
REASONABLE EXPECTATIONS: A REPLY TO ELMENDORF AND SHANSKE 2018

Ethan Hutt
*Morgan Polikoff**

INTRODUCTION

With *Solving ‘Problems No One Has Solved,’*¹ Elmendorf and Shanske have offered an important argument that we hope will provide a framework for new conversations among and between public-interest litigators, educational researchers, and those across the ideological spectrum interested in improving public education for all children. The core of their contribution is rooted in three interrelated insights. First, the issue of causality has been a major stumbling block in both education litigation and research for several decades. Second, recent advances in data collection, computing technology, and research methodology—the “causal revolution”—hold the potential to help (re)move this historic obstacle.² And third, and perhaps most importantly, judges should, in certain cases, facilitate causal research by requiring states, at the very least, to produce the prerequisite data and possibly order the work themselves.

As educational researchers, we agree wholeheartedly with the thrust of this argument: good data on public education in America is too hard to come by, and access to more and better data can help us answer questions that are central to improving education and securing the educational rights of all children. That said, as education researchers, we also recognize the very real limits on the kind of answers that school data—and educational research—can provide.

* Ethan Hutt is Assistant Professor of Education, University Maryland, College Park; Morgan Polikoff is Associate Professor of Education, University of Southern California Rossier School of Education.

¹ Christopher S. Elmendorf & Darien Shanske, *Solving “Problems No One Has Solved”: Courts, Causal Inference, and the Right to Education*, 2018 U. ILL. L. REV. 693.

² For a consideration of how the causal revolution might affect a different aspect of education law, see Ethan Hutt & Aaron Tang, *The New Education Malpractice Litigation*, 99 VA. L. REV. 419 (2013).

In this reply, we elaborate on this view and try to articulate what we see as the key contributions and remaining challenges of Elmendorf and Shanske's proposal. Part I highlights the historical importance and intractability of the problem Elmendorf and Shanske hope to solve. Part II considers the current state of the "causal revolution" in education research in order to highlight the kind of progress that can be made and the issues likely to remain out of reach despite the considerable data, methodological, and computational advances of the last decade. Finally, moving beyond an abstract consideration of the potential of better data, Part III considers Elmendorf and Shanske's proposal in the context of a case—*Williams v California*³—that likewise sought to secure educational rights by requiring the state to produce data on the availability of school resources.

PART I

In his classic 1982 book *Legislated Learning*, Arthur Wise, an early school finance litigation advocate, stressed the importance of distinguishing between the political and technical obstacles to securing educational opportunity. The former, Wise argued, were the kind of problems that might require judicial intervention, while the latter necessitated judicial restraint.⁴ One intriguing part of Elmendorf and Shanske's argument is the extent to which they highlight how deeply intertwined these problems might be in practice: solving the 'technical challenge' of providing high quality schooling for all might stem from the 'political challenge' of securing production and access to high quality information. Elmendorf and Shanske are right to highlight many of the ways in which recent school litigation has been hindered by a lack of access to information—potentially in the state's possession, but certainly within the state's power to collect—that might shed light on the extent of the problem and point toward possible solutions. This challenge, however, has a much longer and more complicated history than even the authors suggest. Securing information about the state of America's schools has been an enduring challenge for lawmakers and would-be reformers for nearly two centuries. It is, indeed, a problem no one has solved.

In this section, we offer a look at some of the enduring challenges in data collection and school statistics. This brief examination is intended to highlight two distinct features of this history. First, the ways in which the politics surrounding data collection of schools has long been shaped by the distributed politics of American school governance—often in a way intended to maintain use-

3. *Williams v. California*, No. 312236 (Cal. Sup. Ct. Aug. 14, 2000).

4. ARTHUR WISE, *LEGISLATED LEARNING: THE BUREAUCRATIZATION OF THE AMERICAN CLASSROOM* (1982).

ful ambiguity. The resulting ‘thin description’—to use Ted Porter’s phrase⁵—provided by school statistics has long projected a sense of rationalization and administrative order while only partially constraining local prerogatives. The second, and related, issue is that, even when improving systematic data collection has been a priority, the usefulness of the data has been limited by the lack of uniformity in the underlying system. That is, while our capacity to produce uniform *statistical* data has improved over time, this should not be confused with the existence of uniform *operational* capacities, environments, or experiences. Despite considerable advances in state and federal administrative capacity over the last century, in American schools, operational localism still reigns supreme.⁶ Placing the problem of data collection in this longer historical perspective points not only to the potential necessity of judicial intervention—and the importance of Elmendorf and Shanske’s contribution—but also to the need to set reasonable expectations about what can be accomplished in doing so.

