 977; Greg J. Duncan & Katherine Magnuson, *Investing in Preschool Programs*, J. ECON. PERSP., Spring 2013, at 112.

[34] *See* DIANE W. SCHANZENBACH, NAT'L EDUC. POL'Y CTR., DOES CLASS-SIZE MATTER? (Feb. 2014), http://nepc.colorado.edu/files/pb_-_class_size.pdf.

[35] *See, e.g.*, DAVID A. FARBMAN, NAT'L CTR. ON TIME & LEARNING, THE CASE FOR IMPROVING AND EXPANDING TIME IN SCHOOL: A REVIEW OF KEY RESEARCH AND PRACTICE 11 (Feb. 2015), http://www.timeandlearning.org/sites/default/files/resources/caseformorelearningtime.pdf; MATTHEW A. KRAFT, HOW TO MAKE ADDITIONAL TIME MATTER: INTEGRATING INDIVIDUALIZED TUTORIALS INTO AN EXTENDED DAY 3 (Apr. 2013), http://scholar.harvard.edu/files/mkraft/files/kraft_-_how_to_make_additional_time_matter.pdf?m=136517 4163s.

charter schools.[36] Yet, no *state* has achieved big, sustained improvements at scale. For example, a recent study found that only 13–14% of the geographic variation among school districts, cities, or counties in the black-white or Latino-white achievement gap represents between-state variation.[37] Thus, in terms of these achievement gaps, each state looks pretty much the same—or 86–87% the same—as every other state.

- School-achievement gaps by socioeconomic status have grown over the last fifty years,[38] even as the state and federal governments have substantially increased their investments in the education of disadvantaged children.[39] Among entering kindergarteners, the school-readiness gap has become modestly smaller recently, but it remains unclear whether this is due to pre-K programs or changes in parenting norms.[40]
- School districts that primarily serve disadvantaged populations used to spend much less per student than districts that serve middle-class populations.[41] Over the last forty years, the between-district spending gap has closed, or even reversed, in many states.[42] The gap remains large in some states, however,[43] and *within-district* disparities in spending on rich and poor students are generally substantial.[44]
- Many studies find little effect of educational spending on student outcomes,[45] but the null results may be due to some unobserved factor that is correlated with spending and outcomes (biasing the estimated effect of spending on outcomes).[46] Estimates of what a state would need to spend in

[36] *See generally* Roland G. Fryer, Jr., *Injecting Charter School Best Practices into Traditional Public Schools: Evidence from Field Experiments*, 129 Q.J. ECON. 1355 (2014). *But see* Atila Abdulkadiroğlu et al., *Accountability and Flexibility in Public Schools: Evidence from Boston's Charters and Pilots*, 126 Q.J. ECON. 699, 745 (2011) (finding that students randomly assigned to union-supported "pilot school" alternative to charter schools failed to realize education gains on par with students randomly assigned to charters).

[37] Sean F. Reardon et al., *The Geography of Racial/Ethnic Test Score Gaps* 25 (Ctr. for Educ. Pol'y Analysis, Working Paper No. 16-10, 2016), https://cepa.stanford.edu/sites/default/files/wp16-10-v201604.pdf.

[38] Sean F. Reardon, *The Widening Academic Achievement Gap Between the Rich and the Poor: New Evidence and Possible Explanations*, *in* WHITHER OPPORTUNITY?: RISING INEQUALITY, SCHOOLS, AND CHILDREN'S LIFE CHANCES 91 (Greg J. Duncan & Richard J. Murnane eds. 2011) ("The achievement gap between children in high- and low-income families is roughly 30 to 40 percent larger among children born in 2001 than among those born twenty-five years earlier . . . [and] the income achievement gap has been growing steadily for at least fifty years.").

[39] Regarding increases in spending on education, see HANUSHEK & LINDSETH, *supra* note 4, at 45, 57–58 (explaining that public-school spending on primary and secondary education in the United States has roughly quadrupled in real terms since 1960 and has also become more equitably distributed across districts with high concentrations of poor and minority students).

[40] Sean F. Reardon & Ximena A. Portilla, *Recent Trends in Income, Racial, and Ethnic School Readiness Gaps at Kindergarten Entry*, AM. EDUC. RESEARCH ASS'N OPEN, July–Sept. 2016, at 13–14.

[41] *See* HANUSHEK & LINDSETH, *supra* note 4, at 57–70; MARGUERITE ROZA, EDUCATION ECONOMICS: WHERE DO SCHOOL FUNDS GO? (2010).

[42] *See* Reardon & Portilla, *supra* note 40, at 3.

[43] *See* NATASHA USHOMIRSKY & DAVID WILLIAMS, THE EDUCATION TRUST, FUNDING GAPS 2015: TOO MANY STATES STILL SPEND LESS ON EDUCATING STUDENTS WHO NEED THE MOST 1 (Mar. 2015), http://edtrust.org/wp/content/uploads/2014/09/FundingGaps2015_TheEducationTrust1.pdf.

[44] *See* ROZA, *supra* note 41.

[45] *See* Julien Lafortune et al., *School Finance Reform and the Distribution of Student Achievement* 2 (Nat'l Bureau of Econ. Research, Working Paper No. 22011, 2016), http://www.nber.org/papers/w22011.pdf (reviewing literature).

[46] *Id.* at 29–30 (discussing risk of downward bias in observational studies of effects of school—bias which may result if governments distribute school funds where they are most needed). The best and most recent

order for schools to meet the state's proficiency standards are extremely sensitive to modeling assumptions.[47] Methods commonly deployed in litigation have yielded estimates that diverge by a factor of ten or more.[48]

- Even when researchers agree on the importance of an educational input, it is often unclear what policy reforms would more effectively distribute that resource to disadvantaged students. For example, researchers concur that teacher quality matters a lot for student outcomes, but there is no established recipe for getting disadvantaged students into classrooms with good teachers.[49] Liberals recommend more spending on teacher salaries and professional development.[50] Conservatives urge more discretion for administrators to fire bad teachers.[51] Both answers are speculative.
- The vast majority of educational research has used test scores, dropout rates, or other near-term outcome measures as the dependent variable.[52] Studies like Chetty's that examine adult outcomes—the outcomes that matter constitutionally—remain uncommon.[53]

Given these stylized facts, what is a court to do when plaintiffs argue that the state has failed poor students and that the education or equal protection clauses require the state to spend more on those students, or to make class sizes smaller, or to reform teacher tenure and seniority protections so that bad teachers can be fired more easily, or to reallocate funds from school-infrastructure projects to other educational programs, or to establish free pre-K for disadvantaged children, or to make it easier for charter schools to form? These questions have

work on the relationship between educational spending and student outcomes finds that the *predicted* increase in spending following a ruling for school-finance plaintiffs is positively correlated with disadvantaged students' future outcomes as adults. *See* C. Kirabo Jackson et al., *The Effects of School Spending on Educational and Economic Outcomes: Evidence from School Finance Reforms*, 131 Q.J. ECON. 157 (2016); *see also* Lafortune et al., *supra* note 45 (using a difference-in-difference approach to estimate effect of state-level finance reforms on student achievement and finding positive average effects for students in low-income *school districts* but not for *low-income or minority students*, whose distribution across districts is not strongly correlated with average district income). But these research designs cannot rule out the possibility that other reforms which tend to happen at the same time as judicial interventions are the cause of the improvement.

[47] These estimates are often produced for litigation, through so-called "costing out studies." For discussion of the standard methods, and the sensitivity of results to modeling assumption, see HANUSHEK & LINDSETH, *supra* note 4, at 178–97; Thomas A. Downes & Leanna Stiefel, *Measuring Equity and Adequacy in School Finance*, *in* HANDBOOK OF RESEARCH IN EDUCATION FINANCE AND POLICY 244, 248–53 (Helen F. Ladd & Edward B. Fiske eds., 2d ed. 2008); Shawna Grosskopf et al., *Efficiency in Education: Research and Implications*, 36 APPLIED ECON. PERSP. & POL'Y 175, 179 (2014).

[48] Jennifer Imazeki, *Assessing the Costs of Adequacy in California Public Schools: A Cost Function Approach*, 3 EDUC. FIN. & POL'Y 90, 97–103 (2008) (reporting cost-function estimate of $47 billion versus production-function estimate of more than $1.5 trillion, using California data).

[49] *See generally* William S. Koski, *Bridging the Teacher Quality Gap: Notes from California on the Potential and Pitfalls of Litigating Teacher Quality*, *in* THE ENDURING LEGACY OF RODRIGUEZ: CREATING NEW PATHWAYS TO EQUAL EDUCATIONAL OPPORTUNITY (Charles J. Ogletree, Jr. & Kimberly Jenkins Robinson eds. 2015); Derek W. Black, *The Constitutional Challenge to Teacher Tenure*, 104 CALIF. L. REV. 75 (2016).

[50] *See* Black, *supra* note 49, at 88, 90 (citing sources).

[51] *See* Koski, *Bridging the Teacher Quality Gap*, *supra* note 49, at 164.

[52] Many researchers have commented on this problem. *See, e.g.*, Drew Bailey et al., *Persistence and Fadeout in the Impacts of Child and Adolescent Interventions* 2 (Lifecourse Centre, Working Paper No. 2015-27, 2015) ("Most interventions targeting children's cognitive, social or emotional development fail to follow their subjects beyond the end of their programs.")

[53] Happily, this number has started to grow, enabled by developments we discuss *infra* Section III.A.

all been litigated in recent years.[54] The next Section briefly summarizes the judicial response.

B. *Intervening (or Not) for Quality Without Knowing How to Achieve It*

Since education-rights litigation got underway in the early 1970s, state constitutional cases have been brought in forty-four of the fifty states,[55] yielding a hugely varied body of law. To motivate the rest of this Article, we need only convey a couple of points about the law. First, evidentiary uncertainties have thwarted many claims. Second, when courts have intervened, they have usually done so in ways that bracket the question of whether particular reforms would materially improve student outcomes.

In seven cases over roughly the last decade, courts vindicated causation arguments against school-finance claims.[56] Other courts rejected new-fangled challenges to teacher tenure and seniority protections, and charter school limits, on causation grounds.[57] This fits into a larger pattern. As Derek Black has shown, causation defenses have stymied a great variety of statutory and constitutional claims by plaintiffs seeking a better, more equal, or more racially integrated education.[58]

[54] See cases cited *infra* Section II.B.

[55] *See* SCHOOLFUNDING.INFO, http://schoolfunding.info/ (last visited Jan. 16, 2018) (hyperlinked map showing court decisions).

[56] Some courts responded favorably by dismissing the claim. *See* Davis v. State, 804 N.W.2d 618, 636–41 (S.D. 2011) (relying on studies showing weak correlation between spending and test scores and lack of apparent improvement in test scores following judicial intervention in New Jersey and Wyoming); Espinoza v. Ariz. State Bd. of Educ., CV-2006-005616, at *10 (Ariz. 2008) ("Plaintiffs have failed to show a causal connection between programs available or unavailable at any given school district and the failure of Plaintiffs in those districts to pass the [exit exam] test."). Others reached the merits but found in whole, or in part, for the defendants because of the uncertain relationship between spending and student outcomes. *See* Morath v. Tex. Taxpayer & Student Fairness Coal., 490 S.W.3d 826, 851 (Tex. 2016) ("[T]he trial court's 'fact' findings as to the specific amount of funding needed to achieve a general diffusion of knowledge are, we think, beyond the current state of science in this field."); Columbia Falls v. State, No. BDV-2002-528, 2008 Mont. Dist. LEXIS 483, at * 72 (Dec. 15, 2008) ("[T]he Court is unclear whether the problems currently experienced by the Plaintiff districts are a result of a constitutionally inadequate funding system or by choices made by the school districts."); Moore v. State, No. 3AN-04-9756CI, at *140–41 (Sup. Ct. Alaska, June 21, 2007) (crediting quantitative and qualitative evidence about weak correlation between spending and test scores, and concluding, "Plaintiffs have failed to demonstrate that the state is inadequately funding public education . . . ."); Cranston Sch. Comm. v. City of Cranston, No. PC/03-5110, 2004 WL 603408, at *6 (R.I. Super. Ct. Mar. 8, 2004) ("Student and school performance measures are arguably relevant [to plaintiffs' case] if a correlation was established between the additional funds sought and how those funds would be allocated in order to achieve . . . results[, but t]he record is devoid of reference as to where these funds would be allocated or how additional funding would achieve the performance standards . . . ."); Hoke Cty. Bd. of Educ. v. State, No. 95CVS1158, 2000 WL 1639686, at *56–57 (N.C. Super. Oct. 12, 2000), *aff'd in part as modified, rev'd in part sub nom,* Hoke Cty. Bd. of Educ. v. State, 599 S.E.2d 365 (N.C. 2004) (agreeing with defendant that additional spending on K-12 schools was unlikely to make a difference, but finding that additional spending on early childhood education would have positive effect on student outcomes).

[57] *See, e.g.*, *Morath*, 490 S.W.3d at 878, (rejecting claim for lifting charter school limits because, *inter alia*, "Intervenors did not quantify the inefficiencies about which they complain, in dollar terms or through some other quantitative measure, or the efficiencies that would be achieved with the additional competition they advocate."); Vergara v. State, 209 Cal. Rptr. 3d 532, 554–57 (Ct. App. 2016) (rejecting facial challenge to teacher tenure and dismissal statutes because plaintiffs failed to show that these statutes "inevitably" resulted in disadvantaged students being taught by a disproportionate number of lousy teachers).

[58] Derek W. Black, *Civil Rights, Charter Schools, and Lessons to be Learned*, 64 FLA. L. REV. 1723, 1731–57 (2012) [hereinafter *Civil Rights and Charter Schools*] (discussing federal constitutional desegregation

At a doctrinal level, courts have often treated the evidentiary problems as a reason to adopt deferential standards of review. A leading example is the U.S. Supreme Court's *San Antonio Independent School District v. Rodriguez* decision, which held that school-finance inequalities trigger only rational basis review under the federal Constitution.[59] Most state courts have followed *Rodriguez* when interpreting the equal protection clauses of state constitutions.[60] Similarly, evidentiary doubts have underwritten state court decisions, adopting deferential standards of review in adequacy cases, or deeming the claims nonjusticiable.[61]

Meanwhile, courts that find for the plaintiffs typically avoid questions about whether particular reforms would materially improve the education of disadvantaged children—let alone whether the incremental educational benefits of plaintiff-urged reforms would outweigh any countervailing state interests.[62] Roughly speaking, two templates have emerged for how to do this.

The first is exemplified by the Kentucky Supreme Court's *Rose v. Council for Better Education* decision,[63] which launched modern educational-adequacy litigation.[64] The *Rose* court began by defining the constitutional standard of adequacy in terms of qualitative criteria cast at a high level of generality, such as "sufficient oral and written communication skills to enable students to function in a complex and rapidly changing civilization."[65] But the court did not proceed to identify particular schools or school districts that were falling short of the standard, or particular state actions that, if implemented, would bring those schools or districts up to par. Nor did the court articulate a standard of review.

Instead, the *Rose* court simply referenced everything in the record that tended to make the state's schools look bad by comparison to schools elsewhere.[66] (Typical factors in *Rose*-style opinions include: disparities among

cases, state constitutional school-finance cases, and cases under federal statues designed to aid disabled children, English language learners, and children stuck in a bad school).

[59] 411 U.S. 1, 55 (1973) ("We are unwilling to assume for ourselves a level of wisdom superior to that of legislators, scholars, and educational authorities in 50 States, especially where the alternatives proposed are only recently conceived *and nowhere yet tested*.") (emphasis added).

[60] *See, e.g.*, Lujan v. Colo. State Bd. of Educ., 649 P.2d 1005, 1018 (Colo. 1982); Koski, *Fuzzy Standards*, *supra* note 11, at 1252 (summarizing cases).

[61] *See, e.g.*, *Morath*, 490 S.W.3d at 886 ("[O]ur lenient[, arbitrariness] standard of review in this policy-laden area counsels modesty."); Comm. for Educ. Rights v. Edgar, 672 N.E.2d 1178, 1191 (Ill. 1996) ("Judicial determination of the type of education children should receive and how it can best be provided would[, if the adequacy claim were justiciable,] depend on . . . whatever expert witnesses the litigants might call . . . . Members of the general public, however, would be obliged to listen in respectful silence."); Koski, *Fuzzy Standards*, *supra* note 11, at 1260–61 (summarizing cases).

[62] As two veterans of school-finance litigation observed, some courts seem to have held the state strictly liable for poor student performance. *See* HANUSHEK & LINDSETH, *supra* note 4, at 105.

[63] 790 S.W.2d 186 (Ky. 1989).

[64] Though *Rose* was an adequacy case, the same template has been used in equity cases. *See, e.g.*, Abbott v. Burke, 575 A.2d 359, 407 (N.J. 1990) (finding school system unconstitutional on the basis of a holistic comparison of educational funding and programs in plaintiff school districts and certain affluent school districts). As Koski and others have observed, the adequacy and equity theories converge in practice. *See supra* notes 15–18 and accompanying text.

[65] *Rose*, 790 S.W.2d at 212.

