MAKING EVIDENCE LOCALLY

RETHINKING EDUCATION RESEARCH UNDER THE EVERY STUDENT SUCCEEDS ACT

by THOMAS J. KANE

THE NEW FEDERAL EDUCATION LAW, the Every Student Succeeds Act (ESSA), envisions a powerful role for states in managing the evidence base behind school improvement efforts. Not only must they certify that interventions meet the "evidence-based" requirements spelled out in the law, they also must monitor and evaluate federally funded school-improvement efforts going forward. There's only one problem: states have never played such a role before.

In order to fulfill this obligation, states will need a scalable model of impact evaluation which could operate at the local level, where decisions are being made.

ILLUSTRATION / JON KRAUSE

States should adopt a simple goal: any major initiative involving more than 100 classrooms should be subject to a local pilot test before being rolled out. In other words, districts should be running their own small-scale impact studies, implementing interventions in a subset of their classrooms, establishing comparison groups, tracking and comparing results, and acting on the evidence. That's been the path to improvement in a variety of fields, from pharmaceuticals to retail sales. Given our incomplete understanding of the way students learn and teachers change their teaching, it is the only path to sustained improvement in U.S. education.

After a decade of investing in state and local data systems, many of the components of such a system—like longitudinal data on individual students and indicators matching students to teachers—have already been built. But some key pieces are still missing. We need a way to pool data among school districts, most of which are too small to assemble sufficient comparison groups on their own. We need a quicker and less expensive route to launch impact evaluation studies rather than the current costly and time-consuming practice of designing each new study from scratch. And local education agencies

need an ongoing analytic partner that can standardize key parts of research analysis, such as how comparison groups are identified. Finally, local leaders need new venues for synthesizing results, comparing notes, and choosing which interventions to test next.

The Every Student Succeeds Act provides an opportunity to put these final pieces in place and spread such an approach nationally. In this essay, I describe how a state could use the authority and resources provided by ESSA to launch a system of "efficacy networks," or collections of local agencies committed to measuring the impact of the interventions they're using. An overlapping system of efficacy networks working with local agencies would create a mechanism for continuous testing and improvement in U.S. education. More than any single policy initiative or program, such a system would be a worthwhile legacy for any state leader.

An organizational mismatch

The United States spends about $620 billion per year on K–12 education nationwide. Only about $770 million of that goes to education research, through the federal Institute of Education Sciences (IES) and the National Science Foundation (NSF)(see Figure 1). There is no estimate of state and local spending on education research because it is nearly nonexistent. Across the economy, our nation spends 2.8 percent of gross domestic product on research and development overall. If we invested a similar percentage of the K–12 education budget on research and development, we would be spending $17 billion per year rather than $770 million. We are clearly under-invested.

Still, education research has yielded some important successes in recent years. Perhaps the most valuable byproduct of the No Child Left Behind Act (NCLB) has been the resurgence of research on the effects of teachers on student achievement, which has informed the redesign of teacher evaluation systems. Moreover, although many have lamented the shortage of interventions with positive results in the What Works Clearinghouse, even null results represent progress. For example, the failure to find positive student-achievement impacts in a series of IES-funded studies of professional development programs has produced a broader appreciation of the difficulty of adult behavior change and more healthy skepticism about the traditional approach to teacher training. A search for more effective models is underway, involving more intensive coaching and feedback, buttressed by strong curricular materials. More recently (some five decades after the Coleman Report told us that schools would be unable to close the black-white achievement gap alone), research has identified charter school models sufficiently powerful to close the gap.

The sizable gap between education research and local decisionmaking calls for a profound shift in strategy aimed at ensuring that our evidence making is better integrated with the way decisions are reached.

In a 2016 survey on research use by state and district decision-makers, only 1 to 4 percent reported that they use federally-funded research sources "all the time."