Though America has a long tradition of local control in schools, state mandates to collect and report school statistics are nearly just as old. In the mid-19th century, local access to state school funds was usually contingent on providing the state board of education with an enumeration of either the total number of school-aged pupils in town or a report of the average daily attendance of the local schools.⁷ Though by the 1880s every state legally required districts to produce an annual statistical report, states differed considerably in what information had to be reported.⁸ While New Jersey, for instance, gave superintendents almost complete discretion in deciding which statistics were relevant to report to the state, Illinois law required the reporting of no fewer than twenty distinct pieces of information ranging from the number of people under the age of twenty-one in each county to the percentage of students taught by male and female teachers at the same time.⁹ Even when states collected and reported the same information, states often differed—and even cities within the same state sometimes differed—on how specific statistics were calculated, making interstate comparisons useless, if not impossible. To take a notorious example, in 1881, there were seventeen different definitions of “legal school age” in use across the states and territories of the United States.¹⁰ Likewise,

5. Theodore M. Porter, *Thin Description: Surface and Depth in Science and Science Studies*, 27 *OSIRIS* 209–226 (2012).

6. DOUGLAS S. REED, *BUILDING THE FEDERAL SCHOOLHOUSE: LOCALISM AND THE AMERICAN EDUCATION STATE* (2014).

7. See, e.g., LEONARD PORTER AYRES, *CHILD ACCOUNTING IN THE PUBLIC SCHOOLS* 22–28 (1915); PAUL HENRY NEYSTROM, *THE SCHOOL CENSUS* 6–47 (1910) (discussing early state practices in child census taking).

8. John Philbrick, *State Reports on Education*, in *REPORT OF THE COMMISSIONER OF EDUCATION, 1884–1885*, XVI (1886).

9. *Id.* at XVII.

10. DEPARTMENT OF THE INTERIOR, BUREAU OF EDUCATION, U.S. COMMISSIONER’S ANNUAL REPORT 1881, 320–321 (1883). This problem, combined with the fact that state definitions of ‘legal school age’—usually defined as the age of compulsory attendance—were not co-extensive with the students eligible to attend school, could result in a state reporting a total school enrollment larger than the total school-aged population.

when it came to calculating “average daily attendance”—a foundational measure of school system quality and efficiency throughout the Progressive Era¹¹—districts and sometimes even principals adopted widely varying calculation methods, making it difficult to use these numbers to compare school systems.¹²

Crucially, these varying practices persisted despite continued attention to their inadequacy. The issue of creating uniform definitions for key school statistics was the subject of repeated pleas from the Bureau of Education and was raised repeatedly at the annual meetings of the National Education Association (NEA), leading to the creation of a National Committee on Uniform Records.¹³ While some of this variation was due to differing levels of administrative capacity¹⁴ and some the result of perceived tensions between superintendents’ professional discretion and the use of standardized measures,¹⁵ it was clear to contemporaries that the lack of statistical comparability also served to make critical evaluation more difficult. As two prominent professors of school administration and leaders in the school survey movement complained in 1921, “astonishing variation exists among school systems with respect to the nature and amount of the data which are being gathered . . . that the success or failure of accepted financial and educational policies or programs may be determined.”¹⁶

For a brief discussion of the history of attendance records, see Ethan L. Hutt, *Measuring Missed School: The Historical Precedents for the Measurement and Use of Attendance Records to Evaluation Schools*, J. EDUC. STUDENTS PLACED AT RISK (forthcoming).

11. ARTHUR B. MOEHLMAN, CHILD ACCOUNTING: A DISCUSSION OF THE GENERAL PRINCIPLES UNDERLYING EDUCATIONAL CHILD ACCOUNTING, TOGETHER WITH THE DEVELOPMENT OF A UNIFORM PROCEDURE (1924).

12. The crux of the issue was how long to leave students on the attendance rolls who were no longer attending school and, even in the case of confirmed transfers, how those students’ absences prior to the receipt of transfer notice should be counted. *Id.* at 21.

13. According to one contemporary account, the issue was the subject of discussion at the annual meeting of the NEA in [CC: need year of this meeting], and was raised at the annual meetings of the National Education Association in 1859, 1860, 1872, 1874, 1887, 1881, 1885, 1886, 1887, 1890, 1891, 1892, and 1895. Halle D. Woods, *School Policy via School Facts*, 13 SCH. REV. 544 (1905). The Committee on Uniform Records produced its final report in 1912 but compliance with its recommendations was mixed. See ARCH OLIVER HECK, A STUDY OF CHILD-ACCOUNTING RECORDS (1925).

It was not until 1953 that the federal government would publish The Common Core of State Educational information—a forerunner to the modern Common Core of Data (CCD). PAUL L. REASON ET AL., THE COMMON CORE OF STATE EDUCATIONAL INFORMATION (1953).

14. The median number of state school department of education personnel at the turn of the century was 2. DAVID TYACK, THOMAS JAMES & AARON BENAVIDE, LAW AND THE SHAPING OF PUBLIC EDUCATION, 1785–1954, 61–62 (1991).

