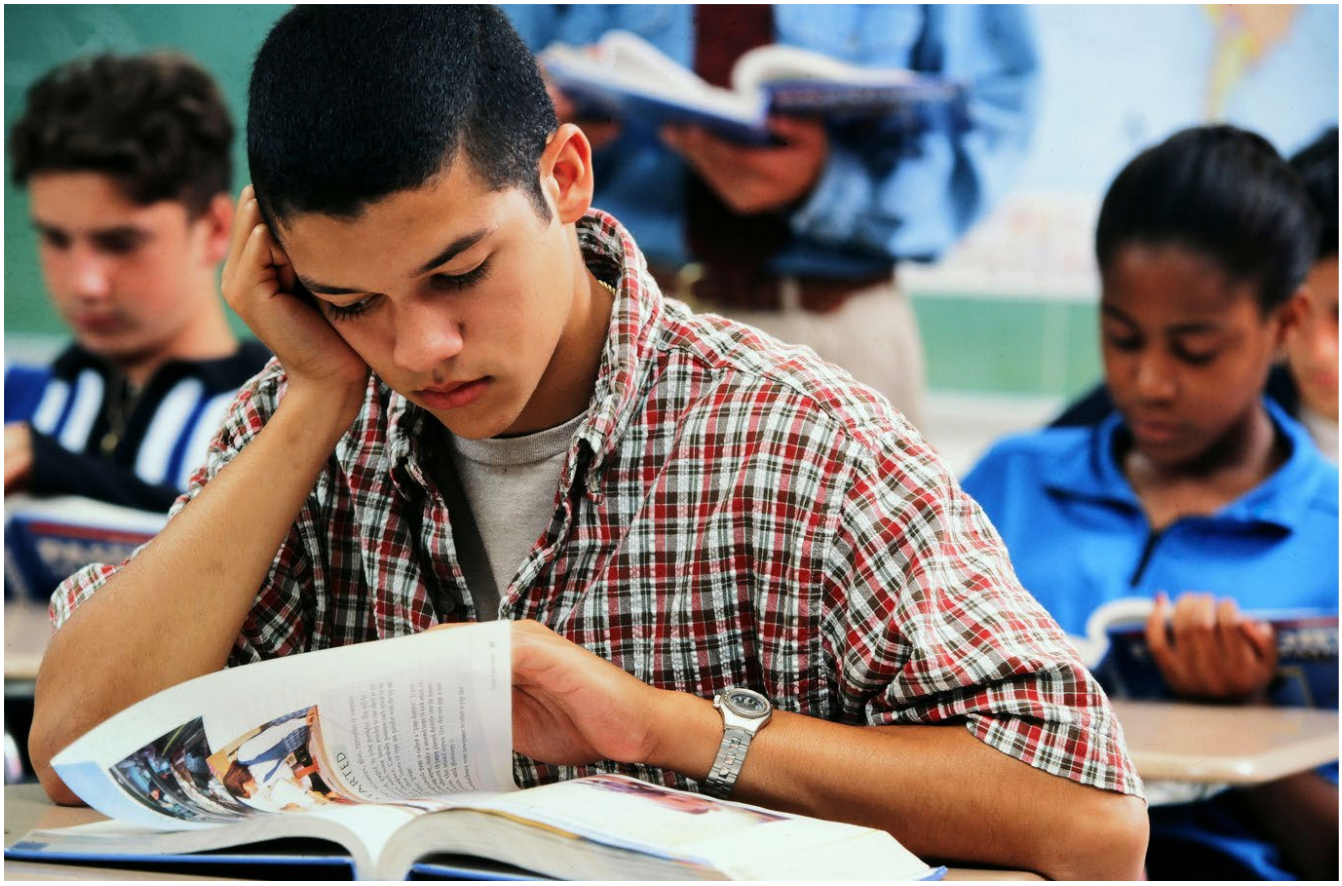
Education Next

FEATURE

The Myth of the Evidence-Based Curriculum

Why calls to choose instructional materials based on student learning data are currently impractical—and what it would take to make them a reality

By Morgan Polikoff and Jake Rowley



ALAMY

Momentum is building in the demand for better evaluation of curriculum materials—textbooks, workbooks, instructional software, and tests—but states and private foundations must lead by supporting data collection.

Curriculum materials are having a policy moment. After more than a quarter-century of standards-based reform, policymakers are recognizing that these reforms will likely not improve student learning without attention to the quality of the materials in teachers' hands.