Despite that progress, we have a long way to go to build an evidence-based culture in the state and local agencies, where most decisions are made. In a 2016 survey on research use by state and district decisionmakers by the National Center for Research in Policy and Practice, more than half reported that they "never" or "rarely" used the federally funded What Works Clearinghouse, the National Center for Education Statistics, and the Regional Educational Laboratories (RELs). Only 1 to 4 percent reported that they used these sources "all the time." Even the most popular source of research, professional associations, was described as being used "all the time" by just 14 percent of respondents. In my work, I have studied school-board minutes in 17 of the 20 largest U.S. school districts between 2010 and 2016, and found that the term "What Works Clearinghouse" appeared only once. The term "Institute of Education Sciences" appeared 15 times in total, but did not appear at all in 11 of the 17 districts.

It would be natural to continue tweaking the parameters of the existing research model: the IES budget, the process for soliciting and evaluating proposals, the process for reviewing and releasing results. However, the sizable gap between evidence and local decisionmaking calls for a more profound shift in strategy. Rather than finding ways to generate more research of the same type, we need to ensure that our evidence making is better integrated

PHOTOGRAPH / RLTHEIS;TWENTY20

with the way decisions are reached. And that requires a different model.

Previously, I have argued that the centralized approach to efficacy measurement used by the IES is insufficient, given the dispersed nature of decisionmaking in U.S. education (see "Connecting to Practice," *features,* Spring 2016). In other fields, such as pharmaceuticals, where supply decisions are federally regulated, such a centralized system of evidence gathering makes sense. However, federal regulators do not choose educational products and strategies—and they never will. Local leaders do. While a federal regulator will care about the average impact of an intervention across a range of sites and subgroups, a local bureaucrat needs evidence to help persuade local interests to invest in a given intervention.

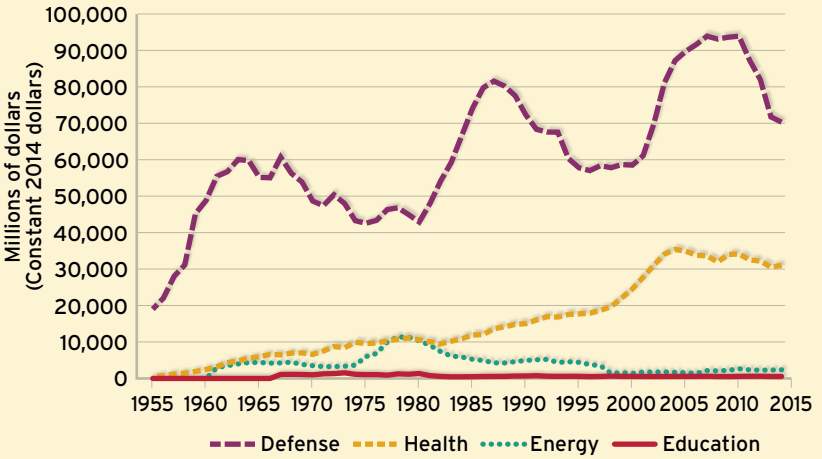
Moreover, when making decisions about the federal programs it controls, the federal government can invest in the knowledge and expertise of a small group of experts to keep itself informed. But a small group of experts cannot possibly advise the thousands of local actors working in school districts, nor give them the evidence they will need to persuade their colleagues.

Following the tradition of James Q. Wilson, we need to understand the system of rewards and constraints within which local decisionmakers work. Especially in an age of outcomes-based accountability, district leaders cannot ignore student achievement. Even where there is limited school choice, school boards and superintendents feel the pressure when outcomes lag. Nevertheless, such pressure is counterbalanced by other, more parochial concerns, such as the preferences of the local school board, superintendent, principals, department chairs, and parents. No chief academic officer ever got fired for choosing an intervention deemed ineffective by the What Works Clearinghouse, but plenty have lost their jobs after a dispute with a school board member or after a committee of

Underinvestment in Education Research (Figure 1)