15. NATIONAL EDUCATION ASSOCIATION, COMMITTEE ON UNIFORM RECORDS AND REPORTS, FINAL REPORT OF THE COMMITTEE ON UNIFORM RECORDS AND REPORTS TO THE NATIONAL COUNCIL AT THE ST. LOUIS MEETING, 42 (1912) (quoting State Superintendent of Public Schools in Maine, Payson Smith: “After we have agreed upon the fundamental points of school reporting and accounting . . . the manner of their presentation and interpretation to the public will constitute a constant challenge to the skill and ingenuity of the superintendent himself”).

16. GEORGE D. STRAYER & NICKOLAUS L. ENGELHARDT, A SCORE CARD AND STANDARDS FOR THE RECORDS AND REPORTS OF CITY SCHOOL SYSTEMS 1 (1923).

This is not to say that school statistics did not improve and become more sophisticated during the Progressive Era—they surely did.¹⁷ But political sensitivities and a desire to produce school statistics, while simultaneously avoiding undesirable public scrutiny, continued to characterize many prominent data collection efforts throughout the 20th century. After World War II, when the federal government began to take an increasingly active interest in educational outcomes, efforts to understand the state of education in the U.S. were repeatedly stymied.¹⁸ For instance, it took an act of Congress to commission the *Equality of Educational Opportunity* study (The Coleman Report).¹⁹ Even then, participation was voluntary—leading many school districts not to participate. This occurred despite promises that information collected would not be disaggregated beyond the regional level or used to compare individual districts.²⁰ The creation and sample design of the National Assessment of Educational Progress (NAEP) in the early 1960s was, likewise, shaped in response to political outcry from state and local officials who did not want the federal government to draw student achievement comparisons between states or individual districts.²¹ The result was a student achievement indicator incapable of being reported at anything other than a national level. Given the total lack of standardization in educational opportunity or curriculum across states, the interpretative value of NAEP scores has, therefore, always been questionable.²²

Nor was this issue merely a function of the direct involvement of the federal government. One of the largest ever research efforts to study the long-term effects of schooling on life outcomes—Project Talent (1960–1974)—had to forgo the collection of student race data to assuage concerned district officials and secure their participation in the study.²³ Likewise, a major long-term longitudinal study of occupational mobility in Wisconsin that began in 1957 was prohibited by state law from collecting information on the respondents' race until the law was changed in the 1970s.²⁴ Wisconsin was far from unique in this respect. Many other states had “sensitive item” policies which prohibited—

17. See, e.g., DAVID TYACK, *THE ONE BEST SYSTEM: A HISTORY OF AMERICAN URBAN EDUCATION* (1974); TRACY L STEFFES, *SCHOOL, SOCIETY, AND STATE: A NEW EDUCATION TO GOVERN MODERN AMERICA, 1890-1940* (2012); RAYMOND E CALLAHAN, *EDUCATION AND THE CULT OF EFFICIENCY* (1964).

18. REED, *supra* note 6; Ethan L. Hutt, ‘Seeing Like a State’ in the Postwar Era: The Coleman Report, Longitudinal Datasets, and the Measurement of Human Capital, 57 *HIST. OF EDUC. Q.* 615 (2017).

19. JAMES S. COLEMAN ET AL., *EQUALITY OF EDUCATIONAL OPPORTUNITY* (1966).

20. *Id.*

21. Lyle V. Jones, *A History of the National Assessment of Educational Progress and Some Questions About Its Future*, 25 (7) *EDUC. RESEARCHER* 15 (1996).

22. It was not until the 1990 that Congress authorized, initially on a voluntary basis, NAEP to collect and report state level data. *Id.* at 17. On the dangers on interpreting NAEP results see Stephen Sawchuk, *When Bad Things Happen to Good NAEP Data - Education Week*, *EDUC. WK.* (Jul. 24, 2013), <http://www.edweek.org/ew/articles/2013/07/24/37naep.h32.html>.

23. DAVID E. KAPEL, *EFFECTS OF NEGRO DENSITY ON STUDENT VARIABLES AND THE POST-HIGH SCHOOL ADJUSTMENT OF MALE NEGROES* iii (1968).

24. William H. Sewell et al., *As We Age: A Review of the Wisconsin Longitudinal Study, 1957–2001*, 20 *RES. IN SOC. STRATIFICATION AND MOBILITY* 3, 43 (2003) (explaining that initial survey design was constrained by state law prohibiting the collection of race data from respondents).

either by law or convention—surveys from asking students about information that might be deemed sensitive. This could include everything ranging from a student's race, religion, or family background to student attitudes towards adults.²⁵ While legitimate privacy concerns informed these practices, states and districts exploited these concerns to stifle efforts to research discrimination and the effects of interventions like desegregation on student attitudes and achievement.²⁶

Although, as Elmendorf and Shanske note, more recent policies like Race to the Top have made investment in data systems an explicit policy focus, we should not assume these new technical developments alone will alter the longer pattern of the selective and strategic dissemination of school statistics. Elmendorf and Shanske's insistence, then, that courts use their power to overcome potential political manipulation or objections to the release of this newly collected data constitutes a valuable intervention in the long history of American school statistics.