[66] *Id.* at 197–99. As the court explained, "[t]he evidence in this case consists of numerous depositions, volumes of oral evidence heard by the trial court, and a seemingly endless amount of statistical data, reports, etc. . . . The tidal wave of the appellees' evidence literally engulfs that of the appellants." *Id.* at 196–97.

states, districts, schools, or demographic groups in test-score results and graduation rates [so-called "output" measures of education quality]; analogous disparities in "inputs," such as funding, teacher salaries, curricula, facilities, and class sizes; government reports on problems with the school system; and evidence of legislative inattention to the problems, sometimes expressed as the legislature's failure to commission a study estimating the cost of achieving the state's educational standards.[67]) Then, having found the state liable, the *Rose* court remanded to the *legislature* for a remedy.[68]

This is a way for courts to put pressure on the legislature, or to provide cover for a reform coalition in the legislature,[69] without resolving any of the difficult empirical and conceptual questions about how best to measure the quality of education being provided to the state's children, or about what funding or policy reforms would most likely improve the education of disadvantaged children. Yet it is only a temporary solution. If the legislated remedy does not satisfy the plaintiffs, the court will have to decide whether to dig into the programmatic weeds of school reform—as a handful of courts eventually did[70]—or to defer to legislative and administrative judgments.[71]

The other template for avoiding the what-works conundrum is to reorient the judicial inquiry from the question of whether the plaintiff schools or school districts are qualitatively good enough, to questions about whether the state legislature established a *reasonable framework* for implementing the education right. Thus, courts have instructed the legislature to "defin[e] or giv[e] substantive content to 'basic education,'"[72] within the meaning of the state constitution; to provide for student testing calibrated to the legislature's gloss on the constitutional standard; and to establish accountability mechanisms.[73] A number of

[67] *See generally* Ryan, *supra* note 16, at 1233 ("Time and again, courts have focused on disparities in funding, curricular and extracurricular offerings, qualified teachers, school facilities, and instructional materials."); Koski, *Fuzzy Standards*, *supra* note 11, at 1230–76 (explaining similarity between Rose and earlier and later equity and adequacy decisions).

[68] *Rose*, 790 S.W.2d at 216. Though concurring and dissenting Justices in *Rose* criticized the legislative remand remedy as a judicial abdication of responsibility, see *id.* at 216–18 (Grant, J., concurring, Wintersheimer, J., concurring, and Leibson, J., dissenting), or impermissibly advisory, see *id.* at 223–25 (Leibson, J., dissenting), the legislative remand is the now-standard initial remedy in school finance cases. *See* Koski, *Fuzzy Standards*, *supra* note 11, at 1241 ("[I]n all nineteen final state supreme court educational finance decisions that favored plaintiffs, the courts issued declaratory relief and ordered the legislature to develop a remedial finance scheme.").

[69] Koski, *Fuzzy Standards*, *supra* note 11, at 1271–72 (summarizing reasons to believe that the *Rose* decision itself is an example of informal institutional coordination).

[70] *See* Benjamin Michael Superfine, *New Directions in School Funding and Governance: Moving from Politics to Evidence*, 98 KY. L.J. 653, 668 (2010) (citing and discussing cases).

[71] In Massachusetts and New York, state courts of last resort reversed trial judges who had reviewed legislative responses to the initial liability ruling *de novo*, rather than according broad deference to the political branches. *See* Campaign for Fiscal Equity, Inc. v. State, 861 N.E.2d 50, 57 (Cir. Ct. App. N.Y. 2006) ("The role of the courts is not, as [the trial court] assumed, to determine the best way to calculate the cost of a sound basic education in New York City schools, but to determine whether the State's proposed calculation of that cost is rational."); Hancock v. Comm'r of Educ., 822 N.E.2d 1134, 1140 (Mass. 2005) (Marshall, C.J., concurring) (reversing trial court determination that legislative response was insufficient because plaintiffs failed to show that the legislature had acted in an "arbitrary, nonresponsive, or irrational way to meet the constitutional mandate").

[72] Seattle Sch. Dist. No. 1 of King Cty. v. State, 585 P.2d 71, 95 (Wash. 1978).

[73] *Id.* at 71; *see also* Moore v. State, No. 3AN-04-9756CI, at *174–88 (Sup. Ct. Alaska, June 21, 2007) (holding that legislature's constitutional duty has four components: promulgating "rational educational standards"; establishing an "adequate method of assessing whether children are actually learning what is set out in

courts have also told the legislature to commission a study estimating the cost of meeting state educational standards, and to establish and justify a school-funding formula in light of that study.[74] The legislature may not shift school funding around willy-nilly in response to interest group or local government pressures.[75] In a nutshell, the legislature must develop a plan to implement the education clauses of the constitution and stick to the plan absent a good reason to deviate.[76]

C. A Partial Defense of the Framework Decisions

We think there is promise in the framework decisions, particularly when coupled with a deferential (but not toothless) standard of review that requires the framework to be substantively non-arbitrary. This approach honors the legislature's important and often textually conferred role in the education domain, while also treating the legislature as a fiduciary that owes the state's children a duty of care.[77] As Scott Bauries has pointed out, the fiduciary duty of care in the

the standards"; ensuring "adequate funding so as to accord schools the ability to provide instruction in the standards"; and providing for "adequate accountability and oversight"); Londonderry Sch. Dist. SAU No. 12 v. State, 907 A.2d 988, 993, 995 (N.H. 2006) (ordering legislature to define constitutional standard of quality); Columbia Falls Elementary Sch. Dist. No. 6 v. State, 109 P.3d 257, 263 (Mont. 2005) (faulting legislature for not defining quality); Claremont Sch. Dist. v. Governor, 795 A.2d 744, 751–58 (N.H. 2002) (deeming the accountability system constitutionally inadequate); DeRolph v. State, 728 N.E.2d 993, 1019–20 (Ohio 2000) (discussing need for standards and standards-aligned assessment tests); Hoke Cty. Bd. of Educ. v. State, No. 95CVS1158, 2000 WL 1639686, at *69 (N.C. Super. Ct. Oct. 12, 2000), *aff'd in part as modified, rev'd in part sub nom*, Hoke Cty. Bd. of Educ. v. State, 599 S.E.2d 365, 396 (N.C. 2004) (reading constitution to require accountability system); Abbott v. Burke, 153 710 A.2d 450, 515 (N.J. 1998), *opinion clarified sub nom*, Abbott ex rel. Abbott v. Burke, 751 A.2d 1032 (N.J. 2000) (addressing accountability systems). *See also* Hancock v. Comm'r of Educ., 822 N.E.2d 1134, 1144 (Mass. 2005) (upholding educational system now that "objective, data-driven assessments of student performance and specific performance goals . . . inform a standardized education policy and direct the Commonwealth's public education resources").

[74] *See, e.g.*, McCleary v. State, 269 P.3d 227, 253–61 (Wash. 2012) (finding school-finance system unconstitutional because legislature had established new performance standards without concurrently updating the funding rules to reflect those standards); Montoy v. State, 102 P.3d 1160, 1164 (Kan. 2005) (explaining that the failure to do any cost analysis and to provide for funding based on such an analysis demonstrates the irrationality of the existing school finance system); Columbia Falls Elementary Sch. Dist. No. 6 v. State, 109 P.3d 257, 262 (Mont. 2005) (faulting legislature for not "link[ing] the [school funding] formula to any factors that might constitute a 'quality' education"); Tenn. Small Sch. Systems v. McWherter, 91 S.W.3d 232, 233–34 (Tenn. 2002) (invalidating school-funding funding formula because it "contains no mechanism for cost determination or annual cost review of teachers' salaries"); State v. Campbell Cty. Sch. Dist., 19 P.3d 518, 526 (Wyo. 2001) (ordering legislature to provide for cost studies updated every five years); DeRolph v. State, 677 N.E.2d 733, 738 (Ohio 1997) (invalidating school-finance system because, inter alia, the "formula amount" was a "budgetary residual," rather than an amount determined on the basis of an estimate of "what it actually costs to educate a pupil").

[75] Conn. Coalition for Justice in Educ. Funding, Inc. v. Rell, No. X07 HHD CV 145037565 S, slip op. at 1 (Conn. Super. Ct. Sept. 7, 2016).

[76] To be clear, the quasi-procedural, legislative-duty theory of liability reflected in the "framework" decisions is not always clearly distinguished from the more substantive, school-quality theory of liability manifested in *Rose*-style opinions. Many of the opinions about framework duties cited *supra* notes 73–74, also include findings about the state's failure to educate disadvantaged students effectively. It is often unclear whether such findings serve as flourishes, as triggering conditions for the legislative duty to promulgate a reasonable framework, or as independent and potentially ongoing bases for liability insofar as the framework fails to generate better outcomes.

[77] For a textual and historical defense of the "fiduciary duty" conception of the education clauses, see Scott R. Bauries, *The Education Duty*, 47 WAKE FOR. L. REV. 705, 719–25 (2012) (summarizing recent academic analysis of these provisions).

private-law context requires "both information gathering and rationality."[78] Translated into the public education setting, substantive rationality might be thought to entail *some* set of proficiency standards and a funding formula shown on the basis of a record to be reasonably calculated to enable schools to achieve those standards. As for information gathering, it is not a big leap to suppose that the fiduciary must at least measure student achievement and determine whether the state's proficiency standards have been met.

To be sure, it does not follow from the existence of a duty of care with respect to education that the courts ought to have much of a role enforcing it. Well-rehearsed institutional-competence and separation-of-powers arguments militate against judicial superintendence of educational policy and budgets. Cutting the other way is perhaps the most widely accepted normative account of constitutional judicial review, John Hart Ely's representation-reinforcement theory.[79]

Though Ely never wrote about education rights, his twin themes of enabling equal participation in the democratic process and, where equal participation has not been achieved, correcting failures of the process *ex post*, suggest that courts ought to play some role enforcing the duty of care—*particularly as to the production of knowledge about the effects of the educational system on economic and political mobility*. There are four related reasons for this. First, children lack the right to vote and, thus, cannot look after their own interests. Second, even if one assumes that parents adequately represent their children,[80] parents have no stake in what researchers can learn *from* their children for the benefit of future children. (By construction, the benefits of research on intergenerational mobility accrue well into the future.) Third, short-term bias is likely to affect political representatives; investing in studies that might benefit children in the next generation can be a hard sell.[81] Fourth, poor adults vote less frequently and are less politically efficacious than affluent adults, and, thus, poor children, especially future poor children, are particularly disadvantaged in the political arena.[82] It follows that judicial enforcement of the education right could rectify a discrete failure resulting from the underrepresentation of poor children—namely, under or mis-investment in education and education research—and, over time, judicial enforcement could give poor families a more equal voice in the political process. Recall that courts have almost universally explained the education right in terms of democratic participation as well as economic opportunity.[83] In sum, it is normatively appealing, and exegetically sound, to interpret

[78] *Id.* at 759. A related, and also supportive, argument has been made by Tang, who argued that the state has failed in its duty to provide a "system" of public education if the state has not articulated educational standards and then taken reasonable steps to "align its school funding structure" to achieving those standards. Aaron Y. Tang, *Broken Systems, Broken Duties: A New Theory for School Finance Litigation*, 95 MARQ. L. REV. 1195, 1201 (2011).

[79] *See generally* JOHN HART ELY, DEMOCRACY AND DISTRUST: A THEORY OF JUDICIAL REVIEW (1980).

[80] This is a problematic assumption unless parents are given an extra vote for each child.

[81] For a recent survey of the literature on the time horizon of democratic politicians, see generally Alan M. Jacobs, *Policy Making for the Long Term in Advanced Democracies*, 19 ANN. REV. POL. SCI. 433 (2016).

[82] For recent evidence and reviews of the literature, see generally Bertrall L. Ross II & Su Li, *Measuring Political Power: Suspect Class Determinations and the Poor*, 104 CALIF. L. REV. 323 (2016); Nicholas O. Stephanopoulos, *Political Powerlessness*, 90 N.Y.U. L. REV. 1527 (2015).

[83] *See supra* note 17 and accompanying text.

the positive command to provide a free public education as representative of a structural decision made within state constitutions to ensure decent opportunities for disadvantaged children.

Judicial enforcement of the education right promises to achieve this goal in a relatively modest way, by giving effect to a narrow textual guarantee, rather than by deeming children or poor people to be a suspect class, a move that could subject much of social welfare, tax, and family law to strict scrutiny.

The "framework" approach to judicial implementation of the education right, coupled with a deferential-but-not-toothless standard of review, is a reasonable way for courts to navigate the crosscutting representation-reinforcement and institutional competence currents. Notably, this approach loosely tracks what courts have done in administrative law, which deals with an analogous set of problems. In the usual administrative law setting, an agency, with more expertise and better fact-finding capabilities than the courts, undertakes to implement a broadly worded statute.[84] Under constitutional provisions that establish a state duty to prepare students for future employment and democratic participation, "the state" in effect becomes "the agency," charged with carrying out a legal mandate defined at a similarly high level of generality. There are, of course, some very important differences between this situation and the prototypical statutory delegations of administrative law, including the lack of a single, unitary implementing agency and the lack of explicit textual hooks for procedural mandates. Courts should not borrow the conventions of administrative law for education-rights cases without attending to these differences.[85] Still, administrative law remains instructive because it furnishes a set of strategies (such as requiring record-building and reasoned explanation) for dealing with situations where a nonjudicial actor with relevant expertise and greater political legitimacy has primary responsibility for effectuating the law's objectives.[86]

[84] As Gersen and Vermeule observed, "[T]he motivating assumption for most of administrative law is information asymmetry." Jacob Gersen & Adrian Vermeule, *Thin Rationality Review*, 114 MICH. L. REV. 1355, 1399 (2016).

[85] We think these differences in context have at least two important implications. First, and most obviously, courts, before imposing procedural requirements, would have to make a prior determination that the state's constitutional text or traditions imply that the courts have authority to create procedural duties or prophylactic rules as a means of indirectly enforcing constitutional provisions that are difficult for courts to enforce directly. Second, because the state constitutions create a positive entitlement to education and do not vest any single branch of government with *exclusive* authority to implement the right, courts, upon finding a violation, should not block the state from acting. Whereas the usual remedy under the Administrative Procedure Act ("APA") is to set aside the agency decision at issue (preventing the agency from acting pending a do-over), violations of the education clauses should be met with "remands without vacation," purely declaratory relief, or an order providing substantive relief to affected school children that would remain in effect until the legislature undertakes to provide a different remedy. On the remands without vacatur in administrative law, see generally Ronald M. Levin, *"Vacation" at Sea: Judicial Remedies and Equitable Discretion in Administrative Law*, 53 DUKE L.J. 291 (2003).

[86] We recognize that many administrative law scholars have been quite critical of the practice of arbitrariness review under the APA, and especially of the courts' occasional efforts to read in procedural requirements to the APA. We think there is little risk of ossification from judicial adoption of the "framework" approach to education rights, coupled with an arbitrariness standard of review. First, as we observe, in note 85, *supra*, the positive nature of the education right means that courts need not—and generally should not—remedy violations by *preventing* legislative action from taking effect. Second, absent an exclusive delegation to the legislature by the education clauses, courts have authority to provide interim, affirmative relief to affected students. Note also that our suggested knowledge-production planning requirement would be a requirement of periodic state action, akin to the statutory deadlines that are often thought to speed up agency rule-making.

Notice also that by cashing out the education clauses in terms of a duty of care, enforced with planning requirements and arbitrariness review, courts can sidestep the difficult causation questions on which education-quality claims have often foundered.[87] If the state fails to exercise appropriate care, via appropriate procedures, in managing the educational system, those failures are unconstitutional *as such*, as breaches of the duty of care, regardless of whether it can be shown that the plaintiffs in a particular case would have obtained a "quality" education if only the state had, say, commissioned a costing-out study prior to updating the school-funding formula. Administrative law is again instructive: Plaintiffs challenging an agency's failure to consider a reasonable range of alternatives, to disclose important studies on which the agency relies, or to respond to significant public comments are not required to prove that the agency would have made a different and more beneficial decision had it followed proper procedures.[88] Normal causation and redressability requirements are relaxed.

But here's the puzzle: In fleshing out the legislative duty of care, courts have required the legislature to act reasonably in light of *then-available* information, *e.g.*, by adopting a record-justified funding formula.[89] Courts have also required the legislature to generate new information going forward about students' academic proficiency (testing). [90] But courts have said nothing—indeed, so far as we can tell, *no court has ever been asked to say anything*—about state duties to help figure out which actual, or potential, educational reforms would improve more constitutionally weighty, long-term outcomes. If the education interest is fundamental because of the presumed importance of education for future participation in economic, civic, and political life, then performance on standardized tests is, at best, a proxy of unknown strength for the constitutionally important outcomes: future employment, income, democratic participation, freedom (non-incarceration), and perhaps health and longevity. Under conditions of high socioeconomic inequality, low mobility, and great uncertainty about the effects of actual and potential educational interventions, the duty of care should probably leave state actors with a lot of discretion about what educational reforms to pursue and how much to spend on them—but not about whether to implement those reforms in a manner conducive to estimating their effects. And, as noted above, the representation-reinforcement case for judicial review is particularly strong with respect to research on intergenerational political and economic mobility.

Yet, the courts have said nothing about this. The tacit assumption has been that the current state of knowledge about "what works" is something for which the state bears no responsibility. Such knowledge either develops or it does not, and there is little that the legislature or the courts can do about it. For instance,

[87] On causation hurdles, see *supra* notes 56–58 and accompanying text.

[88] *See* Lujan v. Defenders of Wildlife, 504 U.S. 555, 572 n.7 (1992) ("The person who has been accorded a procedural right to protect his concrete interests can assert that right without meeting all the normal standards for redressability and immediacy."); Evan Tsen Lee & Josephine Mason Ellis, *The Standing Doctrine's Dirty Little Secret*, 107 NW. U. L. REV. 169, 187–203 (2012).

[89] *See supra* note 19 and accompanying text.

[90] *See* cases cited *supra* notes 56–57.

when the Texas Supreme Court recently concluded that "the scientific community . . . has reached an impasse" on the "intractable" question of whether additional school funding materially improves student outcomes, the court simply rubber-stamped the state's decisions rather than pausing to consider whether the "scientific impasse" might itself be a consequence of Texas's decisions about how to distribute school funding and track student outcomes.[91] Knowledge about what works is assumed to be exogenous to law. Part III explains the error of this assumption.