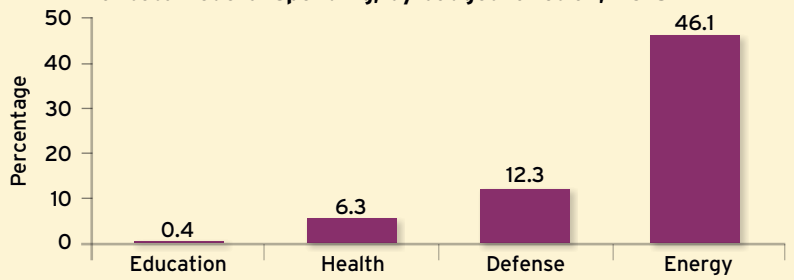
(1a) The United States spends over 130 billion dollars on research and development, but the majority of that spending is devoted to research in defense and health, with less than a billion spent on education research. The marginalization of education research is not a new phenomenon.

Federal funding for research and development, by budget function



(1b) Although greater shares of spending are devoted to research across other budget functions, federal spending on research in education accounted for less than one percent of total federal spending on education. This lack of investment in education research is understated, as 90 percent of education spending in the United States is at the state and local level, where there is almost no spending on research and development.

Federal spending on research and development, as a percentage of total federal spending, by budget function, 2015



NOTE: Figure 1a does not include spending under the American Recovery and Reinvestment Act of 2009. Overall and research and development spending on education includes elementary, secondary, and vocational education, higher education, research and general education aids, training and employment, other labor services, and social services. Health includes health care services, health research and training, and consumer and occupational health and safety. Energy includes energy supply, energy conservation, emergency energy preparedness, and energy information, policy, and regulation. Expenditures must meet the OMB definition of "research and development" to be included. Although the National Science Foundation funds roughly $200 million on education research, such expenditures have not been included in the above.

SOURCE: Office of Management and Budget Historical Tables, and the National Science Foundation, Federal R&D Funding, by Budget Function, NSF 15-306 and NSF 17-305

department chairs complained.

In addition, while regulators care about average effect sizes, practitioners want to know whether a given intervention will work in their own classrooms. Although researchers have historically interpreted such a posture as being parochial and unscientific, it has some justification. We are learning that the answer to a seemingly simple question, "Does a program work or not?" often varies. As discussed by researchers Michael Weiss, Howard Bloom, and Thomas Brock, the impact of a given program could vary for four reasons: variation in the quality or dosage of the intervention being evaluated (for example, due to the skills of the local teachers); variation in the quality or dosage of the services available to the control group; variation in the impact of the treatment for different subgroups of teachers or students; and variations in the context of the intervention.

It turns out that local leaders are correct to wonder whether the national studies apply to them. In a review led by Weiss of 11 multi-site studies involving elementary or secondary students and two postsecondary studies in which treatments were randomly assigned, 9 out of 13 studies had statistically significant differences in impact by site for at least one of the outcomes measured.

If we want local leaders to make decisions on the basis of evidence—and be rewarded for it—we need to provide them with evidence denominated in their local currency—their own students' achievement.

Under ESSA, let a thousand pilots fly

If such evidence would be so valuable, why don't more leaders seek evidence now? There are two primary reasons: First, such pilots are currently not feasible for most districts. Few agencies have the technical staff to create treatment and comparison groups, pull together the data, and analyze the results. And, second, smaller districts—where most Americans attend school—do not have a sufficient sample size within their own data to detect the hoped-for effects. They need an intermediate organizational player to pool their data with other agencies to help them learn.

ESSA is the first federal education law to define the term "evidence-based," which appears 63 times across its various titles and programs.

We should recognize the progress that's been made. The data infrastructure for measuring student achievement over time and linking students to schools and teachers has improved dramatically over the past decade. Prior to 2002, few states had unique student identifiers to track students across time and across school districts, and annual testing in consecutive grades, which shows how student achievement changes over the course of a year, was not mandated. Moreover, before 2009, it was rare for school districts and states to link student data to specific classrooms and teachers. Since many interventions are targeted at or carried out by teachers, we need to know which teachers are working with which students, and we need to identify comparison groups of students to assess different types of interventions. Much of that information is now available.