But while it is easy to imagine how states could collect this data with the proper investment in personnel, development of data infrastructure, and prodding from the judiciary, the history of school statistics suggests that overcoming political objections to data collection is not the only obstacle. Understanding how to capture the complexity of American education in statistical form has always posed its own set of difficulties. Indeed, while lawmakers and scholars have long sought to bring the unruly American educational expanse to statistical heel, the results have often presented more a stylized portrait of school system behavior than an accurate account of its regular state. Statistical categories can obscure as much as they reveal, especially when they seek to make broad generalizations about an inherently eclectic system. The meaning of these categories in the context of the complex operation of schools is often difficult to interpret.

We have already mentioned how decentralized curriculum decisions and concentrated political pressure resulted in a design for NAEP that reflected no particular curriculum and whose score was interpretable only relative to itself. Similarly, efforts throughout the 1960s to develop data on student achievement and access to opportunity simply overlaid uniform statistical categories on top of the messy, idiosyncratic, and non-standardized organization of American schooling.²⁷ For instance, the effort to analyze national course taking patterns—features of important policy conversations throughout the 1980s and 1990s—were an artifact of a standardized coding scheme commissioned by the

25. See, e.g., Richard Dershimer, *Commissioner on the Horns: USOE Should Modify Policy on 'Sensitive Item' Research*, 47 THE PHI DELTA KAPPAN 113 (1965).

26. See, e.g., GREG JACKSON, SUMMARY OF REVIEWERS COMMENTS [OF THE RAND CORPORATION'S DESIGN FOR A LONGITUDINAL STUDY OF SCHOOL DESEGREGATION.] 7 (1974) (indicating that expert reviewers raised concerns about state laws that would prevent researchers from collecting desired demographic information on the theory it invaded student privacy).

27. JOHN C FLANAGAN, DESIGN FOR A STUDY OF AMERICAN YOUTH (1962); Hutt, *supra* note 18.

National Center for Education Statistics, not any actual standardization in state standards, course taking patterns, or curricula.²⁸

Far from being a mere academic or technocratic matter, the contrived uniformity of collected-school statistics had real-world consequences. The push to desegregate schools following *Brown* led to an increased focus on collecting information on school demographics. But statistics indicating school racial balance often masked the resegregation within schools achieved through curricular tracking.²⁹ Subsequent statistical research identifying specific track placement and course-taking patterns that served as gatekeepers of educational opportunity led to calls for detracking and legislating specific course work. These calls were based on the theory, and available empirical evidence, that exposure to more rigorous material had positive effects on student achievement and life outcomes. In California, for instance, this led to a push for “Algebra for all” by 8th grade.³⁰ While these efforts produced greater equity at the level of course-taking statistics—enrollments tripled within a decade—the curricular change produced, on average, a negative effect on students’ tenth grade math achievement scores.³¹ This entire multi-decade episode suggests the iterative challenge of obtaining policy outcomes by legislating to statistical categories. Whether one attributes this history to local resistance to educational equity or inadequate local capacity to meet ambitious policy goals, the disappointing outcome of the “Algebra for all” push offers an important reminder that, while schools can be made to offer the same course or even the same curriculum, providing the same learning opportunities to students is something altogether different.

The point here is not that administrative data is not capable of improving student achievement or that the challenges implied by these examples are insurmountable. Rather, the point is that these challenges are real and will require constant and careful attention.

PART II

Having discussed some of the historical challenges of securing and interpreting school statistics, this section turns to consider the current state of the field. We argue that the educational research enterprise is indeed much better poised than ever before to leverage high quality educational data to evaluate programs and policies. However, we believe Elmendorf and Shanske do not give adequate attention to the legitimate difficulties in using educational data to do the kinds of analyses they suggest. We highlight some of these difficulties and offer thoughts on how to overcome them.

28. MEREDITH J. LUDWIG ET AL., A CLASSIFICATION OF SECONDARY SCHOOL COURSES (1982); CLIFFORD ADELMAN, COLLEGE COURSE MAP: TAXONOMY AND TRANSCRIPT DATA (1990).

29. See, e.g., JEANNIE OAKES, KEEPING TRACK: HOW SCHOOLS STRUCTURE INEQUALITY (2005).

30. Thurston Domina et al., *Aiming High and Falling Short: California's Eighth-Grade Algebra-for-All Effort*, 37 EDUC. EVALUATION & POL'Y ANALYSIS 275 (2015).

31. *Id.*

There is no question that the field of educational research is in the midst of a revolution in its ability to use large-scale datasets to make causal inferences about the effects of programs and policies. This revolution has the potential to dramatically improve our ability to rapidly conduct high-quality evaluations, giving us more and better evidence about “what works” than ever before. Elmendorf and Shanske are right to note that the “causal revolution” is here and lawmakers, scholars, and advocates on all sides should consider how we can make the most of this moment to secure better schools for all children. To understand the opportunities provided by the current moment, we need to understand that it has been made possible by several converging trends.