III. HOW STATES CONTROL THE PRODUCTION OF (LEGALLY URGENT) KNOWLEDGE ABOUT EDUCATION

Because the *constitutional* quality of a state's education system depends on whether it prepares disadvantaged students for a lifetime of gainful employment and effective democratic participation, the state actors who oversee it need to be supplied with a particular kind of research. They need research quantifying the *causal effect* of alternative educational programs or policies on students' *adult outcomes*, particularly for students who are *demographically at risk* of bad outcomes. This is so because to say whether the state's education programs violate the duty of care, one must have some sense of what the state could be doing better, and at what cost.[92]

Consider the issue of reducing class sizes. Small classes are very expensive, but they do seem to have educational benefits, particularly for disadvantaged children.[93] Whether small classes are constitutionally required depends on *how much* they are likely to benefit disadvantaged children and whether similar benefits could be achieved using less expensive interventions.

Our point is not that quantitative research is the only important or valuable kind of education research.[94] Qualitative research is critical for generating hypotheses,[95] for assessing the validity of educational tests,[96] and even for determining whether the results of quantitative studies mean what they purport to mean. For example, the appropriate interpretation of a quantitative study showing that the adoption of a particular curriculum did not result in improved test

[91] *See* Morath v. Tex. Taxpayer & Student Fairness Coalition, 490 S.W.3d 826, 852–55 n.152 (Tex. 2016).

[92] This kind of research is particularly important for judges and legislators, who are unlikely to have experience-based craft or tacit knowledge of how to educate disadvantaged students effectively (unlike, say, a master teacher).

[93] *See* 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, 13–14 (2010); SCHANZENBACH, *supra* note 34. *But see* HANUSHEK & LINDSETH, *supra* note 4, at 205 (arguing that benefits of class-size reduction in California were offset by "commensurately sudden and sharp increase in the demand for teachers," which resulted in "districts serving disadvantaged populations disproportionately los[ing] their most qualified teachers").

[94] The development of better quantitative methods and their increasing influence has perhaps not surprisingly led to a pushback from some education researchers. *See, e.g.*, Superfine, *supra* note 70, at 688–89 (citing and discussing criticisms).

[95] *See id.* at 696.

[96] NAT'L RESEARCH COUNCIL, SCIENTIFIC RESEARCH IN EDUCATION 19 (Lisa Towne & Richard J. Shavelson eds., 2002).

performance depends on whether the new curriculum was properly implemented.[97] Our point is simply that, in a world of finite resources, it is hard to say whether the state is constitutionally obligated to provide certain educational programs or funding levels unless the likely effects of the programs or funding can be estimated.[98]

For the quantitative research enterprise to usefully guide implementation of the education right, three conditions must be satisfied. Think of these conditions as links in the useful-causal-research chain. Unless all three links prove strong, the research enterprise may well mislead more than it illuminates.

The first link is this: Researchers must be able to observe constitutionally important outcomes and to associate those outcomes with the educational "treatments" each student received. Without knowing which students were exposed to different educational programs or policies, and without observing these students' future outcomes, there is no way to estimate the programs' effects. This point may seem too obvious to merit remark, but we will see that it remains a serious practical problem.

The second link in the chain concerns the researcher's response to what statistician Paul Holland famously dubbed "the fundamental problem of causal inference."[99] To explain this, we need to introduce some terminology. Modern statisticians and social scientists employ a conceptual framework for causal inference called the Neyman-Rubin Model (or a generalization of the model).[100] The Neyman-Rubin Model begins with three primitive concepts: *units*, *treatments*, and *potential outcomes*.[101] *Units* are the units of observation.[102] In many education studies, the units are individual students, but in principle, classrooms, schools, or school districts could comprise the units. *Treatments* are actual or potential interventions in the world whose effects on the units the analyst wants to evaluate.[103] In an education study, the treatments might consist of varying

[97] *Cf.* NAT'L RESEARCH COUNCIL, ON EVALUATING CURRICULAR EFFECTIVENESS 43–46 (Jere Confrey & Vicki Stohl eds., 2004) (canvassing implementation issues).

[98] Accordingly, in the remainder of this Article, we will focus on the state role in the production of information useful for quantitative research. A roughly parallel argument can be made about the state-role in the production of useful qualitative research (which is often needed to properly interpret quantitative findings). For instance, states need to provide access to well-qualified researchers to observe whether a given treatment, say a curriculum, is actually being appropriately administered.

[99] Holland, *supra* note 10, at 947.

[100] On Neyman and Rubin, see Donald B. Rubin, *Comment: Neyman (1923) and Causal Inference in Experiments and Observational Studies*, 5 STAT. SCI. 472, 472 (1990). The penetration of the Neyman-Rubin model into economics led to the "credibility revolution" in that field. *See* Angrist & Pischke, *supra* note 93, at 5. The same revolution is taking place among empirically oriented legal scholars. *See, e.g.*, Michael Abramowicz et al., *Randomizing Law*, 159 U. PA. L. REV. 929, 931–32 (2011); D. James Greiner, *Causal Inference in Civil Rights Litigation*, 122 HARV. L. REV. 533, 536 (2008); Daniel E. Ho & Donald B. Rubin, *Credible Causal Inference for Empirical Legal Studies*, 7 ANN. REV. L. SOC. SCI. 17, 20 (2011). Perhaps the most significant generalization of the Neyman-Rubin model is JUDEA PEARL, CAUSALITY (2d ed. 2009). *See also*, STEPHEN L. MORGAN & CHRISTOPHER WINSHIP, COUNTERFACTUALS AND CAUSAL INFERENCE 41–45 (Cambridge Univ. Press 2007), an accessible textbook that relates the Neyman-Rubin Model to Pearl's graphical methods.

[101] For textbook treatments of the Neyman-Rubin model, we recommend MORGAN & WINSHIP, *supra* note 100, at 4, and GUIDO W. IMBENS & DONALD B. RUBIN, CAUSAL INFERENCE FOR STATISTICS, SOCIAL, AND BIOMEDICAL SCIENCES 3 (2015). Another strong textbook on causal inference, emphasizing education applications, is RICHARD J. MURNANE & JOHN B. WILLETT, METHODS MATTER: IMPROVING CAUSAL INFERENCE IN EDUCATIONAL AND SOCIAL SCIENCE RESEARCH 350–68 (2011).

[102] Holland, *supra* note 10, at 945.

[103] *Id.* at 946.

curricula, teachers, classroom sizes, incentive pay structures, funding levels, or even accountability systems. *Potential outcomes* are the outcomes, at some defined time in the future, that each unit *would* realize under different treatments.[104] In an education study, potential outcomes might take the form of test scores, earnings, or rates of truancy, graduation, employment, voting, or incarceration.[105]

The causal effect of Treatment A relative to Treatment B on a given unit is the difference between the unit's potential outcomes under each treatment. The "fundamental problem of causal inference" is that researchers only observe *one* potential outcome for each unit—the outcome the unit realized under the treatment it received.[106] Statisticians, economists, and methodologists in other disciplines have devoted an enormous amount of effort over the last generation to strategies for overcoming the fundamental problem of causal inference. They have made terrific progress, establishing a rough hierarchy of research methods defined by the strength of the assumptions needed to support a causal interpretation of a study's findings; the "check-ability" of those assumptions (whether the assumptions have observable implications that can be cross-checked against the available data); and the populations with respect to which causal inferences can be made. We will say more about these methods momentarily.[107] For now, it is enough to say that the second link in the useful-causal-research chain is stronger insofar as the effect estimates have been generated using well-executed, top-of-the-hierarchy research designs; the link is weaker if the estimates rely on strong, uncheckable assumptions or pertain only to local, unrepresentative populations.

The third link in the useful-causal-research chain concerns the uptake of research findings by state actors and the experts who seek to inform them. Consumers of education research—such as analysts preparing meta-analyses or expert witnesses summarizing the literature for courts—must be able to make reasonable, transparent judgments about whether reported effect sizes and tests of statistical significance in a defined body of work are believable. A traditional, but gravely inadequate, answer to this question is to say that an estimate is believable if it was generated using standard methods by suitably credentialed researchers and was published in a peer-reviewed journal.[108] This rule of thumb is unsound in a world where publication biases, professional incentives, or even ideological homogeneity within the research community can result in a body of research findings that do not mean what they appear to mean.

[104] *Id.*

[105] If classrooms, schools, or districts are the units, the study's outcome measure would be something observed at that level of aggregation, such as the average or median test score in each classroom, school, or district.

[106] Holland, *supra* note 10, at 959.

[107] For a somewhat more technical review, emphasizing work by economists, see Guido Imbens & Jeffrey M. Wooldridge, *Recent Developments in the Econometrics of Program Evaluation*, 47 J. ECON. LITERATURE 5, 20 (2009).

[108] *Cf.* Daubert v. Merrell Dow Pharm., Inc., 509 U.S. 579, 592–95 (1993) (describing hypothesis testing, peer-review, error rates, and "general acceptance" as key factors for determining whether putatively scientific evidence should be admitted at trial).

This Part explains the links in the useful-causal-research chain. The main takeaway is that the strength of each link critically depends on the state.

A. *Observing Outcomes, Linked to Treatments*

Not that long ago, most citizens' lifetime outcomes were essentially unknown to the government. Today, however, governments collect an enormous amount of information about each member of society. Tax agencies know your income, and whether you receive credits for health insurance, earned income, children, etc. Criminal justice administrators know whether you are in jail or on probation and whether there's a warrant for your arrest. Human services agencies know whether you get income support, disability payments, or food stamps. The secretary of state knows whether you are registered to vote and which elections you have voted in. The department of public health knows your birth date and birth weight, and your history of marriage and divorce. Sometime in the future, it will also know the date of your death.[109]

These administrative datasets are potential gold mines for research that could guide implementation of the education right, as it is incredibly difficult to study educational effects on adult outcomes—the outcomes that matter constitutionally—without administrative data. Though a few researchers (and the federal government) have created "purpose built" datasets that track a sample of students over many years, or even decades, such datasets are very expensive to create, and they suffer from attrition over time.[110] (Subjects who initially agreed to participate in a study may withdraw their consent later, or simply disappear from view.) Because they are expensive to create and maintain, purpose-built datasets tend to be small,[111] whereas administrative datasets are huge, often encompassing the entire population of interest.[112] The size of administrative datasets makes it possible for researchers to create much more precise estimates and estimates that can be generalized to the state's population (or to localities

[109] *E.g.*, *Birth, Death, Fetal Death, Still Birth & Marriage Certificates*, CAL. DEP'T PUB. HEALTH, https://www.cdph.ca.gov/Programs/CHSI/Pages/Birth,-Death,-Fetal-Death,-Still-Birth--Marriage-Certificates.aspx (last visited Jan. 16, 2018) ("California birth, death, fetal death, still birth, marriage and divorce records are maintained by the California Department of Public Health Vital Records.").

[110] The federal government has also created some important, purpose-built educational databases that track a sample of the student population over time. *See*, *e.g.*, *Early Childhood Longitudinal Program (ECLS)*, NAT'L CTR. EDUC. STAT., http://nces.ed.gov/ecls/ (last visited Jan. 16, 2018).

[111] For example, the highly influential Abecedarian and Perry studies of the long-term effects of early childhood educational interventions each had fewer than sixty subjects in the treatment condition. For summaries of these programs, see Barnett, *supra* note 33, at 975–77; Duncan & Magnuson, *supra* note 33, at 116–18.

[112] For example, voter registration files are supposed to include all registered voters in the state; tax records, all taxpayers in the state; and social-service agency records, all persons who receive welfare payments from the state. To be sure, these datasets are not error free. *See* Frank Verschaeren, *Checking the Usefulness and Initial Quality of Administrative Data* (Am. Statistical Ass'n, Working Paper No. 302180, 2012), https://ww2.amstat.org/meetings/ices/2012/papers/302180.pdf.

within the state) without making heroic assumptions.[113] For example, the research by Chetty and co-authors on the geography of mobility discussed in Part II would have been impossible to undertake without tax records from the IRS.[114]

Purpose-built datasets are usually constructed from interviews. Research subjects, however, have faulty memories, and people sometimes lie to interviewers.[115] Administrative datasets record events and transactions rather than recollected memories. These datasets certainly are not error-free, but at least the usual errors of recording and transcription are not compounded by errors of memory or a reluctance to answer sensitive questions honestly.

Finally, administrative datasets are usually essential for retrospective studies. A researcher, who decades after the fact learns that an educational reform was rolled out in a particular community, may be able to estimate the reform's effects using administrative data, but if no administrative data are available, she is typically out of luck.[116] She cannot go back in time and collect the data herself.

That administrative datasets have enormous potential for education research is clear, but whether this potential is realized depends on the cooperation of the state. Researchers need to be able to link records of individuals' educational experiences ("treatments") with records of the same individuals' subsequent outcomes in other social, economic, and political domains. For these linkages to be made, state education administrators must maintain detailed, accurate records of students' school and classroom assignments, as well as the assignment of teachers and curricula to classrooms. And, critically, the school records must contain identifiers that allow students to be matched to their future and past selves in other administrative datasets.[117] Finally, there must be a procedure in place for researchers to obtain matched records from the state, with individual identifying information removed to safeguard privacy interests.[118]

There are huge differences across states—and nations—in the degree to which education records can be matched to other administrative datasets. For

[113] When researchers observe outcomes for the entire state population (rather than a sample), researchers can make inferences about the effects of treatments on the state population without making assumptions about the representativeness of a sample. Samples are only representative (in expectation) if each member of the population has an equal probability of being included in the sample, and because different members of the population are sure to differ in their ease of being reached and willingness to participate in a study, researchers who want to make inferences to the full population must reweight the sample to approximate the population. This reweighting depends on a fair amount of guesswork because the attributes that the researcher observes (and can therefore reweight on the basis of) may not include key attributes that are both correlated with the probability that an individual is included in the study and correlated with the individual-level treatment effect (difference in potential outcomes). For further discussion, see Erin Hartman et al., *From Sample Average Treatment Effect to Population Average Treatment Effect on the Treated: Combining Experimental with Observational Studies to Estimate Population Treatment Effects*, 178 J. ROYAL STAT. SOC'Y: SERIES A 757, 765 (2015).

[114] Chetty et al., *Where is the Land of Opportunity?*, *supra* note 21, at 1554.

[115] David N. Figlio et al., *Education Research and Administrative Data* 2 (Nat'l Bureau of Econ. Research, Working Paper No. 21592 2015) [hereinafter Figlio et al., *Education Research*].

[116] To be sure, a number of important retrospective studies have been conducted using purpose-built datasets created by the federal government. *See*, *e.g.*, Jackson et al., *supra* note 46, at 161 (using state-level panels from the National Assessment of Educational Progress, created in 1990 by the U.S. Department of Education, to estimate effect of judicially induced spending increases on achievement gaps). But this just reinforces the importance of the government role.

[117] For more on record matching, see *infra* Section IV.B.

[118] Records have been "matched" when those persons who appear in both datasets are identified as being the same individual. Figlio et al., *Education Research*, *supra* note 115, at 23.

example, when the Nordic countries established social security systems in the 1960s, they assigned each citizen a unique identifier and quickly began using this identifier in numerous administrative databases, including birth and death records.[119] Researchers who want to use these data may petition the national statistical office, and if their project is approved, the office matches the requested records by social security identifier and creates a replacement identifier in the dataset provided to the researcher.[120] The social security identifier is not released.[121]

This arrangement has enabled education research that would be difficult or impossible to carry out elsewhere, such as studies of the causal effect of mandatory schooling on educational attainment, employment, and criminal activity *in the next generation* (parents and their offspring can be linked through the administrative records),[122] and studies of the effect of publicly provided day care on labor market outcomes decades later.[123] The linkage to birth records has also allowed researchers to relate birth weight and newborn health indicators to future educational outcomes,[124] an important development that could help schools identify "at risk" children early on.

Within the United States, education agencies in Florida, Texas, and North Carolina in the late 1990s developed systems that allow students to be tracked over time and matched to classrooms and teachers.[125] Florida's Department of Education has also matched student records to post-secondary school data on military service, criminal justice, and labor market outcomes,[126] and Florida's Department of Health cooperated with its Department of Education to match birth records from 1992–2002 to school records.[127] Critical to this exercise was the fact that Florida—like the Nordic countries, but uniquely among the fifty states—assigns a social security number to each newborn[128] and uses this number to track students through the education system.[129] These pioneering efforts have already generated an important paper on the effects of quality teachers on students' future earnings, but the study in question only had teacher data from a single urban school district.[130] Researchers will not be able to learn whether the results generalize until they can link student records to adult-outcome records in other states.

[119] *Id.* at 7.

[120] *Id.* at 8.

[121] *Id.*

[122] *Id.* at 8, 14.

[123] *Id.* at 9 (citing studies).

[124] *Id.* at 30–31.

[125] *Id.* at 17–18.

[126] *Id.* at 18.

[127] *Id.* at 19.

[128] *Id.* at 28.

[129] For an illustration of why this is so important for record matching, see David Figlio et al., *The Effects of Poor Neonatal Health on Children's Cognitive Development*, 104 AM. ECON. REV. 3921, 3924 n.6 (2014) (comparing match rate achieved in Florida, where social security numbers are included in the administrative datasets, with match rates achieved in North Carolina, where this identifier is lacking).

[130] Raj Chetty et al., *Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates*, 104 AM. ECON. REV. 2593, 2593–94 (2014).