The term "evidence-based" appears 63 times across the various titles and programs of ESSA. Although an analogous term, "scientifically based research," appeared in the preceding NCLB, it was never effectively defined. In contrast, ESSA defines four levels of "evidence-based" practices: "strong," with at least one well-designed and well-implemented experimental study with a statistically significant, positive effect; "moderate," with at least one well-designed and well-implemented quasi-experimental study such as a matched-comparison group; "promising," with at least one well-designed and well-implemented correlational study with statistical controls for selection bias.

Crucially, the law also establishes an "under evaluation" category, to describe interventions that do not yet meet the more stringent standards but have either some research-based rationale or are subject to ongoing evaluations. Evidence-based interventions in the "under evaluation" category may be implemented in all but the lowest-performing schools under Title I—either those in the bottom 5 percent of all schools, or those slated for "targeted assistance" after at least one subgroup of students falls short of state benchmarks. At those schools, interventions must meet the criteria in the other three categories, a requirement that will inevitably drive traffic toward the federal What Works Clearinghouse.

However, the evidence requirements will amount to nothing without state leadership. The law leaves it to the states to decide how much they want to build an evidence base and nudge districts toward choosing more effective strategies. No doubt, many states will turn the "evidence-based" requirement into an empty compliance exercise, describing evidence-based requirements so broadly that districts will find it easy to fit any intervention plan within them. But state leaders who want to do more could do so in two ways.

First, ESSA leaves it up to state agencies to determine which interventions meet the "strong," "moderate," or "promising" evidence standards, as well as what constitutes a "well-designed and well-implemented" study. By controlling evidence requirements, states have an opportunity to filter how federal funds can be used. To bolster the legitimacy of those decisions, they could impanel external experts to sift through the evidence and identify school turnaround models, professional development programs, and high school–dropout prevention strategies with the strongest evidence of impact.

Second, and more important over the long term, states could encourage or even require local agencies applying for competitive funds (such as school improvement funds, teacher and school leader incentive funds, and early-childhood literacy programs under Title II) to participate in an "efficacy network." An efficacy network would be a collection of local agencies committed to measuring the impact of interventions they're using. Independent organizations such as universities, research firms, or other nonprofits would apply to the state to organize and support a network, which could be organized by region, outcome, or type of intervention. In joining a network, an agency would agree to pool its data, collect common outcomes such as a common interim assessment or teacher surveys, and work with the network organizer to establish a comparison group for each major intervention it implements. Finally, it would agree to share its findings and compare notes with others in the network.

States could pay for efficacy networks with the 5 percent of school improvement funds expressly set aside for evaluation and dissemination, or other administrative set-asides. Moreover, they could provide regulatory guidance allowing districts to use a portion of their share of federal dollars to pay the cost of participation.

Even with an enhanced state and local role, there is also a role for federal research programs to play. IES could support states by offering guidance on how they might review and approve interventions meeting the top-three evidence categories. Through its state and local partnership grant program, IES also could support researchers working with a state to review current interventions or evaluate future ones. Most importantly, the federally managed Regional Educational Laboratories could support states to assess existing evidence and evaluate ongoing school improvement efforts.

To have an 80 percent chance of detecting the impact of an intervention that raises student achievement by an average of 2 percentile points over the course of a year in elementary math classrooms in New York City, one would need roughly 200 classrooms.

Local but not anecdotal

The primary audience for the efficacy networks must be district or charter-network leadership, not schools or teachers. I say that not due to any preference for a hierarchical approach to school administration, but in recognition of the need for aggregation in identifying effective interventions. There is an inherent tradeoff in conducting research on a scale small enough to identify local impact and large enough to discern a true impact from statistical noise. In striking the necessary balance, there is no avoiding a level of aggregation above the school and classroom level, because the impact of most education interventions is small relative to the variation in student achievement at any point in time.