First, there is more and better educational data available. State longitudinal datasets have grown out of *No Child Left Behind* and other accountability reforms. Prior to NCLB it was not possible in most states to get reliable annual student achievement data, let alone data disaggregated by demographic categories. The move away from percent proficient accountability measures to “growth” or value-added measures has pushed states to develop data systems that allow us to track children’s performance (and teacher performance) over time in unprecedented ways. Other student, teacher, and school-level datasets (such as those from the Office of Civil Rights) similarly allow us to track trends across the entire population of schools and districts in the nation. And Stanford’s SEDA Archive has put every district’s (and soon every school’s) performance on a single national scale, permitting ever more comprehensive policy analyses. Researchers have even made linkages from education to taxes to estimate the long-term impacts of teachers and schools.³²

Second, computational power continues to increase dramatically and be more widely distributed. While several decades ago only the most sophisticated researchers at major research institutions could run complex statistical models, these tools are now available to the average researcher with the right training, statistical software, and access to a reasonably powered laptop. For even the largest and most complicated datasets and analyses, technological capacity is simply no longer an impediment.

Third, the quality of quantitative educational research has dramatically improved in recent years due to improvements in training (largely a result of the Federal Institute of Education Sciences providing training grants to universities) and access to the datasets described above. As applied economists have realized the quality and sophistication of educational data systems, they have swarmed to, and begun dominating, quantitative research in the field. While academic and think-tank researchers still lead the field, many school districts now even have their own in-house research outfits staffed with researchers capable of producing tailored analyses on the fly.

32. Raj Chetty, John N. Friedman & Jonah E. Rockoff, *Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood*, 104 AM. ECON. REV. 2633 (2014).

In short, it is clear that the capacity to conduct large volumes of high quality causal research exists in the education research enterprise. While Elmendorf and Shanske make an important contribution by helping us think through how the legal system might leverage this revolution to address the problems of causality that have stymied education litigation, we think there are at least three major challenges to the agenda they lay out.

First, there is a well-documented history of educational policies and reforms being implemented with meager fidelity.³³ Thus, while judges could order states to make administrative data more accessible, they could not order that the data produced be trustworthy. In large swathes of educational-impact evaluations, it is unclear whether a null effect is due to an intervention that does not work or an intervention that was never really implemented. For example, the standards-based reform movement that has dominated U.S. education policy for several decades has produced, at best, modest positive effects on student learning.³⁴ One likely explanation, though, is that the standards have been misinterpreted and weakly implemented in classrooms.³⁵

It is likely impossible that the issue of implementation fidelity could be addressed through administrative data alone—understanding implementation requires qualitative or survey data. While the article acknowledges that qualitative work will be a necessary complement to large-scale quantitative analyses, these cannot be devised as an afterthought; they must be considered an integral part of the general viability of these proposals, and judges will have to be prepared to compel states to develop the infrastructure to do these analyses well and routinely. If not, we can expect litigants on both sides to seize on this uncertainty to reproduce the current state of affairs, where ambiguity is selectively leveraged to suit an advocate's preferred legal theory.

Second, while the ability to do sophisticated impact analyses has undoubtedly increased, there are many messy complications that make it challenging to construct a “what works” agenda in education. For instance, many educational policies and interventions work in some settings but not others: some charter schools are more effective than their local traditional public schools, while others are worse.³⁶ Vouchers seem to work in some states but not others.³⁷ Even for something as simple as comparing the effectiveness of

33. See, e.g., Milbrey W. McLaughlin, *Learning from Experience: Lessons from Policy Implementation*, 9 EDUC. EVALUATION & POL'Y ANALYSIS 171 (1987).

34. Thomas S. Dee & Brian Jacob, *The impact of No Child Left Behind on student achievement*, 30 J. POL'Y ANALYSIS & MGMT. 418 (2011).

35. Morgan S. Polikoff, *Instructional Alignment under No Child Left Behind*, 118 AM. J. EDUC. 341 (2012); JAMES P. SPILLANE, *STANDARDS DEVIATION: HOW SCHOOLS MISUNDERSTAND EDUCATION POLICY* (2009).

36. Center for Research on Education Outcomes (CREDO), *National Charter School Study* (CREDO, 2013), <http://credo.stanford.edu/documents/NCSS%202013%20Final%20Draft.pdf>.

37. David Figlio & Cassandra Hart, *Competitive Effects of Means-Tested School Vouchers*, 6 AM. ECON. J.: APPLIED ECON. 133–56 (2014); Jonathan N. Mills & Patrick J. Wolf, *Vouchers in the Bayou: The Effects of the Louisiana Scholarship Program on Student Achievement After 2 Years*, 39 EDUC. EVALUATION & POL'Y ANALYSIS 464–484 (2017).

two textbooks, studies will often return conflicting results.³⁸ The fact that the conclusions are causal does not change the fact that the effects of the interventions are unclear. When states decide to persist or terminate policies in the face of these results, judges will still likely face difficult calls about whether a state has reasonably calculated its efforts to secure educational opportunity for all children.