Despite a slow and bumpy start,[131] state record-keeping and record-linkage policies are evolving rapidly, and mostly in a salutary direction. In 2005, a new national nonprofit group, the Data Quality Campaign ("DQC"), began advocating for better educational data systems—including matching educational and workforce records—with the goal of helping states use longitudinal data to improve student achievement.[132] Congress, through the America Competes Act of 2007, largely embraced the DQC's agenda,[133] and Congress required states to improve their educational data systems as a condition for receiving fiscal stabilization funds under the American Recovery and Reinvestment Act of 2009.[134] DQC has conducted annual surveys of states to measure progress.[135] According to the most recent survey, nineteen states have securely linked K-12 educational records to early learning, postsecondary, and workforce record systems;[136] forty-three states have established a cross-agency data governance committee with authority to regulate agency practices; and twenty-eight states have reasonably transparent policies about who is authorized to use the data and for what purposes.[137] Pew Charitable Trusts has also launched a major project on state data systems and record linkage, with the goal of pinpointing differences among the states and identifying best practices.[138]

The perspective of researchers on the ground, however, is somewhat less sanguine. Economist David Figlio reports that many states prohibit the use of social security numbers in connection with educational records,[139] and more than half forbid the linkage of educational records with other records.[140] DQC leaders acknowledge that their periodic surveys may not fully capture the reality on the ground.[141] Education agencies that are supposed to match and share records often stymie researchers, raising sometimes meritless objections under state law or the federal Family Educational Rights and Privacy Act,[142] perhaps to avoid

[131] For the history of state attempts to implement "No Child Left Behind," including its record keeping requirements, see Gail L. Sunderman & Gary Orfield, *Domesticating a Revolution: No Child Left Behind Reforms and State Administrative Response*, 76 HARV. EDUC. REV. 526, 552 (2006).

[132] DATA QUALITY CAMPAIGN, FACT SHEET, http://2pido73em67o3eytaq1cp8au.wpengine.netdna-cdn.com/wp-content/uploads/2016/04/DQC-Fact-Sheet-4-28-16.pdf (last visited Jan. 16, 2018) [hereinafter DATA QUALITY CAMPAIGN, FACT SHEET].

[133] *Id.*

[134] *State Progress*, DATA QUALITY CAMPAIGN, https://dataqualitycampaign.org/why-education-data/state-progress/ (last visited Jan. 16, 2018).

[135] DATA QUALITY CAMPAIGN, DATA FOR ACTION 2014: PAVING THE PATH TO SUCCESS 21 (Nov. 2014), http://dataqualitycampaign.org/resource/data-action-2014-paving-path-success/ [hereinafter DATA QUALITY CAMPAIGN, DATA FOR ACTION].

[136] The quality of these matches varies from state to state.

[137] DATA QUALITY CAMPAIGN, FACT SHEET, *supra* note 132, at 19.

[138] Tom Conroy, *Big Data Is Big News*, PEW CHARITABLE TRUSTS (Dec. 4, 2015), http://www.pewtrusts.org/en/research-and-analysis/analysis/2015/12/04/big-data-is-big-news.

[139] Figlio et al., *Education Research*, *supra* note 115, at 23.

[140] *Id.* at 29 ("[T]he majority of the U.S. states [] have regular student assessments but do not allow merging different databases."). We asked David Figlio about the basis for this assertion; he replied, "reports from researchers from all over the country who have been rebuffed by state agencies." Email from David Figlio to Chris Elmendorf (Apr. 28, 2016).

[141] Telephone Interview with Paige Kowalski, Vice President for Policy & Advocacy, Data Quality Campaign, & Elizabeth Dabney, Assoc. Dir., Research & Policy Analysis, Data Quality Campaign (May 17, 2016).

[142] 20 U.S.C. § 1232(g) (2012) (with implementing regulations in title 34, part 99 of the Code of Federal Regulations).

scrutiny. Also, some states flatly prohibit the use of critical outcomes datasets, such as records of voter registration and turnout, for research purposes.[143]

Absent state-enabled record linkage, the outcomes that researchers tend to study are the outcomes found in the state's education databases or federally funded surveys, such as scores on standardized tests. Courts, similarly, have used test-score outcomes as a constitutional guidepost.[144] But it's far from clear that standardized test scores represent a good proxy for the *constitutional* quality of the educational system.

In a series of papers, the Nobel Laureate James Heckman argued that the education of young children ought to focus on the development of personality traits, such as conscientiousness, rather than cognitive or analytic ability.[145] Similarly, psychologist and McArthur Fellow Angela Duckworth argued that "grit" and self-control are keystone competencies that schools ought to instill.[146] A number of studies have shown that such personality traits are more predictive of adult outcomes than IQ scores or performance on standardized tests, and a huge body of work has emerged over the last two decades on instructional programs for socioemotional development.[147] If Heckman, Duckworth, and others in their camp are correct, reforms that improve performance on standardized tests could actually worsen lifetime outcomes. Worse lifetime outcomes are likely to result if teachers respond to testing incentives by substituting test-prep drills for more productive, grit-enhancing forms of instruction.

To sum up, it is not news that the government controls the data that it collects. What we are emphasizing—and what is not as well understood—is that the massive datasets assembled by state governments for administrative rather than research purposes have enormous potential to guide implementation of the education right. Yet this potential often goes unrealized because of prohibitions on record linkage or seemingly minor flaws in the state's record-keeping or data-sharing architecture.

[143] For a breakdown of the allowable uses of state voter files, see *Full List Facts and Info*, VOTER LIST INFORMATION, http://voterlist.electproject.org/full-list-purchase-facts-and-info (last visited Jan. 16, 2018). Massachusetts is an example of a state that flatly prohibits research use. *Massachussets*, VOTER LIST INFORMATION, http://voterlist.electproject.org/states/massachussets (last visited Jan. 16, 2018); Alaska is an example of a state that permits research use while withholding information that would greatly aid linkage to education records, such as social security numbers and birth dates. *Alaska*, VOTER LIST INFORMATION, http://voterlist.electproject.org/states/alaska (last visited Jan. 16, 2018).

[144] *See supra* note 18 and accompanying text.

[145] *See, e.g.*, Flavio Cunha et al., *Estimating the Technology of Cognitive and Noncognitive Skill Formation*, 78 ECONOMETRICA 883, 920 (2010); Mathilde Almlund et al., *Personality Psychology and Economics* (Inst. for the Stud. Of Lab. ("IZA"), Discussion Paper No. 5500, 2011); James J. Heckman et al., *A New Cost–Benefit and Rate of Return Analysis for the Perry Preschool Program: A Summary* (Nat'l Bureau of Econ. Research, Working Paper No. 16180, 2010).

[146] For a popular treatment, see ANGELA DUCKWORTH, GRIT: THE POWER OF PASSION AND PERSEVERANCE (2016).

[147] For a recent literature review, see Joseph A. Durlak et al., *The Impact of Enhancing Students' Social and Emotional Learning: A Meta-Analysis of School-Based Universal Interventions*, 82 CHILD DEV. 405, 401 (2011); *see also* Michael Baker et al., *Non-Cognitive Deficits and Young Adult Outcomes: The Long-Run Impacts of a Universal Child Care Program* 3 (Nat'l Bureau of Econ. Research, Working Paper No. 21571, 2015) (finding that universal pre-K program in Quebec had contemporaneous negative effects on socio-emotional outcomes and long-run negative effects of criminal behavior and health outcomes).

B. The Problem of Causal Inference

Let us assume that the researcher has access to relevant administrative datasets and can link records across the datasets. Her next task is to tackle Holland's "fundamental problem of causal inference."[148] Again, the researcher wants to learn the difference between units' potential outcomes under various treatments, but for each unit, she observes only one outcome—the outcome the unit realized under the treatment it received. So, to estimate the effect of treatment A relative to B *on a given unit*, the researcher would need to estimate the unit's outcome under the treatment it did not receive. Social scientists can do this but only with a lot of guesswork, such as by assuming that potential outcomes are identical among units that share certain pre-treatment traits, or that potential outcomes bear a particular functional relationship to observed traits and are not related to any unobserved traits.

One of the great contributions of Rubin (of the Neyman-Rubin Model) was to show that under certain well-specified conditions, the *average* causal effect of a treatment on a group of units can be estimated *without making any assumptions about how potential outcomes vary or do not vary with background characteristics of the units.*[149] The critical conditions are, first, that the researcher knows each unit's *ex ante* probability of receiving the treatment; second, that these probabilities, sometimes called "propensity scores," are bounded between zero and one; and, third, that the potential outcomes of each unit are independent of the treatments received by the other units.[150] When these conditions hold, the difference between weighted averages of the *observed* outcomes in each treatment condition provides an unbiased estimate of the average treatment effect.[151]

The intuition is fairly straightforward. The difference between average observed outcomes in two groups, only one of which received a treatment, could, in theory, result from: (1) the treatment, (2) the sorting of units into treatment groups in a manner correlated with potential outcomes, or (3) some combination of the treatment and sorting. Recall Chetty and Hendren's study of the effect of localized geographic societies on intergenerational economic mobility. In their study, the treatment consisted of a child's years of exposure to a given locality, that is, the number of years the child spent growing up in that community.[152] A simple comparison of the average outcome (economic transitions) of children who spent *x* years in locality A and with the average outcome of children who spent *x* years in locality B is likely to yield a biased estimate of the average

[148] *See* Holland, *supra* note 10; *supra* text accompanying note 99.

[149] *See* Paul R. Rosenbaum & Donald. B. Rubin, *The Central Role of the Propensity Score in Observational Studies for Causal Effects*, 70 BIOMETRIKA 41, 43–44 (1983).

[150] For the canonical proof that balancing treatment and control units on "propensity scores" (probability of assignment to treatment) provides balance in expectation on all covariates that may be correlated with potential outcomes, see *id.* at 51. For an accessible textbook treatment of the Neyman-Rubin framework, see MORGAN & WINSHIP, *supra* note 100, at 4.

[151] In their canonical paper, Rosenbaum and Rubin proposed stratifying on propensity scores. *See* Rosenbaum & Rubin, *supra* note 149, at 48–53. The reweighting result owes to Guido Imbens. *See* Guido W. Imbens, *The Role of the Propensity Score in Estimating Dose-Response Functions*, 87 BIOMETRIKA 706, 706 (2000). For an illustrative simulation of the reweighting result, see MORGAN & WINSHIP, *supra* note 100, at 100–04.

[152] Chetty & Hendren, *supra* note 23, at 15.

treatment effect of A relative to B. If A is a good neighborhood for mobility, those low-income families who invest a lot of effort into moving their children up the economic ladder are probably more likely to choose to live in A rather than B, and the children of these families probably have better outcomes than other low-income children independent of the treatment.

By contrast, a comparison of the average outcomes of the children who spent *x* years in A with the average outcomes of the children who spent *x* years in B would yield an unbiased estimate of the treatment effect if each child had been randomly assigned to spend *x* years in A *or* B (and had "complied" with this assignment).[153] Where every unit has an equal probability of being assigned to each treatment condition, as in the classic randomized controlled trial, the distribution of background characteristics is, in expectation, the same in both groups. Where assignment probabilities are unequal, achieving expected balance in background characteristics between the groups is just a matter of reweighting each unit to account for its assignment probability.[154]

Nonrandom sorting into treatment and control groups is a ubiquitous problem for education research. For example, parents who invest a lot in their children (and whose children therefore have the best potential outcomes) are surely more likely than other parents to finagle a spot for their child in the "best" schools, which means that the causal effects of a school cannot be reliably estimated by comparing average outcomes across schools. Principals may assign easy (or difficult) students to teachers as a way of rewarding or punishing teachers, or because certain teachers are thought to have a particular gift for educating difficult students. This makes it perilous to gauge teacher effects by comparing average student gains across classrooms. Education agencies or school districts may direct supplemental funds to schools with particularly challenging student populations, biasing downward the estimated effect of spending on student outcomes. Or, parents of raring-to-learn students may lobby successfully for additional funding for their schools, resulting in a positive correlation between spending and potential outcomes, and, thus, an upwardly biased estimate of the effect of spending on student outcomes.

Given the pervasiveness of sorting, when can it be said that a study is well designed for causal inference? Formally, any empirical study that advances a causal claim rests on what researchers call *identifying assumptions*. Identifying assumptions are posited facts about the world that, if true, mean that the study's estimate of the treatment effect is unbiased.[155] The *strength* of a study's causal claim is inversely proportional to the *weakness* and *plausibility* of the identifying assumptions.[156]

[153] Provided that each unit's potential outcome is independent of the other units assigned to the treatment condition. This assumption may well be violated in the Chetty & Hendren study (because of peer effects), unless the number of units assigned to each commuting zone is small relative to the population of the zone. *Id.*

[154] Technically, the reweighting is by the *inverse* of assignment probabilities. *See* Imbens, *supra* note 151, at 707–08.

[155] *See generally* PAUL R. ROSENBAUM, DESIGN OF OBSERVATIONAL STUDIES (2010).

[156] Researchers can quantify the sensitivity of an ostensibly causal finding to violations of the identifying assumptions. A study's causal claim is strong if it would take a big violation of the identifying assumptions to vitiate the effect. *See id. at* 257–74. But it still remains to ask how *plausible* such a violation is. In some designs, there are strong *a priori* reasons to believe the identifying assumptions to be satisfied. The canonical example is

Roughly speaking, one can arrange research designs for causal inference on a spectrum from strong to doubtful. Anchoring the "strong" end of the spectrum is the randomized controlled trial ("RCT") with full compliance.[157] In a randomized trial with perfect compliance, the treatment assignment probabilities are *known*. Thus, the average causal effect of the treatment on the population of interest can be estimated (unbiasedly) without making any assumptions about the relationship between units' background characteristics and potential outcomes, or about the relationship between those characteristics and the units' probabilities of treatment assignment. The only identifying assumption for causal inference from an RCT that cannot be verified is the assumption that each unit's potential outcomes are independent of the assignment of other units to the treatment conditions.[158] This "non-interference" or "no-spillover" assumption may be violated in some education research designs. Peer pressures, for example, may cause a student's outcomes to depend on which other students end up in the same treatment group. The spillover problem can often be dealt with by redefining the "unit" as the classroom, school, or school district and randomizing treatment assignment at these levels, where spillover effects among units are less likely.

Though RCTs are the gold standard for causal inference, it is important to recognize that they are not always golden in practice, even if the non-interference assumption is satisfied. *Selective attrition* or *noncompliance* with assigned treatments can bias the results.[159] To illustrate, imagine that a school district wants to figure out whether Magnet High does a better job educating a target student population than Yeoman High. All children in the district take a standardized test at age eighteen, and the administrators are confident that the test scores measure proficiency with respect to the competencies they care about. Since the district sets the rules for student assignment, it could elect to randomly assign members of the target population to Magnet or Yeoman. But if Magnet is widely seen as a better school, ambitious and connected parents of students assigned elsewhere may pressure school district functionaries to bend the *de jure* rules and reassign their student to Magnet. Parents who fail in these lobbying efforts may opt out of public school altogether, going private or choosing home schooling (disappearing from the study). If such selective attrition occurs, and if the children of these ambitious, connected parents tend to have better potential outcomes than other children, then the *observed* average outcome of the students who complete high school at Magnet is likely to be better than the *observed*

the randomized controlled trial, in which the researcher herself assigns units to treatment conditions with explicit, known probabilities.

[157] "Compliance" means that each unit assigned to a treatment condition receives the treatment to which it was assigned.

[158] *See supra* notes 150-52.

[159] For a nice example of the selective attrition problem, see Jennifer Weuve et al., *Accounting for Bias Due to Selective Attrition: The Example of Smoking and Cognitive Decline*, 23 EPIDEMIOLOGY 119, 119 (2012). On noncompliance, see JOSHUA D. ANGRIST & JÖRN-STEFFEN PISCHKE, MOSTLY HARMLESS ECONOMETRICS: AN EMPIRICIST'S COMPANION 117 (2009); IMBENS & RUBIN, *supra* note 101, at 513–59. There are certainly other potential pitfalls too. For an overview, see Diane W. Schanzenbach, *Limitations of Experiments in Education Research*, 7 EDUC. FIN. & POL'Y 219, 225 (2012).

average outcome of the students assigned elsewhere, even if the schools provide exactly the same quality of education.[160]

Researchers can check for selective-attrition problems in a randomized trial by comparing the distribution of background characteristics among students in the treatment and control groups, or among students who received their initially assigned treatment and those who dropped out.[161] But the utility of these checks depends on whether the state's educational records include a rich and relevant array of background information on each student, or, better yet, an identifier that allows the education records to be matched to tax records, criminal justice records, welfare records, and the like.

At the other end of the causal-inference spectrum from randomized trials are research designs that rely on assumptions about how outcomes or probabilities of treatment assignment vary with the background characteristics of the units. For example, in a world where spots at Magnet are assigned on a first-come, first-served basis, a researcher could try to estimate the causal effect of Magnet on the target population by running a regression in which test scores are modeled as a function of the student's race, sex, free-lunch status, and school of attendance. This will give us a predicted "effect" of attending Magnet, but the prediction is only as good as the model. If unobserved student characteristics are correlated with "attending Magnet" and "testing well," our estimate will be biased. Almost surely there are such unobserved traits. Conditional on race, sex, and free-lunch status, students who are more highly motivated, who receive more parental encouragement, and who have greater academic ability are probably more likely to attend Magnet (under a first-come, first-served enrollment rule) and to test well regardless of the school they attend.

Alternatively, we could estimate propensity scores—the *ex ante* probability that a student enrolls in Magnet—using the same demographic characteristics, or simply "match" each student in Magnet High to the most demographically similar student in the non-Magnet group, positing that the units in a matched pair have the same propensity scores or potential outcomes.[162]

But matching and propensity-score methods do not resolve the problem of unobserved traits that may be correlated with treatment assignment and potential

[160] Under certain conditions, it is possible to use treatment assignment as an instrument, producing consistent estimates of the average treatment effect on the population of "compliers." *See* ANGRIST & PISCHKE, *supra* note 159.

[161] The Institute for Education Sciences ("IES"), with the U.S. Department of Education, has developed standards for presumptively acceptable attrition rates and for covariate-balance checks in randomized trials. *See* WHAT WORKS CLEARINGHOUSE: PROCEDURES & STANDARDS HANDBOOK 11–15 (3d ed.), https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_procedures_v3_0_standards_handbook.pdf [hereinafter, WWC HANDBOOK].