This shift in focus began in the early years of the Common Core standards, when researchers questioned publishers' claims that their products aligned with the standards. The movement gained momentum as external organizations—most notably EdReports—began rating curriculum materials for alignment and usability, with their judgments influencing adoption decisions in many states. And now, as states

increasingly emulate Mississippi's science of reading reforms, a growing number of them are getting into the business of evaluating curriculum materials, examining how well they align with science of reading principles, encouraging schools to adopt highly rated materials, and ensuring teachers receive aligned professional development. This approach will almost certainly expand to subjects beyond reading. We applaud these developments.

Operating in parallel is a growing push for robust evaluation of the curricula, interventions, and technologies that schools use. This movement emphasizes evaluation methods that can reliably show whether a curriculum or intervention improves student learning—such as those adhering to the widely used research standards and procedures of the federal What Works Clearinghouse, which evolved into the influential "tiers of evidence" codified in the Every Student Succeeds Act (ESSA).

In an ideal world, schools and districts would be able to turn to WWC, or to privately funded efforts like the Evidence for ESSA project at Johns Hopkins University, to identify curriculum materials and other resources shown to be effective in improving student outcomes. Schools and districts adopting more evidence-based materials and practices seems like an unequivocally good thing.

But these two promising trends—states' growing interest in effective curricula and the movement toward better evaluation of them—have not combined to drive widespread implementation of more-effective materials, including textbooks, workbooks, instructional software, and tests. As a recent viral article on Curriculum Associates' iReady assessment system argues, technology products in schools are rarely evaluated using the highest tiers of evidence outlined in ESSA, with studies often conducted by the companies themselves using what the author asserts are weaker designs that produce questionable evidence. A second viral article went further, trashing iReady research and questioning the value of edtech generally in the face of what the author sees as weak evidence of efficacy.

Curriculum Associates is far from the only company in this position—it's common throughout edtech—but we don't think the industry is to blame. Based on our experience consulting with curriculum companies, we believe most are committed to improving their products and subjecting them to rigorous evaluation. Overall, these companies are doing their best to create high-quality materials, respond to ever-evolving state and district requirements, and rapidly demonstrate their products' effectiveness, despite practical and financial challenges.

While it's tempting to blame the curriculum companies and technology providers, the problem is more nuanced: States have failed to track which curricula schools are adopting, leaving independent researchers without the data needed to study what works. At the same time, the dominant standards used to evaluate education research set unrealistic expectations that are inconsistent with how that research is actually conducted. To establish a functional system for understanding the effects of curricula and other interventions on student outcomes, states and private foundations will need to lead the way.

A Primer on the ESSA Evidence Tiers

The tail that wags the dog is the Every Student Succeeds Act evidence tiers. Enacted in December 2015, ESSA replaced No Child Left Behind and introduced a formal, tiered definition of the term "evidence-

based," which the U.S. Department of Education has since fleshed out in guidance:

- Tier 1, the highest standard, requires a statistically significant positive effect on student outcomes drawn from a rigorously conducted experimental study with at least 350 participants who are comparable on baseline characteristics, across more than one district or school.
- Tier 2 applies the same criteria but uses a quasi-experimental research design (that is, one that attempts to produce causal evidence using observational data).
- Tier 3 evidence requires the same positive, statistically significant finding but can be drawn from a correlational study that attempts to control for selection bias, or from an experimental or quasi-experimental study that does not meet the criteria for sample size or setting.
- Tier 4 allows the combination of a logic model grounded in rigorous research plus a plan to conduct an impact evaluation as the product or intervention is implemented.

These tiers are intended to help leaders cut through marketing claims and base decisions on what is known about an intervention (such as a new curriculum) and its impact on student achievement. While resources such as EdReports provide valuable ratings of curricula based on their alignment to standards and usability by teachers, they stop short of where the rubber hits the road: evidence of effects on student achievement. Systematically studying effects on student learning could provide a critical, complementary resource districts could use in selecting new materials. Districts and states could go a step further and require such evidence for a curriculum's inclusion on their lists of approved instructional materials. In a loose comparison between education research and the medical model, this stipulation would be akin to requiring evidence of drug effectiveness before making it available for doctors to prescribe.

The ESSA evidence tiers have an important role to play in guiding curriculum adoption decisions, but they don't apply perfectly to the evaluation of core curriculum materials. A few examples show why not:

- The evidence tiers privilege randomized experiments—and the viral critiques mentioned above stress their primacy—but we will argue below that strong quasi-experimental designs may well confer at least as much value, if not more, in this context.
- A curriculum can attain Tier 1 status through an experimental study of just 350 students across two sites—far too small a scale to produce confidence in the validity of results or to examine school-to-school variation in implementation and effects.
- The federal guidance does not address the issue of study duration, and the influential Evidence for ESSA project requires a minimum of a mere 12 weeks "from program inception to posttest"—not enough time for testing the effects of core curriculum materials.