Another messy complication is that even advanced econometric techniques sometimes fail. The most common reason for this failure is when there is selection on unobservables. When schools or districts choose which policies or programs to implement, we are sometimes able to model that selection process. Often, though, we are not, in which case our econometric tools cannot overcome the challenges to causal inference. To be sure, Elmendorf and Shanske's suggestion that judges consider randomization is a good one, and it might successfully mitigate some of these challenges. But even still, randomization often only reflects an intent-to-treat; implementation is often another matter. Parents sometimes prevail on principals to adjust class assignments,³⁹ and teacher and student attrition can thwart even the best-designed intervention. Furthermore, unless school districts are compelled to participate in these randomized studies, there is no guarantee of the generalizability of results from volunteer districts to non-volunteers (and if compelled, the results may not generalize to districts that subsequently have choices as to whether or not to adopt a particular program).

Even when research can be done and returns consistent findings about impact, it often leaves many important questions about tradeoffs unanswered. For example, the tradeoffs of choosing one policy over another are rarely addressed clearly in research. Cost-benefit analyses are sometimes included in research, but not always—these would need to become routine, and standards for when interventions were “worth it” would need to be created. Research cannot always speak to scale. What works with careful attention in one site may not work when brought to a larger scale, and efforts to rapidly scale a project may outstrip existing capacity and result in a much lower quality treatment in later years. To be clear, these challenges do not invalidate the important work that can be done with high quality educational data. They merely point to the challenges of relying on causal research to inform policy and practice.

Third, the timeline for quality research using educational data often is very short—test score effects in the next year, effects on grade promotion, graduation from the current school, etc. Gains in the short term are important, but fade-out is a problem in much of educational research.⁴⁰ Even still, some-

38. Rachana Bhatt & Cory Koedel, *Large-Scale Evaluations of Curricular Effectiveness: The Case of Elementary Mathematics in Indiana*, 34 *EDUC. EVALUATION & POL'Y ANALYSIS* 391 (2012).

39. Barbara Nye, Larry V. Hedges & Spyros Konstantopoulos, *The Effects of Small Classes on Academic Achievement: The Results of the Tennessee Class Size Experiment*, 37 *AM. EDUC. RES. J.* 123–151 (2000).

40. Janet Currie & Duncan Thomas, *School Quality and the Longer-Term Effects of Head Start*, 35 *J. HUM. RESOURCES* 755 (2000).

times short-term effects do not align with the long-term effects of various policies (for instance, some policies seem to boost test scores in the short term but not attainment or other long-term outcomes).⁴¹ Elmendorf and Shanske recognize these issues and rightly argue that we should do what we can to determine the effects of schools on lifetime outcomes—college attendance, income, civic participation, and general welfare in adulthood. But in some respects, this longer timeline only compounds these challenges while adding new ones. For example, Raj Chetty’s groundbreaking work on the long-term effects of having a highly effective teacher has been criticized because the study began before the current accountability era and was based on a low-stakes exam.⁴² While it may be the case that teacher effects on current state exams have the same degree of long-term persistence, we of course cannot know this until enough time has elapsed. Then the question arises: how should we interpret—let alone apply—decades-old findings in a different policy context? If we found that two decades ago having access to computer courses and typing skills resulted in those students having a disproportionately high percentage of high paying technology jobs, would we still believe that this was the case two decades later? If we thought it was actually the skills in the classes that were valuable, then perhaps. But it seems just as likely that there was a first mover advantage; the value was in the novelty of the skills. The longer the timeline and the more complex the outcome measure (income is easy to measure but a limited measure of thriving democratic citizenship), the more difficult these measurement and interpretive issues become.

We hope that our raising of these issues will not be mistaken for pessimism about the utility of the authors’ proposed legal intervention; quite the contrary. If enacted, the proposal has the potential to advance our knowledge and the conversations within the field of education research (and litigation) in substantive ways. But while we are confident that the authors’ proposal would result in better descriptive information, we are less sanguine about the possibility of better access to data being a silver bullet for improving our knowledge of “what works” and taking that to scale.

PART III

Though the challenges we have raised in Part II are serious and they ought not be overlooked, there are real potential gains to be made from the authors’ proposal. In this final section, we discuss a case where implementation of the authors’ proposal would have made a meaningful difference. Examining this case helps us clarify the potential gains and the lingering challenges of what Elmendorf and Shanske put forth.

41. Rebecca Unterman et al., *Going Away to School: An Evaluation of SEED DC* (MDRC, 2016).

42. Michael Winerip, *Study on Teacher Value Uses Data From Before Teach-to-Test Era*, N.Y. TIMES, January 15, 2012, at A13.