[162] *See* MORGAN & WINSHIP, *supra* note 100, at 117–18. The matching-based estimator would also yield an unbiased estimate if matched units had the same propensity score (even if not the same potential outcomes). Matching and propensity-score methods represent a modest advance over regression-based methods, in that they mitigate the so-called "common support problem," which occurs when units in the treatment group have no similar counterpart in the control group, or vice versa. *See id.* at 117. To illustrate, in our running example, the regression could yield very misleading results if black men and black women have systematically different potential outcomes and few black men attend Magnet while few black women attend Yeoman. Matching on observables or propensity scores mitigates the common-support problem because units with no counterpart in the other condition are excluded from the analysis.

outcomes. To continue with our example, if the school district tracks only student race, sex, and eligibility for free lunches, matching strategies for estimating the effect of Magnet High will be more of a stretch than if the district also records, say, each student's performance on an eighth-grade exit exam, the student's record of disciplinary infractions in grade school, and parental income and education. The reason is that, within pairs of units matched on race, sex, and free-lunch status, there is likely to be a lot of variation in both potential outcomes and probability of attending Magnet—with the students who have better potential outcomes being more likely to attend Magnet. If the researcher can match students as well on eighth-grade academic performance, junior-high discipline, and parent income and education, there will probably be much less within-pair variation in true propensity scores and potential outcomes.

In general, then, a richer administrative dataset may allow for better matching or propensity-score estimation. But it is important to recognize that matching and propensity-score estimation are as much arts as they are science.[163] No theorem establishes that matching on more background traits produces better balance in terms of potential outcomes than matching on fewer traits. To the contrary, it is well known that matching on certain kinds of pre-treatment characteristics tends to *reduce* potential-outcome balance between the treated units and the matched controls.[164]

The research-design midfield between randomized trials (gold standard) and outcome- or propensity-score estimation consists of studies that leverage an event or discontinuity in the world that *plausibly* results in the assignment of units to treatments independent of potential outcomes. These include so-called *regression discontinuity* ("RD") designs, *instrumental variable* ("IV") designs, and *natural experiments*.[165] The key advantage of this family of "quasi-experimental" designs over matching, propensity-score estimation, and outcome-estimation approaches is the existence of a *design-based reason* for believing that the groups of treatment and control units are balanced on *unobserved* background characteristics.[166] This is an important advantage because when real-world processes result in sorting correlated with potential outcomes, researchers are unlikely to observe the full set of background characteristics that correlate with potential outcomes.

To illustrate design-based inference with observational data, imagine that all students within a geographic zone, whose family income is less than $25,000/year, are assigned to Magnet. If this were so—*and if families do not*

[163] Propensity score estimation requires arbitrary choices among estimation strategies. For a simulation illustrating bias from mis-specified propensity score models, see *id.*, at 100–04. Matching requires arbitrary choices among algorithms and distance (similarity) metrics. *See* Imbens & Wooldridge, *supra* note 107, at 35–58; MORGAN & WINSHIP, *supra* note 100, at 105–16.

[164] For example, matching on an "instrumental variable"—a variable that is correlated with treatment assignment but not with potential outcomes except through the mechanism of treatment—increases bias. *See* Jeffrey M. Wooldridge, *Should Instrumental Variables Be Used as Matching Variables?* (July 2009) (unpublished paper, Mich. St. Univ.), http://econ.msu.edu/faculty/wooldridge/docs/treat1r6.pdf. The problem faced by researchers who have observational data with information on a set of covariates is that it is not known whether a covariate is an instrument. *See* MORGAN & WINSHIP, *supra* note 100, at 196–97 (illustrating why the identifying assumptions for instrumental variables methods cannot be tested with observational data).

[165] *See* sources cited *supra* note 101.

[166] *Id.*

strategically adjust their income to get their children into Magnet—a researcher might be able to figure out the causal effect of attending Magnet on students in this geographic zone by comparing the outcomes of Magnet students, whose family income at the time of admission was just under $25,000, with the outcomes of Yeoman students, whose family income was just over $25,000. The latter provide the counterfactual, "control condition" outcome for the former. This design will yield an unbiased estimate of the *local average treatment effect* of attending Magnet if potential outcomes vary smoothly with family income at the cutoff—which is plausible provided there is no sorting at the cutpoint.[167] The estimated effect is "local" to families near the cutoff; the regression discontinuity design does not shed light on the effect of attending Magnet for students whose family income is well below, or far above, $25,000.

Regression discontinuity designs can also be used if the administrative breakpoint induces a shift in the probability that a unit is treated, rather than assigning treatment deterministically. In this case, the income cutoff creates an *instrumental variable*—a variable that is correlated with treatment assignment but uncorrelated with potential outcomes, except through the mechanism of treatment.[168] Instrumental-variable designs are widely used in social scientific research. The recent paper by Jackson et al., using court rulings in school-finance cases as an instrument to identify the effect of spending on student outcomes, is a good example.[169] Like regression-discontinuity designs, instrumental-variable approaches yield "local" estimates of treatment effects. The estimate is specific to the subpopulation whose assignment to treatment or control conditions is affected by the instrument.

Though they have a critical, design-based advantage over matching, outcome-estimation, and propensity-score analyses of observational data, RD and IV methods are distinctly second-best compared to the gold-standard design for causal inference, a randomized controlled trial. First, in RD and IV designs, the critical identifying assumptions are always a matter of some conjecture, whereas in RCTs the researcher *knows* that treatment assignment is (in expectation) independent of potential outcomes. Second, the local estimands of RD and IV designs are usually less normatively important than the average treatment effect estimand of an RCT. We want to know whether Magnet High leads to better outcomes for a wide swath of eligible students, not just those who are near the cutoff for admission or whose attendance decision is affected by an instrument. Finally, RD, and many IV designs, tend to have limited statistical power. So, even if Magnet High really does provide big benefits for students near the admissions cutoff, the study may fail to detect it.[170]

* * *

[167] The "essentially random" assumption is the identifying assumption for so-called "fuzzy" regression discontinuity designs, and the "potential outcomes vary smoothly at the cutoff" assumption is the identifying assumption for "sharp" designs. *See* Imbens & Wooldridge, *supra* note 107, at 62–63. Note that if some families adjust their reported income strategically to be just below the cutoff, the identifying assumptions are likely to be violated.

[168] On instrumental variable designs, see IMBENS & RUBIN, *supra* note 101, at 513–84; MORGAN & WINSHIP, *supra* note 100, at 187–217; Imbens & Wooldridge, *supra* note 107, at 56–61.

[169] *See generally* Jackson et al., *supra* note 46.

[170] The same goes for instrumental variable designs. *See* MORGAN & WINSHIP, *supra* note 100, at 198.

How does the problem of causal inference implicate the state? Government decisions largely determine whether the practicable set of research designs for estimating the causal effects of educational policies and programs include "strong" designs. This is so, first, because of state actors' control over educational treatment assignment, and, relatedly, because state policies determine whether it is easy or hard for units to escape from their assigned treatments. Principals determine the assignment of teachers and students to classrooms. School districts and state departments of education assign curricula and professional development opportunities to teachers. State legislatures assign teacher tenure rules, school-level accountability systems, and funding. In everyday usage, it is strange to describe teachers, schools, curricula, budgets, peers, disciplinary protocols, labor laws, and school accountability systems as "treatments" when they are not assigned pursuant to an explicit experimental protocol, but from the point of view of researchers (or judges) trying to figure out whether such state actions cause certain outcomes, treatments are what they are.

For purposes of enabling causal inference, two treatment-assignment decisions are particularly critical: Whether to assign treatments to units with explicit, known probabilities and whether to use numerical cutoffs to determine assignment. If the state uses the former approach, then researchers will be able to estimate average treatment effects on the population of units with treatment assignment probabilities between zero and one, subject only to the non-interference assumption. If the state uses the cutoff alternative, researchers may be able to estimate local treatment effects using the second-best alternative of a regression-discontinuity design. If treatment assignment is neither explicitly probabilistic nor tied to hard cutoffs, causal inference will depend on the guesswork of matching, propensity score estimation, or regression. The worst situation of all is universal assignment, *i.e.*, all units receiving the treatment. In this case, causal inference is likely to depend on strong assumptions about time trends in potential outcomes.[171]

An appreciation of the fundamental problem of causal inference also reinforces the importance of rich, record-linked administrative datasets, for reasons that go well beyond the identification of relevant outcome variables. Researchers using observational data need to be able to match students or estimate propensity scores or outcomes using pre-treatment characteristics. How plausible these efforts are depends on the richness of the educational database, in combination with any linkable dataset. Similarly, for researchers analyzing experimental or quasi-experimental data, a rich dataset of pre-treatment covariates allows for stronger checks of whether randomization worked as planned, whether units "sorted" at the cutpoint in a regression discontinuity design, and whether a putative instrument induced some variation in treatment assignment that appears to be uncorrelated with other characteristics of the units.

The existence of big administrative datasets can also enable retrospective studies that are more causally credible than matching or propensity-score studies. Examples include regression-discontinuity designs and "natural experi-

[171] *See id.* at 244–49 (discussing interrupted time-series designs).

ments" leveraging events that may have affected treatment assignment independent of potential outcomes.[172] Regression-discontinuity designs tend to have little statistical power, so they depend on large datasets. For similar reasons, studies that leverage idiosyncratic local events as "instruments" for treatment assignment require large datasets with broad geographic coverage.[173] Chetty's work on the geography of socioeconomic mobility is again instructive. Because he had a huge taxpayer dataset covering the entire nation, he could use observations from localities that had suffered adverse economic shocks to construct plausible estimates of the causal effect of locality on mobility.[174] If he were using a purpose-built dataset consisting of a random sample of the U.S. population, there would probably be just a few observations from localities that had suffered big economic shocks.

Finally, the strength of the feasible set of research designs depends upon the state because the state determines not only the content of educational tests, but also who must take these tests and under what conditions. Some educational treatments may induce certain students to sort among different types of schools, for example, by switching from public to private schools. Some educational treatments may give administrators an incentive to reclassify students in ways that could affect which tests the students take and under what conditions. For example, the school-based accountability rules established by the federal No Child Left Behind Act seem to have encouraged schools to reclassify weaker students as disabled.[175] Any sorting into or out of testing that is correlated with both treatment assignment and potential outcomes will bias treatment effect estimates. For purposes of the education right, research designs that use test scores as the outcome variable are a pale substitute for designs that use adult outcomes, and the substitute is much paler yet if students sort into or out of testing in ways that are correlated with treatment assignment and potential outcomes.

C. Bias in the Research Enterprise

The barriers to causal inference we have discussed thus far are largely barriers *to what a researcher can learn*, given the available data and the rules governing treatment assignment. But for courts or legislatures to enforce the education right, a further, and subtly different, question must also be considered: *what can third parties learn from researchers' findings*? Of particular concern are biases in the research enterprise, which may cause the distribution of published findings to present a misleading picture of actual treatment effects. Such biases can result from elite journal norms about what makes a study publication-worthy, from academics' professional incentive to produce the kind of work that top journals want to publish, and from the existence of ideological uniformity within social scientific disciplines. As we will see, the state can play an important role

[172] Figlio et al., *Education Research*, *supra* note 115, at 10.

[173] The credibility of these studies, as manifested through checks of covariate balance between treatment and control groups, also depends on the information in the state's databases.

[174] Chetty et al., *Where is the Land of Opportunity?*, *supra* note 21, at 1553.

[175] For a review of the literature, see David Figlio & Susanna Loeb, *School Accountability*, *in* 3 HANDBOOK OF THE ECONOMICS OF EDUCATION, 383, 395 (Eric A. Hanushek et al. eds., 2011).

counteracting bias in the research enterprise through the terms on which it provides access to administrative data.

Journal Norms & File Drawers.[176] It is common wisdom that the publication or placement value of an empirical study depends on whether the findings are statistically significant and novel.[177] Even if no researcher fudges anything, the requirement of a "significant" finding means that the distribution of *published* effect sizes and statistical significance levels will be exaggerated and unrepresentative of the distribution of effect sizes and significance levels *actually obtained* in the full set of statistical analyses run by researchers. Studies that find small and insignificant effects get stashed away in the file drawer; studies that find big, significant effects get published. When this phenomenon occurs, reported tests of statistical significance are not credible.

The standard representation of statistical significance—a "p-value"—quantifies how improbable the observed data would be if the null hypothesis (usually the hypothesis of no treatment effect) were true.[178] A p-value of 0.05, the conventional cutoff for statistical significance, means that if the null hypothesis of no effect were true and the study was replicated twenty times, then, on average, only one of those twenty replications would yield an estimated treatment effect as big as the effect reported in the original study.[179]

Now, imagine that Treatment A has no average effect relative to Treatment B on the population of, say, elementary school students in the United States. In a world with many researchers, twenty independent investigations of Treatment A might well be undertaken. In expectation, one of these twenty estimates will be significant at the 5% level. If this study gets published and none of the others see the light of day, policy-makers, judges, and other third-party consumers of the research literature are likely to mistakenly infer that Treatment A has a real effect relative to B.

Now consider a different world, in which Treatment A has a real, positive effect relative to B. The first researchers who study these treatments will probably figure this out and publish the results, after which additional studies showing positive effects of A relative to B will be considered less interesting and will be harder to publish. On the other hand, a study finding that Treatment B outperforms Treatment A—a study that runs against previous findings—would be novel and very publishable. With enough replications of the experiment, eventually random chance will result in an unusual distribution of units across treatment conditions (with potential outcomes much lower in the group assigned to

[176] The problem described in the next two paragraphs has been documented in numerous scientific and social scientific disciplines. See, for example, the studies cited in Justin McCrary et al., *Conservative Tests Under Satisficing Models of Publication Bias*, PLOS ONE, Feb. 22, 2016, at 1 and the papers collected in PUBLICATION BIAS IN META-ANALYSIS: PREVENTION, ASSESSMENT AND ADJUSTMENTS 111–12 (Hannah R. Rothstein et al. eds., 2005).

[177] The common wisdom is backed by at least one quality experiment. *See* Gwendolyn B. Emerson et al., *Testing for the Presence of Positive-Outcome Bias in Peer Review: A Randomized Controlled Trial*, 170 ARCHIVES INTERNAL MED. 1934, 1934 (2010).

[178] For an accessible explanation of this idea, see SEAN GAILMARD, STATISTICAL MODELING AND INFERENCE FOR SOCIAL SCIENCE 236–89 (2014).

[179] *Estimated* average treatment effects will bounce around from replication to replication whenever there is heterogeneity in potential outcomes across the units, assuming random assignment of units to treatments.

A than to B), leading to an estimated *negative* effect of A relative to B that is "statistically significant."

The upshot is that in the world where A has no effect relative to B, the initially published studies may well show that it does, and in a world where A has positive effects relative to B, the distribution of published effects may present a more ambiguous picture.

Professional Incentives. Professors have financial and reputational incentives to produce the kind of work that elite journals value. Because statistically insignificant findings are hard to publish, researchers are encouraged to do what critics call "p-hacking"—conducting multiple hypothesis tests, constructing lots of different statistical models, and slicing and dicing the data in lots of different ways until a result pops up whose nominal p-value is less than 0.05.[180] When researchers run multiple statistical tests they *should* make adjustments to account for the likelihood of chance results,[181] but because these adjustments often reduce the test statistic to insignificance, the researcher has little incentive to make the adjustment unless a third party insists on it. Similarly, when researchers subset the data to look for heterogeneous treatment effects, they *should* report their initial plan and the number of subgroups they examined. But third parties have no way of observing how many tests the researcher ran, how many model specifications the researcher tried, or how many ways the researcher subsetted the data before discovering the results that appeared in the published paper.[182]

The upshot is that journal publication norms not only result in a distribution of published findings that is unrepresentative of the distribution of actual findings, but also in a distribution of actual findings that is unrepresentative of the findings that would be produced in a world where researchers were rewarded for, say, the quality of their research designs rather than the putative "statistical significance" and the novelty of their findings.[183]

These are not hypothetical problems.[184] Researchers across a number of scientific and social scientific fields have documented selective reporting and publication of results.[185] The distribution of reported p-values in published papers is often abnormally clustered just below 0.05, the conventional threshold for a publishable result.[186] A massive recent project to replicate one hundred leading studies in social psychology found average effect sizes only half as large

[180] *See* Uri Simonsohn et al., *P-Curve: A Key to the File-Drawer*, 143 J. EXPERIMENTAL PSYCHOL.: GEN. 534, 534 (2014).

[181] For an overview, see FRANK BRETZ ET AL., MULTIPLE COMPARISONS USING R xiii (2011).

[182] John P.A. Ioannidis, *Why Most Published Research Findings Are False*, PLOS MED., Aug. 2005, at 701 ("[U]sually it is impossible to decipher how much data dredging by the reporting authors or other research teams has preceded a reported research finding.").

[183] *Id.* at 698 ("[T]here is strong evidence that selective outcome reporting, with manipulation of the outcomes and analyses reported, is a common problem even for randomized trails.").

[184] For recent and reasonably exhaustive lists of research on publication bias and p-hacking across the scientific and social scientific disciplines, see C. Glenn Begley & John P.A. Ioannidis, *Reproducibility in Science: Improving the Standard for Basic and Preclinical Research*, CIRCULATION RES., Jan. 2, 105, at 119 tbls.1 & 2; McCrary et al., *supra* note 176, nn.2–30.

[185] *See, e.g.*, An-Wen Chan et al., *Empirical Evidence for Selective Reporting of Outcomes in Randomized Trials: Comparison of Protocols to Published Articles*, 291 J. AM. MED. ASS'N 2457 (2004) (surveying clinical trials in medicine); Annie Franco et al., *Publication Bias in the Social Sciences: Unlocking the File Drawer*, 345 SCIENCE 1502 (2014) (surveying experiments in social science).