For instance, detecting a math achievement impact equivalent to being assigned an experienced teacher as opposed to a novice teacher—typically .08 standard deviations—requires compiling the results of roughly 200 classrooms: 100 in an intervention group and 100 in a comparison group. With such a snapshot, an evaluator

PHOTOGRAPH / ISTOCK

would have an 80 percent chance of detecting such an effect, assuming that one could control for students' baseline achievement. Because such sample sizes are critical, a key role of the networks would be to pool together the minimum number of classrooms to reliably discern reasonably sized impacts and to coordinate the choices of schools to form an intervention group, while remaining small enough to provide results that are authentically local.

I've seen the potential for districts to gather local evidence on the efficacy of their programs in this way through the Proving Ground project at the Center for Education Policy Research (CEPR) at Harvard University, where I am director. In 2015, CEPR began working with 13 school agencies to develop a model to easily conduct low-cost, local pilots. By pooling data across a network of agencies, we help smaller districts and charter organizations meet the 200-classroom target sample size, with 100 classrooms each in the treatment and control groups. That enables them to evaluate interventions aimed at increasing student achievement by 2 percentile points, or .08 standard deviations. We also provide a network for sharing lessons and facilitating discussions about what districts might try next, based on the evidence they have collected.

From an initial assessment of their data, district leaders develop hypotheses about interventions they wish to try. CEPR helps them decide how many classrooms to include in a pilot. If the district is willing to use random assignment, we provide them with the software tools to select treatment and control groups. If not, we use algorithms to identify comparison students, employing a standard approach to matching on prior test scores and achievement.

Despite its attractiveness, the idea that small groups of teachers working together will deliver improvements over time is an illusion. Of course, some will correctly identify effective interventions by chance. But many more will draw the wrong conclusion and implement strategies that do harm. Indeed, we have ample evidence of the limits of such school and classroom-level trial and error, given the historical record of American education. We need new organizational structures such as the efficacy networks to provide some level of aggregation, while preserving the local context.

If we make it possible for local leaders to pilot their initiatives on a limited scale first, we would provide both more upside potential and less downside risk—and local leaders would be in a better position to claim credit for effective interventions.

Conclusion

In the next phase of U.S. education reform, we need to integrate evidence making with decisionmaking. Local bureaucracies administer our school systems, and local bureaucracies need locally generated evidence in order to make the case for change. Our current research infrastructure is not providing the type of evidence that persuades at the local level.

By necessity, the superintendency is, in part, a political job—and no politicians are eager to see an initiative in which they have invested their political capital fail. Participation in an efficacy network would reduce risk and increase the appetite for reform. If we make it possible for local leaders to pilot their initiatives on a limited scale first—say, by launching an intervention in 100 treatment classrooms and tracking results relative to a set of 100 comparison classrooms—we would provide both more upside potential and less downside risk. When an intervention is shown to be effective, local leaders would be in a better position to plausibly claim credit. And when an intervention fails, a leader may even be able to claim credit for stopping the process before rolling it out more broadly.

Education research can no longer be seen as solely a federal responsibility. Nevertheless, as noted above, there will still be an important role for the federal government to play. One will be to support states as they seek to implement ESSA. In addition, the type of careful, long-term intervention studies managed by IES could be used to validate the most promising forces bubbling up from the state and local learning networks.

The evidence requirements in ESSA provide state leaders with a powerful lever. To use it, they simply need to recognize that a system for testing ideas is more valuable than any single idea they might champion. There is no initiative that would have a similar effect 20 years from now, nor is there any worthwhile policy proposal—such as preschool education—that would not be enhanced if such a system were in place. The only thing required is longer-term vision, and faith in the desire of local actors to improve.

Thomas J. Kane is professor of education and economics at the Harvard Graduate School of Education and faculty director of the Center for Education Policy Research. Mallory Perry played the lead role in analyzing school committee records for discussions of research.