In 2000, a group of schoolchildren from the San Francisco Unified School District sued the state of California, arguing that the state had failed to provide them with equal access to adequate school facilities, curriculum materials, and high-quality teachers. Four years later, the state opted to settle the case, *Williams v. California*, by passing legislation that laid out new requirements, including establishing a right to sufficient textbooks for all California students and requiring yearly self-evaluations of textbook sufficiency for each school, the results of which must be publicly reported in a school accountability report card (SARC).⁴³ The basic theory of the case—enacted in the settlement legislation—was that the state has a responsibility to provide all children access to basic educational necessities and that part of this responsibility includes developing administrative systems capable of monitoring that access. The settlement gave local communities additional leverage to hold the state accountable for these obligations by requiring that the monitoring data be public and by establishing procedural rights for parents and community members seeking to redress reported inadequacies.⁴⁴

This emphasis on administrative data systems and production of information about school conditions shares important similarities with the proposal envisioned by Elmendorf and Shanske. Though the goal in *Williams* was not specifically to allow for causal analysis of the effects of these basic educational necessities, its data reporting requirements represent similar thinking about the importance of state-produced data and represent similar administrative burdens for districts. In this sense, we think that the experience of *Williams* offers a useful example for thinking through, in ways more concrete than the prior two sections, the potential benefits and challenges of Elmendorf and Shanske's new proposal.

An important component of the monitoring system created by the *Williams* settlement involved tracking districts' access to standards-aligned textbooks. Though textbook titles are a weak proxy for educational opportunity, district reports of textbook titles offer a measure of their baseline ability to produce basic information with high fidelity. We say baseline because, unlike many kinds of educational opportunity data involving considerable administrative complexity, evaluative frameworks, or subjective judgment, the information requested in this case is clear, easy to ascertain, and definitive (i.e., not

43. General information about the *Williams* case and settlement can be found at <https://www.cde.ca.gov/eo/ce/wc/wmslawsuit.asp>. For more detail on the specific requirements on textbook reporting, see Sally Chung, *Williams v. California: Lessons from Nine Years of Implementation* (ACLU, 2013), http://decentsschools.org/settlement/Williams_v_California_Lessons_From_Nine_Years_Of_Implementation.pdf.

44. William S. Koski, *Achieving 'Adequacy' in the Classroom*, 27 BC THIRD WORLD L.J. 13, 42 (2007) (The *Williams* deal clarified the State's obligation to prevent, detect, and correct the denial of basic educational necessities and provided children and children and their communities both a monitoring system and procedural rights to hold the State accountable for that obligation").

open to interpretation).⁴⁵ Despite the statutory reporting requirement and straightforward reporting task, recent research finds that the data produced by the state under the settlement were limited in a number of ways.⁴⁶ For example, about 10% of schools are not reporting *any* data at all to the state. Many other schools provide textbook titles that do not include the detail necessary to determine whether the textbook aligns with state curricular standards.⁴⁷ The large majority of schools provide textbook information in a format that requires substantial expert knowledge to interpret (for instance, listing just a publisher and the most recent adoption year and requiring the reader to infer that it refers to the most recent state-adopted book by that publisher).⁴⁸ And, while reporting the data is a statewide requirement, the data reporting itself has not been centralized: the data are available only from non-standardized, district-created PDFs. Creating a statewide picture of textbook use, therefore, requires considerable effort to gather and clean for analysis. The consequence of these limitations is that, despite more than a decade of data collection, very little work has actually examined the extent to which the deficiencies that triggered the lawsuit have been addressed by the settlement.⁴⁹ Furthermore, while the methodological tools to analyze textbook data to investigate impacts on student achievement have existed for several years,⁵⁰ California's data had never been used to investigate these issues until very recently.

In retrospect, it appears that the 2004 *Williams* settlement was inadequate along several dimensions:⁵¹ data fidelity, value, and use. But suppose the state's settlement had compelled the kind of data collection and use described in Elmendorf and Shanske's proposal. What would have been different?

First, the state would have collected better data on school districts' textbook adoptions over the past thirteen years. The specific charge of collecting data for the sake of answering questions about the provision of education means that the effort would have to have been more standardized and centralized from the beginning—an improvement on the idiosyncratic reporting conducting under *Williams*. Though data on textbook collection was required by law, it is likely that a judge directing the collection of these data would have intervened as evidence of non-compliance began to accumulate. The built-in audience for the data, as opposed to data production for generic transparency or

45. Efforts to examine the distribution of “highly qualified” or “high quality” teachers would be an example where the administrative complexity is high. For instance, it is not clear to define high quality (credentials versus teaching behaviors vs. test score impacts), and if either of the latter two definitions are chosen there is a great deal of measurement difficulty.

46. Cory Koedel et al., *Mathematics Curriculum Effects on Student Achievement in California*, 3 AERA OPEN 1 (2017).

47. For instance, some districts listed ‘Houghton Mifflin’ as a mathematics book when that publisher produces a half dozen elementary textbook series—only some of which are on the approved list.