[186] *See, e.g.*, Simonsohn et al., *supra* note 180, at 534.

as those originally reported.[187] Pharmaceutical companies regularly undertake their own replications of basic research before investing in drug development because published findings in top scientific journals so often fail to replicate.[188]

Ideology. It is no secret that political liberals are vastly overrepresented in the social sciences, relative to their numbers in the U.S. population.[189] This provides a further reason to be suspicious of published findings, at least if the findings corroborate liberal orthodoxies. A small body of work shows that researcher ideology (or personal characteristics associated with ideology) is suspiciously correlated with researchers' findings.[190] Indeed, asked to review otherwise identical studies that differ in their results, researchers tend to deem the study more credible if the results accord with their ideology or theoretical perspective.[191] Thus, if most researchers in a field subscribe to similar political ideologies, the likely result is further inflation of reported effect estimates with respect to treatments favored by the dominant ideology. This inflation of ideologically congruent "effects" will occur even if no researcher intentionally manipulates her analysis to corroborate her ideological preferences. It is a natural byproduct of researchers' tendency to accept ideologically congruent results uncritically while probing incongruent findings to see if they are robust.[192]

* * *

School boards and education departments are certainly not responsible for bias in the research enterprise. But, because these entities control access to educational data and provide some research funding, they are well positioned to mitigate the problems. To see this, consider the features that would render a study credible, given the biases we have described. The most credible studies have the following characteristics:

- The results of the study were guaranteed to be publicly available regardless of what the researchers found. (Nothing left in the file drawer.)

[187] Open Science Collaboration, *Estimating the Reproducibility of Psychological Science*, 349 SCI. 943, 943 (2015).

[188] Begley & Ioannidis, *supra* note 184, at 117; *see also* Bruce Booth, *Academic Bias and Biotech Failures*, LIFE SCI. V.C. (Mar. 28, 2011), https://lifescivc.com/2011/03/academic-bias-biotech-failures/ (arguing that campus technology transfer offices should invest in study reproduction by independent labs).

[189] *See* NEIL GROSS, WHY ARE PROFESSORS LIBERAL AND WHY DO CONSERVATIVES CARE? 63 tbl.1.1 (2013) (summarizing research and reporting that, per 2006 survey, the greatest concentration of progressive and radical faculty within the academy is in the social sciences).

[190] For a review focusing on social psychology, see generally Philip E. Tetlock & Gregory Mitchell, *Why So Few Conservatives and Should We Care?*, 52 SOC'Y 28 (2015). For evidence from economics, see generally Zubin Jelveh et al., *Political Language in Economics* (Columbia Bus. Sch., Research Paper No. 14-57, 2015), http://dx.doi.org/10.2139/ssrn.2535453.

[191] *See, e.g.*, Stephen I. Abramowitz et al., *Publish or Politic: Referee Bias in Manuscript Review*, 1975 J. APPLIED SOC. PSYCH. 187, 187 (1975); Stephen J. Ceci et al., *Human Subjects Review, Personal Values, and the Regulation of Social Science Research*, 40 AM. PSYCHOLOGIST 994, 994 (1985) (reporting results of field experiment showing political bias by institutional review boards); *see also* Michael J. Mahoney, *Publication Prejudices: An Experimental Study of Confirmatory Bias in the Peer Review System*, 1 COGNITIVE THERAPY & RES. 161, 161 (1977).

[192] Social psychologists call this "confirmation bias" or "motivated reasoning." *See, e.g.*, José L. Duarte et al., *Political Diversity Will Improve Social Psychological Science*, 38 BEHAV. & BRAIN SCI. 1, 7–8 (2015). As Duarte et al. also note, ideological homogeneity within a social scientific field may also affect the questions that get investigated and, of course, the interpretation of results. *Id.* at 4.

- Any record-linkage, matching of treatment and control units, propensity score estimation, or modeling of potential outcomes was done by researchers who were blind to the data they would need to jury-rig the findings.[193]
- The researchers described their hypotheses and plan of analysis before they saw the outcome data. Such "pre-analysis plans"—which have become commonplace in medicine, but are only starting to take hold in the social sciences—allow third parties to distinguish credible results from post-hoc explorations of the data, and to judge the appropriateness of any adjustment (or lack of adjustment) for multiple-hypothesis testing.[194]
- The researchers' reward for conducting the study was not dependent on journal placement or otherwise tied to the size, direction, or statistical significance of the findings.
- The sample size was large, and the principal finding was an average effect for the full sample, rather than a subgroup effect.[195]
- The research team was ideologically diverse.

So what can state education departments do about this? At the very least, the department can insist that any researcher who wishes to use its administrative data register a pre-analysis plan prior to receiving the data. The Society for Research on Educational Effectiveness ("SREE") is currently developing a study registry and pre-analysis standards for education research.[196] School districts and state education agencies will soon be able to piggyback on the SREE protocols. Education departments can also set up procedures for staggered data releases, thereby enabling researchers to credibly signal that record-linkage, matching, propensity score estimation, or outcome estimation was carried out blind to the information the researcher would have needed for an unprincipled, results-driven analysis.

State education agencies can also mitigate bias in the research enterprise through their funding and grant-making programs. If the agency wants to figure out whether particular programs are actually working, the agency may do better to hire a research company like Mathematica or the Rand Corporation rather than sourcing the work to academics, as the research firms are probably less affected by incentives to publish findings in a top journal or to confirm a theory for which the researcher has become known.

[193] Matching or propensity-score estimation should be executed by researchers without access to the outcome variable. If the study design requires estimation of counterfactual outcomes, the research team might be provided initially with data just for the control units (or just for the treatment units). Only after the team finalizes its model for counterfactual outcomes based on units in one treatment condition would the researchers be allowed to see the outcome data for units in the other condition.

[194] *See generally* Benjamin A. Olken, *Promises and Perils of Pre-Analysis Plans*, 29 J. ECON. PERSP. 61 (2015).

[195] *See* Ioannidis, *supra* note 182, at 697 (demonstrating that, other things equal, smaller samples reduce the "post-study probability" that a positive finding is a true finding).

[196] *See Funding Opportunities: Strengthening Education Research Through Professional Development and a Trial Registry*, INST. EDUC. SCI. (2015), https://ies.ed.gov/funding/grantsearch/details.asp?ID=1757 (last visited Jan. 16, 2018); *Panel Discussion Introduction: A Registry of Effectiveness Studies in Education*, SOC'Y FOR RES. ON EDUC. EFFECTIVENESS (Spring 2016), https://www.sree.org/conferences/2016s/program/downloads/abstracts/1849_organizer.pdf; *Unsolicited and Other Awards: Strengthening Education Research Through Professional Development and Trial Registry*, NAT'L CTR. FOR EDUC. RES. (2015), http://ies.ed.gov/ncer/projects/grant.asp?ProgID=25&grantid=1757 (last visited Jan. 16, 2018).

Finally, by providing researchers with access to administrative datasets, state education agencies can enable high-power, large-N research designs in which findings of putatively big effects are less likely to be chance or researcher-manipulated occurrences.

D. Summary

In light of the manifest limitations of much putatively scientific education research, it is perhaps understandable that courts in education-quality cases have often rested their decisions on decidedly nonscientific grounds.[197] But as this Part has shown, the availability or absence of credible estimates of the causal effects of educational programs and policies on the outcomes that ground the education right is not simply a brute fact over which courts and other state actors have no control. The state affects both the availability and the credibility of such estimates, with policies (or a lack of policies) that bear on each link in the useful-causal-research chain.

Our account of the useful-causal-research chain, from which we derive our claims about state control over the production of constitutionally relevant knowledge, is hardly idiosyncratic. Though lawyers and legal scholars working on education rights have not focused on these issues,[198] policy-makers, government agencies, foundations, and social scientists are very actively grappling with

[197] *See supra* Section II.B.

[198] For example, the first education law casebook, DEREK W. BLACK, EDUCATION LAW: EQUALITY, FAIRNESS, AND REFORM (2013), spends many pages discussing evidentiary and institutional competence issues in education-quality cases but does not address state control over the production of constitutionally relevant knowledge, and only fleetingly alludes to problems of causal inference. *Id.* at 249–53 (discussing "disagreement" about "[h]ow to measure funding gaps and their importance," without addressing problem of causal inference or need for linked administrative datasets"); *id.* at 253–57 (discussing research on long term effects of education without addressing administrative datasets, record linkage, or problems of causal inference); *see also* Black, *supra* note 58, 1757–67 (arguing that education cases present exceedingly difficult causal questions owing to the complexity of education and the multiple factors that may influence student outcomes, but failing to consider whether—and if so, how—the state determines what can be learned about these causal questions). Similarly, the prominent book about education rights by Michel Rebell, a leading plaintiff-side litigator, says nothing about administrative data systems, requirements for causal inference, or bias in the research enterprise. *See generally* MICHAEL A. REBELL, COURTS AND KIDS: PURSUING EDUCATIONAL EQUITY THROUGH THE STATE COURTS (2009). The principal law review article devoted to evidentiary issues in education-rights cases is Superfine, *supra* note 70. It nicely summarizes the history of the federal role in education research, *id.* at 672–88, and debates about what constitutes "science" in education research, *id.* at 689–92, but it fails to identify and discuss the forms of control that the state exercises over the production of constitutionally relevant knowledge. Prescriptively, Superfine urges education researchers to engage in "deep[] analy[sis]" sensitive to "nuanced differences in context," *id.* at 690–91, in order to generate "broad, evidence-based principles that can be actively applied by educators to . . . contextual interactions and ever-changing environments," *id.* at 696, and he proposes that courts regulate education "governance" to ensure that state and local decisions about the allocation of educational resources are responsive to these principles, *id.* at 697–700. In addition to overlooking the particular forms of state control over constitutionally relevant knowledge, this prescription puts cart before horse: the deep, general, convincingly proven "principles" about education desired by Superfine do not exist, so instead of imagining how courts might hold states and school districts to such principles, it is more productive to ask how courts might *concretely* enable the development of constitutionally relevant knowledge. *See supra* Part IV; *see infra* note 211(discussing limits of the "democratic experimentalism" theory to which Superfine ties his prescription). The book-length critique of education-rights jurisprudence co-authored by Eric Hanushek, the leading defense-side expert witness, pays attention to "link 2" in the useful-causal-research chain and also touches on the importance of state data systems, but he fails to consider whether the state may have a justiciable duty to improve the base of evidence. *Compare* HANUSHEK & LINDSETH, *supra* note 4, at 211–16, 247–50 (discussing limitations of available evidence and opportunities for improvement), *with* HANUSHEK & LINDSETH, *supra* note 4, at 281–

them. As we noted above, the Data Quality Campaign ("DQC") has, since 2005, been advocating for policies to improve educational records and enable linkage of educational and workforce databases.[199] Congress, through the America Competes Act of 2007, endorsed most of DQC's recommendations,[200] and Congress required states to commit to many such policies as a condition for receiving fiscal stabilization funds under the American Recovery and Reinvestment Act of 2009.[201]

Meanwhile, the fundamental problem of causal inference has been central to the work of the recently established Institute for Education Sciences ("IES"), within the Department of Education.[202] As the IES's founding director Grover Whitehurst quipped, the No Child Left Behind Act's mandate that states use the "best scientific research" was akin to "growing food by decree in the old Soviet Union."[203] Most research produced by schools of education simply had not attended to problems of causal inference.[204] IES responded to this state of affairs by publishing research method guidelines,[205] by reviewing and rating published research for consistency with scientific research norms,[206] and by disseminating high-quality studies of what works to practitioners and policy-makers.[207] Recently, IES has turned its attention to the deep threats posed by bias in the research enterprise, teaming with the Society for Research on Educational Effectiveness to develop pre-analysis registries and standards for education research.[208]

But is any of this of *legal* relevance? Do state constitutions create justiciable obligations to facilitate research that would illuminate the constitutional quality of educational systems? We turn to this next.

87 (arguing for an advisory judicial role in which courts would make "findings about mismanagement, waste, inefficient practices, constraints imposed by collective bargaining agreements, state tenure laws, and so on," but not order any remedies). Note the irony: the question of whether tenure laws, for example, undermine the quality of instruction presents exactly the same difficult problem of causal inference as the question of whether ostensibly insufficient funding undermines education quality. Hanushek and Lindseth pillory courts for their reliance on dubious costing out studies to make findings about the effects of money, *see* HANUSHEK & LINDSETH, *supra* note 4, at ch. 7, but treat the effects of labor laws as an issue that courts can simply resolve using their "investigative and fact-finding expertise and authority," *id.* at 285.

[199] *See supra* notes 132–43 and accompanying text.

[200] *See* DATA QUALITY CAMPAIGN, FACT SHEET, *supra* note 132, at 2.

[201] *See* DATA QUALITY CAMPAIGN, *supra* note 134.

[202] *See generally* WWC HANDBOOK, *supra* note 161 (explaining criteria by which the IES's public-information wing "scores" published studies).

[203] Whitehurst, *supra* note 31, at 107–13.

[204] *Id.*

[205] *See* WWC HANDBOOK, *supra* note 161.

[206] For reviews and ratings, see *What Works Clearinghouse*, *supra* note 32.

[207] *See generally* Whitehurst, *supra* note 31. The IES has been criticized by some researchers for being insufficiently attentive to qualitative research (including the need for a strong qualitative component in many experimental designs) and for focusing too much on what can or cannot be learned from individual studies rather than the "big picture" of the literature as a whole. *See, e.g.*, Alan H. Schoenfeld, *What Doesn't Work: The Challenge and Failure of the What Works Clearinghouse to Conduct Meaningful Reviews of Studies of Mathematics Curricula*, 35 EDUC. RESEARCHER 13 (2006). These criticisms may have some validity, but the IES has still played a hugely important role in propounding good research norms and highlighting important studies.

[208] *See supra* note 185 and accompanying text.

IV. CONSTITUTIONAL DUTIES AND THE PRODUCTION OF KNOWLEDGE ABOUT EDUCATION

The legal implications of the state's control over the production of knowledge about how to educate disadvantaged children effectively depend, of course, on one's theory of the right to education. For reasons explained in Section II.C, we generally agree with courts that have treated the education clauses as creating a justiciable duty of care and that have adopted deferential but not toothless standards of review, akin to administrative law's arbitrary-and-capricious standard.[209]

The state could violate this duty of care by failing to adhere to certain substantive "best practices" about how to fund and operate a system of public schools, or by failing to investigate (or to enable others to investigate), which actual or potential reforms would make the schools work better for disadvantaged children. Let's call these the *operational* and the *knowledge-frontier* components of the duty of care. The state's balance of responsibilities between these components depends on two factors: (1) the current state of knowledge about how to educate disadvantaged children effectively, vis-à-vis the constitutional objectives of the education system, and (2) the extent of state control over production of that knowledge.

In a world where the state had no control over the frontier of knowledge, the duty of care would only require the state to attend to operational matters, which by definition the state controls.[210] Similarly, in a world in which the effects of the set of practicable reforms were very well understood, the state's knowledge-frontier responsibilities would be very weak, and the state's duty to adopt operational best practices would be commensurately strong. But in a world of great uncertainty about how to help disadvantaged children realize better long-run outcomes, the state's obligation to adopt what are regarded at a given point in time as best-practices is quite weak, and the duty of reasonable care requires the state to devote a lot of attention to figuring out what might be done better—insofar as this is within the state's control. *This is our world today*. As Part II explained, a number of recent studies point to tantalizing possibilities for improving the education of disadvantaged children, but there is a dearth of knowledge about long-run effects, about scaling up interventions that have worked at small scales, and about the policy levers that would most effectively distribute critical inputs, such as high-quality teachers, to disadvantaged students. And per Part III of this Article, advances along the knowledge frontier absolutely depend upon the state.

In this Part, we discuss how the duty of care with respect to knowledge production could be enforced through litigation. Three lines of attack are open today. First, analogizing to "framework" responsibilities that courts have already recognized, litigants could argue that state failures to adopt non-arbitrary *knowledge-production plans* violate the duty of care. Second, litigants should be

[209] *See* Bauries, *supra* note 77, at 706.

[210] State constitutions require a system of *public* schools. *See* Emily Parker, *Constitutional Obligations for Public Education*, EDUC. COMM'N OF THE STATES (Mar. 2016), https://www.ecs.org/ec-content/uploads/2016-Constitutional-obligations-for-public-education-1.pdf.

able to attack discrete components of state and local educational and record-keeping systems on the ground that they hinder the production of knowledge for no good reason. Finally, in certain hard cases, where litigants have a strong argument for a proposed intervention but have been unable to produce credible estimates of its likely effect because of the state's control over treatment assignment, it may be appropriate for courts to issue a *temporary-experiment remedy*, rather than entering a permanent, state-wide injunction or letting the state off the hook entirely.[211]

A. Framework Duties

We observed in Part II that several courts have demanded that the state promulgate educational standards and establish testing protocols calibrated to those standards.[212] These precedents recognize that the state must make some effort to gather information about the performance of the educational system. Put slightly differently, the duty of care is not only a duty of operational care; it also requires that the state act reasonably with respect to the knowledge frontier.

This is a sensible conception of the legislative duty, but the courts' focus on testing is far too narrow. Essentially, the courts have addressed *one component* (observing outcomes) of *one link* (the first) in the useful-causal-research chain, ignoring the rest of it. Yet unless those outcomes can be associated with treatments, and the effects of the treatment on the outcomes credibly estimated, the state's gathering of outcome data is not going to help the state, or anyone else, learn whether the educational system is performing as well as it reasonably could perform. Moreover, scores on standardized tests are, at best, intermediate outcomes on the road to the constitutionally important outcomes: participation

[211] The proposals in this Part can be understood as refining an idea developed some years ago by a group of scholars generally known as democratic experimentalists, including Michael Dorf, James Liebman, Charles Sabel, William Simon, and others. Sabel and Simon famously asserted that courts should undertake to destabilize and reform public institutions, such as schools, police forces, and prisons, upon finding that the institution had suffered gross failures of performance and has become insulated from correction through the normal democratic process. *See, e.g.*, Michael C. Dorf & Charles F. Sabel, *A Constitution of Democratic Experimentalism*, 98 COLUM. L. REV. 267, 267 (1998); James S. Liebman & Charles F. Sabel, *A Public Laboratory Dewey Barely Imagined: The Emerging Model of School Governance and Legal Reform*, 28 N.Y.U. REV. L. & SOC. CHANGE 183, 184 (2003); Charles F. Sabel & William H. Simon, *Destabilization Rights: How Public Law Litigation Succeeds*, 117 HARV. L. REV. 1015, 1017 (2004). At the remedy stage, the court is to convene a working group of stakeholders to develop performance metrics and adequacy standards, all of which would remain provisional and subject to revision as better information becomes available. Sabel & Simon, *supra*, at 1067–73. We share the democratic experimentalists' hope for a court-prodded process of learning about how to operate public schools more effectively, but we part ways with the experimentalists as to doctrine and remedies. In a world of very limited information about the causal effect of alternative educational programs on constitutionally relevant outcomes, it is a dubious task for courts to say whether an educational system has "grossly failed." A school system has failed badly only if some alternative way of running the schools would have resulted in much better outcomes for the students, and by hypothesis the causal effects of those alternative arrangements on student outcomes are not well understood. Nor do we think that judicial convening of a working group of stakeholders will generally suffice to bring into being the information needed to improve a public institution. Better information may be very threatening to some stakeholders. Rather than relegating knowledge production to a consensual, remedial process, we would focus liability-stage adjudication on concrete failures of the state to enable the development of evidence about the causal effects of educational policies and programs on the adult outcomes that ground the education right.

[212] *See supra* notes 89–91 and accompanying text.

as an adult in economic, political, and civic life. Because the relationship between the intermediate and adult outcomes is uncertain, and because high-stakes testing may induce teachers to substitute test-prep lessons for more productive forms of instruction,[213] courts should give legislators and school administrators a very wide berth to decide what educational standards to adopt and how to measure student achievement.

Though the states should be given considerable leeway as to the details of standards and measurements, courts should not permit the state (bound by the duty of care) to say, in effect, "we're just not interested in whether our educational programs are actually helping disadvantaged students develop into productive members of society, who participate fully in economic, civic, and political life."

Thus, rather than compel the state to establish educational standards and a testing regime calibrated to those standards, courts should require the state to adopt and periodically update a reasonable *knowledge-production plan* concerning the constitutional quality of the educational system. Standardized tests might be one component of this plan, but they would not comprise its central focus. An adequate knowledge-production plan would identify major gaps in the state's current understanding of the effects of educational policies and programs on disadvantaged children, diagnose the state's role in enabling or thwarting the filling of those gaps,[214] and explain what the state plans to do about it. Courts would enforce the state's duty of care with respect to the knowledge frontier largely through arbitrariness review of this plan, and of the steps that the state takes, or fails to take, to implement it.

This is an auspicious time for state courts to establish a knowledge-production planning requirement, because of related institutional developments at the federal level. First, as we noted above, the IES within the U.S. Department of Education is creating benchmarks for high-quality research on which state courts can piggyback.[215] The second, and related, development has been the improvement of state data systems, with support from the federal government and the nonprofit DQC.[216] States need not march in lockstep with the IES, let alone the DQC, but it would probably be arbitrary for a state's knowledge-production plan not to reckon with IES-promulgated best practices, congressionally approved standards for data systems, and the IES's account of major gaps in research.[217]

[213] *See supra* notes 144–47 and accompanying text.

[214] *See supra* notes 93–98 and accompanying text.

[215] On IES benchmarks, see Superfine, *supra* note 70, at 686; Whitehurst, *supra* note 31, at 110–11. Superfine also recognizes that state courts may, for some purposes, draw upon IES benchmarks, though he warns against this if the standards are "politicized." Superfine, *supra* note 70, at 699. He rather puzzlingly derides the privileging of causally oriented research as an instance of inappropriate "politicization." *Id.* at 692.

[216] *See supra* notes 132–43 and accompanying text.

[217] When No Child Left Behind was passed in 2001, many commentators observed that the standards included in the law could become the Archimedean point that gave a clear standard to state-level adequacy litigation. *See, e.g.*, Michael Heise, *Adequacy Litigation in an Era of Accountability in* SCHOOL MONEY TRIALS, *supra* note 16, at 262–77 (critically noting trend of plaintiffs appealing to federal standards); Martin R. West & Paul E. Peterson, *The Adequacy Lawsuit: A Critical Appraisal*, *in* SCHOOL MONEY TRIALS, *supra* note 16, at 8. One way to frame our point here is to say that we think that the federal efforts to improve the production of information about education is better positioned to serve as a more modest Archimedean point.

The state interests weighing against the development of a knowledge-production plan are very weak—certainly much weaker than the interests weighing against mandatory testing, which many courts have already required.[218] The issuance of a plan puts nobody's privacy at risk, requires no one to take a demoralizing test, and obligates no schools to use particular curricula. And it need not cost the state a lot of money. Much of the architecture for such a plan could be taken "off the rack" thanks to the work of the IES and DQC. Education is the single largest line item in most states' budgets.[219] Surely it is worth devoting a few million dollars and some executive and legislative attention to figuring out what the state might do to spend the education budget more effectively.

B. Arbitrariness Claims

Whether or not courts recognize a legislative duty to promulgate a framework plan for knowledge production, litigants should be able to attack discrete components of state or local educational and data systems on the ground that they arbitrarily hinder the production of knowledge about how the educational system affects the constitutionally relevant outcomes of disadvantaged students.

For example, flat prohibitions on linking educational records to other administrative records ought to be held unconstitutional as violating the state's role as a fiduciary under the education clauses.[220] Similarly, states that presently disallow research using critical outcomes datasets, such as voter registration files, would have to make exceptions for research about the effects of educational policies and programs.[221] It is arbitrary for a state that has an affirmative duty to prepare children for future employment and democratic participation to prevent researchers from using the state's data to study whether the various educational treatments to which students are exposed affect their future employment and propensity to vote.

A judicial determination that categorical prohibitions on record-linkage are unconstitutional would not require the legislature to spend any money or to comply with any judge-made procedures. Nor would this determination depend on the court's evaluation of any legislatively or administratively prescribed arrangement for balancing research interests and privacy interests.

More difficult cases may arise concerning the terms on which the state provides researchers with access to administrative data, or about the state's failure to include certain unique identifiers, such as social security numbers, in administrative records.[222] Though the state must make *some* effort to enable research about the effects of educational policies and programs on lifetime outcomes, the state also has a legitimate interest in protecting sensitive personal

[218] *See supra* cases cited in notes 73–4.

[219] *State & Local Government Finance Snapshot*, U.S. CENSUS BUREAU (July, 2016), https://www2.census.gov/govs/g13-alfin.pdf.

[220] *Cf.* Figlio et al., *Education Research*, *supra* note 115.

[221] For example, Massachusetts currently allows political parties, but not researchers, to use the voter registration files. *See Massachussets*, *supra* note 143. Other states make the voter files available to researchers but withhold information that would greatly aid linkage to education records, such as social security numbers and birth dates (Alaska, for example, see *Alaska, supra* note 143).

[222] *Cf. supra* Section III.A (discussing record linkage).

information from accidental disclosure. This is an area in which both the risks from data leaks and the opportunities for enabling research while protecting privacy are changing rapidly.[223] So long as the legislature or education department attends to these developments and gives a reasonable explanation for its choice, courts should hesitate to displace it.

Occasionally, state decisions that affect the assignment of educational treatments may also prove arbitrary. Consider school lotteries. Many school districts give parents and guardians of school-age children some element of choice with respect to the school their child attends.[224] Spaces in oversubscribed schools are often allocated via complicated decision rules that blend lottery elements with defined priorities.[225] These systems presumably were not designed to enable research on the causal effect of the offer of admission to a prized school, but researchers have discovered that they provide a rare opportunity to generate gold-standard estimates of this treatment's effect.[226]

The causal inferences one can make from school-lottery data depend, however, on whether the school district runs a single, unified lottery or separate lotteries for traditional public schools and each charter school. The "treatment effect" estimate for a given school is an average with respect to those students whose probability of being assigned to the school was between zero and one. Unified lotteries, which are less cumbersome for parents, result in a much larger share of the student population—and especially students from disadvantaged families—having positive probabilities of being assigned to a charter school *and* to a traditional public school.[227] This enables researchers to estimate the treatment effect of charter schools on a much wider swath of the student population,

[223] Computer scientists are developing data-encryption methods that allow personal identifiers to be used in matching but never revealed to the researcher. *See generally* C. Quantin et al., *Automatic Record Hash Coding and Linkage for Epidemiological Follow-Up Data Confidentiality*, 37 METHODS INFO. MED. 271 (1998); C. Quantin et al., *How to Ensure Data Security of an Epidemiological Follow Up: Quality Assessment of an Anonymous Record Linkage Procedure*, 49 INT. J. MED. INFO. 117 (1998). Statisticians are working on privacy protection techniques that combine encryption with the temporary supplementation of real data with fake data, so that a leak of, say, names in a registry of felons, would not allow anyone to confidently infer that the listed people actually are felons. *See generally* H.C. Kum et al., *Privacy Preserving Interactive Record Linkage (PPIRL)*, 21 J. AM. MED. INFO. ASSOC. 212 (2014). States can also minimize risks by disclosing to researchers certain publicly available information to be used for record linkage (e.g., name, sex, date of birth) decoupled from sensitive outcome information (test scores, income, incarceration, etc.). This both protects privacy and—very importantly—prevents the researcher from playing around with different record-linkage algorithms until she jury-rigs a substantive treatment-effect "finding" she had hoped to discover. *See supra* notes 170–78 and accompanying text. Similarly, the Census Bureau is developing protocols for releasing perturbed (statistically blurred) sensitive data to researchers, who use the perturbed data to write code for analyses that the Census Bureau then runs on the researcher's behalf using the real data. *See* Susan Dynarski, Building Better Longitudinal Surveys (on the cheap) Through Links to Administrative Data 12 (Dec. 2014) (unpublished paper prepared for Nat'l Academy of Educ.) (available at https://naeducation.org/wp-content/uploads/2016/10/dynarski-nces-longitudinal-surveys.pdf) (discussing these protocols and extensions for use in education research).

[224] *See generally We Can Help You Build a Unified Enrollment System in Your City*, INNOVATION IN PUBLIC SCHOOL CHOICE, www.iipsc.org (last visited Jan. 16, 2018).

[225] For example, siblings or nearby residents may be given priority over other applicants. *Interdistrict Public School Choice*, STATE OF NEW JERSEY DEP'T OF EDUC., http://www.state.nj.us/education/choice/ (last visited Jan. 16, 2018).

[226] *See, e.g.*, A. Abdulkadiroglu et al., *Research Design Meets Market Design: Using Centralized Assignment for Impact Evaluation* (Nat'l Bureau of Educ. Research, Working Paper No. 21705, 2015), (including sources cited therein).

[227] *See* Marcus A. Winters, *Narrowing the Charter School Enrollment Gap* (Civic Rep. No. 106, Manhattan Institute, Dec. 2015), https://www.manhattan-institute.org/sites/default/files/R-MW-1215.pdf (finding

not just the children of exceptionally motivated parents, without the strong assumptions of matching or regression-based research designs.[228]

But when the school superintendent in Oakland, California recently proposed a unified lottery, the teachers' union and allied school board members fought back.[229] Presumably they were worried that the simplified lottery would make charter schools more attractive, drawing students (and eventually jobs) away from the unionized public-education sector. The state may well have a legitimate interest in capping the number of charter schools within a district,[230] but once the state has decided to allocate spaces by lottery, it has no legitimate interest in making lottery participation cumbersome. When a district can adopt a unified lottery at no cost, as Oakland was positioned to do,[231] courts should not hesitate to require it.

C. *Randomized Remedies*

Education-rights plaintiffs typically seek remedies that only the state can provide. Liberal plaintiffs want billions more in spending, class-size limits, expanded preschool programs, and sometimes socioeconomic integration. Conservatives want to reform teacher pay and tenure and to enable poor children to attend charter or private schools. Because such policy interventions are controlled by the state, the standard defense-side argument that "the evidence is too shaky" to warrant judicial intervention elides the central issue. The evidence is shaky in large measure *because* the state has not elected to randomize the allocation of funding, or tenure laws, or charter-school policies across schools or school districts. In the absence of a set of credible experiments—conducted, of necessity, in cooperation with the state—the evidence on each side will consist of selective gleanings from a body of literature that, owing to publication incentives and researcher discretion, contains a *predictably inconsistent set of findings*.[232]

School-finance plaintiffs often have tried to sidestep the morass of conflicting empirical results by asking courts to, in effect, privilege the craft

that Denver's adoption of common enrollment was followed by a "substantial[] increase[in] the proportion of students enrolling in charter kindergartens who are minority, eligible for free/reduced-priced lunch, or speak English as a second language").

[228] *See* Abdulkadiroğlu et al., *supra* note 226 (estimating effects of charter schools in Denver). The reason is that estimates from lottery studies are only informative about students with propensity scores (*ex ante* probability of admission to the treatment school) bounded between zero and one. If only a small, self-selected group of students bothers to apply to a charter school, researchers can learn whether the charter school makes a difference for these students, but not whether that type of school would help the larger population of students who do not bother to apply.

[229] Jill Tucker, *Charter School Advocates Push Enrollment Shift in Oakland*, S.F. CHRON. (Dec. 1, 2015), http://www.sfgate.com/bayarea/article/Charter-school-advocates-push-enrollment-shift-in-6668509.php.

[230] *Cf.* Kate Zernike, *A Sea of Charter Schools in Detroit Leaves Students Adrift*, N.Y. TIMES (June 28, 2016), https://www.nytimes.com/2016/06/29/us/for-detroits-children-more-school-choice-but-not-better-schools.html?_r=0.

[231] A charitable foundation had offered to pay for the cost of the new lottery system. Tucker, *supra* note 229.

[232] *See supra* Section III.C.

knowledge of professional educators over the technical knowledge of economists and statisticians.[233] The plaintiffs present teachers, principals, and school district administrators who testify about all the wonderful things they could and would do for disadvantaged children, if only they had more money.[234] This testimony is often complemented by a "professional judgment" costing-out study.[235] In this kind of study, a panel of educators from around the state is asked what features a model school ought to have, and the cost of establishing such model schools is then estimated, taking account of regional variation in the price of inputs.[236] Plaintiffs then argue that the state must provide local school districts with enough money to convert extant schools into model schools.

Defendants respond, correctly, that it is entirely speculative whether such an increase in funding would actually be used to convert extant schools into model schools.[237] Defendants also point out that the educators' putative craft knowledge may consist of little more than hunches or, worse, self-interested rationalizations.[238] (In contrast to claims based on well-designed studies conducted pursuant to pre-analysis plans, there is no way for third parties to verify whether the craft-knowledge beliefs are correct—short of testing them with a well-designed empirical study.)

Yet, plaintiffs are on equally solid ground criticizing the defense experts' statistical "demonstrations" of a weak or non-existent correlation between spending and outcomes. First, the observed range of spending variation across poor schools or school districts may be small compared to the size of the remedy proposed by plaintiffs. The effect of modestly increasing or decreasing funding to a school may tell little about the effect of a big increase, particularly if the programs and school features that administrators would buy with a big cash infusion have complementary, mutually reinforcing effects. Second, and crucially, the observed variation in spending may be correlated with potential outcomes, obscuring the causal effect of spending on outcomes.[239]

[233] *See, e.g.*, Montoy v. State, No. 99-C-1738, 2003 WL 22902963, at *47–48 (Kan. Dist. Ct. Dec. 2, 2003) ("Defendants attempted to discount any connection [between spending and student outcomes] with expert witnesses some of whom hinted that "money didn't matter" in student performance. Controverting these Ivory Tower views . . . were the impressive and credible experiences of many Kansas educators . . . . One by one, these unsung heros [sic] in the daily battle against ignorance looked the Court straight in the eye and said we know how to do it, we simply lack the resources to do what we know how to do. . . . In a word, this Court believed them.").

[234] *Id.*

[235] For overviews of the accepted costing-out methods, see Thomas A. Downes & Leanna Stiefel, *Measuring Equity and Adequacy in School Finance*, *in* HANDBOOK OF RESEARCH IN EDUCATION FINANCE AND POLICY 244, 248–53 (Helen F. Ladd & Edward B. Fiske eds., 2d ed. 2008). HANUSHEK & LINDSETH, *supra* note 4, at 178, (stating that the "professional judgment approach" is the most widely used costing-out technique).

[236] *See supra* note 227 and accompanying text.

[237] *See* HANUSHEK & LINDSETH, *supra* note 4, 178, 187 (observing that in Wyoming, where courts and the legislature relied on a model-schools costing out study, the new funding was not used to create the model schools). Lindseth is a leading defense-side lawyer and Hanushek is the leading defense-side expert witness. In contrast to claims based on well-designed studies conducted pursuant to pre-analysis plans, there is no way for third parties to verify whether the educators' beliefs are correct—short of testing them with well-designed empirical studies. *Id.*

[238] *See id.* at 180–84.

[239] Some evidence for this conjecture comes from recent studies that use judicial decisions as an instrument for plausibly exogenous spending changes. *See* Lafortune et al., *supra* note 45.