48. *Id.*

49. For notable exceptions see, Chung, *supra* note 45.

50. Bhatt & Koedel, *supra* note 39; Rachana Bhatt, Cory Koedel & Douglas Lehmann, *Is curriculum quality uniform? Evidence from Florida*, 34 ECON. OF EDUC. REV. 107–121 (2013).

51. For a more sanguine assessment of the *Williams* settlement see, Chung, *supra* note 45.

public use, likely matters when it comes to producing a high-quality dataset. Still, the *Williams* case offers a cautionary tale about the need to build administrative capacity around data collection. The assumption that districts can produce standardized information—even concerning something as straightforward as textbook titles—is likely to be violated. This is no small matter given that the data collection envisioned by Elmendorf and Shanske is exponentially more expansive and complex. Tracking lifetime outcomes will require faithful reporting across a wide variety of areas over an extended period of time and likely across a range of local and state jurisdictions. The inability of California districts to report textbook information reliably, despite a legal requirement to do so, should give us serious pause.

Second, more complete and systematically collected data would have allowed researchers or advocates to investigate certain key research questions over that period. Elmendorf and Shanske specifically envision judges issuing “temporary experiment remedies” that would explicitly call for the evaluation of implemented remedies—thereby addressing potential issues around subsequent attention and publication of the research. Judges could require the state make available the necessary data for studying the matter and ensure that the findings do not get locked away in a file drawer.⁵² In the context of *Williams*, scholars would have been able to investigate whether curriculum materials really were equitably distributed, and we would, therefore, have been able to address inequities stemming from access to these resources. As it stands, we still do not really know whether equitable access has been achieved. We also would have been able to go beyond questions of access to questions of effects on student achievement. Conducting textbook impact analyses over the duration of this period might have allowed us to evaluate the effectiveness of specific textbooks or specific pedagogical approaches embedded in textbooks, though, as we describe above, there are several technical challenges that make this work difficult. As it currently stands, all textbook impact analyses have been completed too late to inform any district adoptions (because standards or available textbooks have changed).

Third, and very much related, the Elmendorf and Shanske proposal recognizes the importance of both the production and consumption of information. That is, judges should not simply direct the state to produce information but must also serve as an audience for the results. This would represent a major advance over what occurred in the *Williams* settlement where it was assumed that the production of information would automatically ensure its consumption and use by the public. The volume of missing and uninterpretable data and lack of official or public notice of these deficiencies over more than a decade suggests that this was not the case. The Elmendorf and Shanske proposal tightens this loop considerably by explicitly tying the production of information to a specific

52. See Christopher S. Elmendorf & Darien Shanske, *Solving “Problems No One Has Solved”: Courts, Causal Inference, and the Right to Education*, 2018 U. ILL. L. REV. 693, 742.

empirical, and potentially judicial, question. Though we agree with Elmendorf and Shanske that court direction of these research endeavors—especially around assigned randomization—will minimize political influence on these investigations, honest disagreements about the practical significance of research findings remain considerable, even in instances when political influence or ideological predispositions have been minimized. For instance, few contest the findings that class size reduction has an effect on student achievement, but questions about whether the gains justify the immense cost are another matter. The same might be said of debates about charter schools, which often turn on questions of the relative weight given to potentially non-commensurable values. Regardless of their ultimate interpretation, however, bedrock findings about major educational phenomenon are difficult to produce and Elmendorf and Shanske are right to note that state bears a large responsibility for creating or eliminating this difficulty.

CONCLUSION

The field of education research has long been maligned as offering findings that are short on rigor and relevance while being long on idealism and ideology. Decades of efforts to improve the quality of research and make the field more “scientific” have met with mixed results. According to one prominent scholar, education research remains the “elusive science”⁵³; in the words of another, the immense difficulty of advancing knowledge in field is one of its defining characteristics: “if Sisyphus were a scholar, his field would be education.”⁵⁴ The challenges inherent to the field are difficult enough without being compounded by states’ lack of attention to data collection and access. This is especially true at a moment when developments in technology and experimental and quasi-experimental methods are providing new opportunities to understand the effects of key aspects of our education system on student learning and life outcomes. By raising the issue of what judges can do in the context of education litigation to direct the collection and analysis of information and how they might use that to break a decades old logjam of remedy claims of uncertain value, Elmendorf and Shanske have made a major contribution to the field. Even if advocates and judges take up their proposal, it may be that the answers to the “problems no one has solved” may remain elusive—but at least they will have ensured that the data won’t be.

53. ELLEN CONDLIFFE LAGEMANN, AN ELUSIVE SCIENCE: THE TROUBLING HISTORY OF EDUCATION RESEARCH (2002).

54. David F. Labaree, *Educational Researchers: Living With a Lesser Form of Knowledge*, 27 EDUC. RESEARCHER 4, 9 (1998).