An entirely analogous set of issues arise when conservative plaintiffs push for tenure and teacher-compensation reforms and defendants complain that plaintiffs have not *proven* that relaxing tenure or seniority protections, or tying salaries to performance, would improve the outcomes of disadvantaged students. Of course they have not. Plaintiffs have litigated teacher tenure in states like California that require *all* school districts to follow the same tenure and seniority rules.[240] There is no between-district variation in the practices at issue, and without variation, statisticians cannot estimate the effect of tenure reform on student outcomes. One might try to estimate the effect of tenure reform by making comparisons over time between states that alter their tenure laws and states whose tenure laws remained unchanged (a so-called "difference in difference" design), but this approach is only viable if some states have actually adopted the reform sought by plaintiffs, and it rests on a *big*, uncheckable, and not-very-plausible assumption.[241] Specifically, the causal estimate is unbiased only if time trends in student performance would have been the same in the states that did and did not reform tenure (but for the change in the tenure rules).[242] This assumption is implausible because changes in tenure rules are probably correlated with changes in the political strength of the state's teacher unions, and union strength may have good or bad consequences for student performance through many mechanisms other than tenure laws.[243]

So, what should a court do when plaintiffs argue that the state must spend a lot more on disadvantaged students, or must liberalize tenure and seniority protections? Under an arbitrariness standard of review, it is doubtful that the court should set aside the legislature's judgment when the educational payoff is speculative. But in some cases, circumstantial danger signs may indicate a potentially significant breach of the state's duty of care.[244] For example, the state may be a negative outlier per Chetty and Hendren's estimates of socioeconomic mobility,[245] while also providing much less educational funding or establishing much stricter tenure rules than other states. And perhaps there is an overwhelming consensus among people who may have pertinent craft knowledge, such as public-school principals, that the funding limitations or labor laws have seriously stymied their efforts to educate disadvantaged students. If so, a more activist stance might be in order, considering the fundamental status of education

[240] Vergara v. State, No. 484642 (Cal. Sup. Ct., Cnty. of Los Angeles, Aug. 27, 2014), *rev'd* 246 Cal. App. 4th 619 (2016).

[241] For good discussions of these issues, emphasizing the sensitivity of results to modeling assumptions, see ANGRIST & PISCHKE, *supra* note 159, at 161–66, 227–46; MORGAN & WINSHIP, *supra* note 100, at 251–75.

[242] *See* sources cited *supra* note 241.

[243] Another way researchers might try to study the effects of tenure is by comparing outcomes of students at charter schools (which are exempt from tenure laws that apply to regular public schools) and regular public schools. But these schools differ from one another in lots of other ways too.

[244] In the domain of election law, where government regulation is similarly pervasive and constitutional values are also at stake, the U.S. Supreme Court has implicitly (and sometimes explicitly) used "danger signs" as a trigger for heightened scrutiny. *See generally* Christopher S. Elmendorf, *Structuring Judicial Review of Electoral Mechanics: Explanations and Opportunities*, 156 U. PA. L. REV. 313 (2007).

[245] To get at this question, one could construct weighted averages of the effects (per Chetty & Hendren, *supra* note 23) of each state's "commuting zones" on socioeconomic mobility, with weights corresponding to the fraction of the state's low-income population in each commuting zone.

and the representation-reinforcing arguments for judicial review. But what is the activist court to do?

It should *consider* issuing what we call a "temporary experiment remedy"—a court order requiring the state to implement the plaintiff-sought remedy in a randomly selected subset of schools or school districts, for a fixed period of time.[246] The temporary experiment remedy has a number of attractive properties.

Initially, it is more respectful of the legislature's policy-making and fiscal prerogatives than a universal remedy. Rather than permanently displacing legislative judgments, the remedy displaces the legislature's judgment only as a to some localities, and only for a period of years. Second, the court will be able to learn which remedies make a difference, terminating those that prove ineffectual. Freed-up resources can then be reallocated to help disadvantaged students in other ways. Third, in hashing out the design of the temporary experiment, the parties and their expert witnesses will take *ex ante* positions on the quality of the remedy's design for causal inference. The methods are committed to in advance, before any interested party sees the outcome data. And because of the court's involvement, the eventual findings will be made public rather than buried in a file drawer. This largely resolves the third-link issues in the useful-causal-research chain—that is, lack of credibility owing to publication incentives and results-oriented data analysis.

This is not to say that temporary experiment remedies are appropriate for all cases in which the plaintiffs have a strong "danger signs" argument but cannot provide credible estimates of causal effects owing to the state's control over the assignment of educational treatments. In some cases, there may be reasonable ethical objections.[247] In other cases, the court might conclude that the plaintiffs' theory cannot be tested satisfactorily with a randomized remedy, perhaps because of a risk of strategic behavior by actors with a stake in the experiment's results, because the theory underwriting the reform is a theory about a new equilibrium that would be reached only after a long period of adjustments, or because of other reasons.[248] Then again, even if some of these factors could bias the findings, a court might conclude that the experiment would be *informative enough* that the state must give it a try, given the plaintiffs' argument against doing nothing.

[246] Alternatively, the court might issue a legislative remand, while signaling that one way for the legislature to come into compliance is through the enactment of temporary-experiment remedies.

[247] In our view, the ethical objections to "experimenting on students" are generally weak, except in cases where the treatment is known or discovered to have a substantial positive effect (much as medical ethics require drug trials to be canceled if it becomes clear that the treatment works much better than the placebo).

[248] For a succinct overview of the limits of experiments in education research, see Schanzenbach, *supra* note 159, at 220. One possible objection that we *do not* think courts should credit is that education is just too "complex" or "nuanced" for researchers to learn much from quantitative studies. (For suggestions to this effect, see Superfine, *supra* note 70, at 689–92 and Black, *supra* note 58, at 1757–64.) Medicine is similarly complicated. Drugs interact with one another and with idiosyncratic patient characteristics in complicated ways, but learning about *average* treatment effects has nonetheless been hugely important for improving medicine. *See* FREDERICK M. HESS & MICHAEL J. PETRILLI, NO CHILD LEFT BEHIND PRIMER 94 (2006). Also, the asserted "complexity" of education is to a large extent an empirical proposition, which can be tackled using experiments (e.g., subgroup analysis) and which can inform the design of interventions (e.g., governance or incentive structures designed to harness teachers' and principals' craft knowledge of their students' particular needs).

To illustrate, consider *Vergara v. California*, a recent case about teacher tenure and seniority protections in California.[249] California's short tenure clock and stringent seniority protections make it a national outlier, and surveys of California principals show that they regard the state's tenure rules as a major impediment to improving the education of disadvantaged students.[250]

Yet, treatment-effect estimates from loosening tenure and seniority protections in a random subset of school districts could be quite misleading. For example, teachers' unions in the control districts might become temporarily more cooperative with respect to the dismissal of poor teachers, in the hopes of causing the treatment to have no measured effect. Principals might also change their behavior during the period of the remedy because they want to affect the findings. Perhaps they would work extra hard to fire bad teachers in the treatment districts and slack off in the control districts. Principals in the treatment districts may also find it more difficult to hire and retain teachers than they would under a universal remedy, as teachers may prefer jobs in the control districts (strong tenure protections) to jobs in the treatment districts (weaker protections). Teachers may sort by quality, with stronger teachers electing to stay in the treatment districts and weaker teachers moving to the control districts. Parents may relocate their families from control to treatment districts if they perceive the treatment districts as providing a better education, and this sorting of families among districts is likely to be correlated with potential outcomes.

Still, a court *might* conclude that the randomized remedy would be sufficiently instructive that the state must adopt it. For example, if the job choices of new teachers during the period of the remedy did not reveal a strong preference for control districts over treatment districts, this would substantially undermine the state's argument that loosening tenure protections would diminish the supply of quality teachers. If the number of treatment and control districts were large, the court might conclude that no individual principal or union would have much incentive to behave strategically in the hopes of affecting the court's eventual findings about the temporary remedy. And, with respect to student migration, the court might side with researchers who argue that the average effect of the treatment on children living in each district at the time the court announces the remedy (prior to any treatment-induced migration) would be a lower-bound estimate of the average effect of the treatment on the full population.[251]

[249] Vergara v. State, 209 Cal. Rptr. 3d 532 (Cal. Ct. App. 2016).

[250] Researchers at Stanford and UC Berkeley surveyed a stratified random sample of several hundred California public school principals about barriers to achieving the state's academic standards. *See* BRUCE FULLER ET AL., CALIFORNIA PRINCIPALS' RESOURCES: ACQUISITION, DEPLOYMENT AND BARRIERS (2006). Principals were asked about the need for "change and improvement" in ten areas, such as adding more teachers (positions) to the school, budgetary flexibility, funding for professional development, technical assistance with student data, and discretion to dismiss ineffective teachers. *Id.* at 42. Respondents ranked "more freedom to dismiss ineffective teachers" as the area where change was most urgently needed. *Id.* at 43, fig.11. Principals were also asked whether they could increase the amount of instructional time spent on reading or lengthen the school day, reforms that have often been promoted as a means of bringing disadvantaged students up to grade level. Labor rules and constraints on the expenditure of categorical funds were reported to hinder such reforms. *Id.* at 43–44. Another survey revealed that California principals are more likely than principals in most other states to report barriers to the dismissal of underperforming teachers. *Id.* at 24–28.

[251] For similar arguments in an instrumental-variable study of the effects of school funding, see Lafortune et al., *supra* note 45.

The case for a randomized remedy is certainly much stronger, however, where the political actors with a big stake in the remedy's success or failure are not positioned to intervene strategically, and covertly, in ways that could affect the outcome of the experiment. To illustrate, imagine a more conventional school-finance case in which plaintiffs seek supplemental funding for certain high-poverty schools or districts. The state distributes funds to schools using a weighted student formula, which provides 10% more funding for low-income than middle-class students. The plaintiffs want the average low-income "weight" to be increased from 1.1 to 2, and they want the weights to vary with family income, rather than being fixed at the same level for all students who qualify for free or reduced-price lunches.[252] The plaintiffs request a universal remedy, but as a fallback, they argue in the alternative for a ten-year temporary experiment remedy, randomized at the level of the school district. Statewide taxpayer groups oppose these interventions. To successfully throw the experiment, however, the taxpayer interests would have to find some way to make additional spending by the treatment schools or districts temporarily unproductive, or to temporarily improve the control schools, and the taxpayer groups would have to do all this covertly, such that the court and other public actors *believe* the experiment's null finding.[253] It seems unlikely that the taxpayer groups could pull this off. Also, anticipating possible shenanigans, the court could build prophylactic safeguards into the remedy, such as requiring judicial approval of any new state-imposed restrictions on school administrators' use of the supplemental funds.[254]

To be clear, it is absolutely not our position that courts *should* enter temporary experiment remedies in challenges to teacher tenure and seniority protections, or in cases where plaintiffs argue that the state must dramatically increase spending on the education of the poor. Our point is simply that these remedies should be considered, and right now, they are entirely missing from the discussion. We also think that courts could induce a salutary change in litigation strategy by signaling their willingness to consider temporary-experiment remedies. Plaintiffs hoping to solve "problems no one has solved" will begin to ask themselves not simply, "Which educational reforms would I most like a court to order?" but also, "Which reforms would be testable, with the cooperation of the state, using well-designed experiments?" A shift in focus may ensue, with litigants shying away from large-scale reforms that may have complicated spillover effects in favor of more modest and harder-to-game interventions, such as: extension of the school day, targeted tutoring or dropout prevention programs,

[252] *Cf.* BRUCE BAKER ET AL., IS SCHOOL FUNDING FAIR? A NATIONAL REPORT CARD 4–6 (2012) (comparing weights used in different funding formulas and noting lack of empirical justification for the weights).

[253] If the taxpayer group's intervention was not covert, the judge would not be deceived by the remedy's apparent lack of impact, and the judge might well respond with a permanent remedy that the taxpayer group strongly disfavors.

[254] To be sure, strategic intervention by taxpayer groups is not the only possible cause of misleading results. There may be spillover effects, for example, if the additional funding allows treatment districts to bid away top teachers from the control districts. Or principals in the control districts, who want the experiment to succeed so that they too can get the extra funds, may throw their students under the bus (so to speak) in the hopes of the court finding a big "treatment effect" from the additional funds. Before entering the remedy, the court would have to consider these risks and whether they can be mitigated.

early childhood interventions in low-income neighborhoods, and possibly socioeconomic integration programs.[255]

* * *

Although, in the short-term, students in the control groups would not benefit from judicial interventions, the emergence of the temporary experiment remedy should greatly benefit disadvantaged children in the longer term. Courts will be able to see whether their remedies make a difference and abandon those that prove ineffectual. The emergence of the temporary-experiment remedy as a familiar, practicable response to education-quality claims may encourage courts that are wavering about whether to find for the plaintiffs in a given case to provide relief, because the relief need not be as drastic as a permanent injunction or a declaration that the school system is unconstitutional from top to bottom.[256]

The evidence developed through temporary-experiment remedies in jurisdictions whose courts enforce the education right relatively aggressively should prove enormously helpful to plaintiffs in other jurisdictions whose courts have been more skeptical. As we noted in Part II, a number of recent education-quality claims have been rejected on causation grounds.[257] Insofar as temporary-experiment versions of a plaintiff-sought remedy have proven effective in other states, the causation hurdles erected by the skeptical courts should be easy to overcome.

Beyond causation, the evidence developed through temporary-experiment remedies in states with good, linkable administrative databases should enable plaintiffs to make much more convincing arguments that certain policies are *substantively* arbitrary. For example, if it could be shown that the cost of intensive early childhood interventions for disadvantaged children is substantially less than the present value of future public savings on welfare and criminal justice programs, then it would probably be arbitrary for the state not to provide those programs.[258] Or if (as a recent study with a good design suggests) academic tracking substantially improves the outcomes of low-income African American students who are placed on the high track without damage to students on the lower tracks,[259] then it might be arbitrary for states not to provide an academic-

[255] Of these examples, socioeconomic integration is perhaps the most challenging because middle class families have historically opposed efforts to diversify middle-class schools, and they might try to strategically undermine an integration-oriented temporary experiment remedy. If the remedy were to be kept in place for a fairly long period (say, ten–fifteen years), however, it seems unlikely that "throwing the experiment" (by willfully trying not to educate or to demoralize the randomly assigned low-income students) would be an attractive option for the middle class families. Low performance by the low-income students would bring down test scores and other measures of the "quality" of the middle-class school, with adverse consequences for the middle-class families, particularly if they own a home nearby.

[256] *Cf.* Daryl J. Levinson, *Rights Essentialism and Remedial Equilibration*, 99 COLUM. L. REV. 857, 922 (1999) (arguing that judicial perceptions about the feasibility, costs, and benefits of potential remedies pervasively shape the courts' definition of constitutional rights).

[257] *See supra* notes 56–57 and accompanying text.

[258] There is some evidence to this effect, but it comes from small-scale demonstration projects. *See* Sneha Elango et al., *Early Childhood Education* 40–43 (Nat'l Bureau of Econ. Research, Working Paper No. 21766, 2015) (noting that only the Perry and Abecedarian studies tracked enough outcomes, for a sufficient period of time, to enable a cost-benefit analysis that accounts for most anticipated benefits). It is conceivable that a state could defeat the claim of "substantive arbitrariness" by showing a tradeoff between the pursuit of measurable and unmeasurable education objectives, but this question will only become relevant once good data about causal effects on the measurable outcomes becomes available. (Thanks to David Schleicher for raising the question.)

[259] *See* David Card & Laura Giuliano, *Can Tracking Raise the Test Scores of High-Ability Minority Students?* (Nat'l Bureau of Econ. Research, Working Paper No. 22104, 2016).

tracking option in schools that serve high concentrations of poor African American students. We are wary of courts controlling the size and allocation of education budgets, but to the extent that courts intervene on a more-than-temporary basis, courts should be doing so on the basis of credible estimates of the effects of mandated interventions on the adult outcomes that ground the education right.[260]

V. CONCLUSION

In 1973, the U.S. Supreme Court in *Rodriguez* stressed epistemic uncertainties in holding that Texas's system of public education passed muster under the federal constitution.[261] In 2016, following decades of state constitutional litigation, the Texas Supreme Court in *Morath* similarly threw up its hands, declaring that research on the "intractable" question of whether additional school funding would improve student outcomes had reached an "impasse," and refusing to take sides in the litigants' dispute over whether to increase school funding, relax teacher-tenure and seniority protections, or remove charter-school caps.[262] It might seem that little has changed.

But the *Morath* Court overlooked something important: education research is undergoing its own "scientific revolution."[263] Informed by the logic of randomized controlled trials, this revolution tracks broader developments in the social sciences and has been aided and abetted by the federal government. Owing to this revolution, statisticians and social scientists today have a vastly better understanding (compared to the early 1970s) of *barriers* to credibly answering questions about whether one or another educational intervention is likely to substantially improve the outcomes of disadvantaged children. We have argued that many of these barriers can be overcome—with the cooperation of the states. Insofar as state constitutions impose upon state actors a justiciable duty of care with respect to education, courts ought to scrutinize state failures to enable the production of knowledge about the impacts of education policies and programs on the adult outcomes of disadvantaged children. In time, the resulting body of research may lead even cautious courts and policymakers to reconsider their epistemic pessimism and make real the constitutional promise of a quality education for all.

[260] Temporary-experiment interventions may also enrich the quality of public discourse about what works in education, which is itself a key goal of democratic experimentalism. *See* Dorf & Sabel, *supra* note 211, at 473.

[261] San Antonio Indep. Sch. Dist. v. Rodriguez, 411 U.S. 1, 33–34 (1973).

[262] Morath v. Tex. Taxpayer & Student Fairness Coal., 490 S.W.3d 826, 853 (Tex. 2016).

[263] *See* Robert E. Slavin, *2002 Dewitt Wallace-Reader's Digest Distinguished Lecture: Evidence-Based Education Policies: Transforming Educational Practice and Research*, 31 EDUC. RES. 15, 16–17 (2002).