Better guidelines for evaluating curricula would emphasize longer time horizons (at least a full academic year) and larger sample sizes (especially more "sites," that is, the units adopting the curriculum under study). They would also place more value on carefully conducted quasi-experimental studies and research that shed light on how curricula are used in schools when implemented at scale.

Setting aside the questionable criteria codified by the tiers, we are laughably far from having the research to enable evidence-based curriculum decisions. The decade since ESSA's enactment has produced vanishingly little evidence on which curricula truly work best. A recent review by one of us (Jake) of 36 publicly available studies of the effects of core K–5 English language arts curricula on student achievement found pervasive methodological problems: Five studies simply examined student scores before and after the curriculum's adoption without comparing students to a control group. Sixteen used statistical methods that failed to take into account that students experiencing the curriculum were grouped within classrooms or schools. Others exhibited clear evidence of selection bias or puzzling exclusions of data. In the end, we concluded that we could trust only a handful of these studies, only one of which included data from more than one school year.

Furthermore, not one of the studies in Jake's review included a comparison curriculum identified by name. All comparison groups were described as using unspecified "business-as-usual" curricula or a composite of multiple curricula. This omission may be driven in part by publishers, when evaluating their materials, attempting to avoid the impression they are "calling out" competitors' products. But the comparison group matters—districts adopting curricula are almost always choosing among a small set of options, and they want to know which of these curricula boosts student learning the most. A generic comparison group does not help answer this question.



EVAN VUCCI / ASSOCIATED PRESS

President Barack Obama signs ESSA into law in 2015. Try as it might have to assess curriculum quality with its evidence-based tiered system, the law stopped short of guiding schools to determine what helps student learning.

What Might Have Been

In response to ESSA's passage a decade ago and the growing push for effective curriculum materials, states could have taken the lead. They could have begun collecting and sharing data on school district textbook adoptions. They could have conducted evaluations based on these data or partnered with trusted external researchers to do so. (There is no shortage of qualified researchers desperate for useful data.) Researchers have long been advocating for this approach, sometimes even spelling out the technical and logistical issues that states would need to overcome in order to harness these data to drive change. There are real obstacles to gathering these data and maintaining them over time, but states today collect all kinds of data and surely could devise systems to track changes in curriculum materials with reasonable efficiency.

But state officials didn't really do any of this—certainly not our home state of California. A decade after Morgan began critiquing the state's curriculum data reporting, which was required as part of a court settlement about inequitable access to curriculum materials, Jake found the same problems still exist. It took hundreds of hours to assemble curriculum data and make it even close to usable for a planned evaluation of literacy materials. Neither have other states taken the necessary steps to generate evidence on curriculum effects that could inform district decisions. Yes, some states have started collecting data on curriculum adoptions and making it available, but they haven't advanced to using those data for evaluations.

Given that states have not invested in collecting and learning from curriculum data, publishers have stepped in to fill the void. Currently only some states, including California, Florida, and Texas, require impact evidence. Publishers need to sell books to stay in business, and these most populous U.S. states hold the greatest market share. To get on these state- or district-approved lists, publishers have little choice but to perform the evaluations to meet the evidentiary standards in use by the field. Conducting evaluations—especially experimental studies, but even those using quasi-experimental methods and secondary data—takes time and money. Funders are hardly leaping at the chance to bankroll dozens of curriculum evaluations, so publishers typically pay for studies themselves. They might do the evaluation work internally or hire a contractor to carry it out (for a veneer of independence). Either way, this situation creates a conflict of interest and an incentive to do the minimum amount of work required to produce positive findings to bring the product to market.

Ideals, Meet Reality

In short, because states didn't take on this data-collection and evidence-building work, yet some large-market states require it, publishers are performing these tasks. It's fair game to scrutinize their efforts—and they could certainly invest more in careful research both to improve their products and to demonstrate effectiveness. But many obstacles impede that vision and underscore the need for more independent, centralized curriculum evaluations, ideally led by states or foundations.

Unrealistic timelines present one set of obstacles. Relatively few states require that curriculum materials meet ESSA evidence standards for statewide approval, but the number is not zero. And, of course, districts are making purchases all the time, often corresponding to their states' curriculum adoption cycles, so publishers need to produce evidence quickly. But how can they evaluate their materials if they can't sell them to schools and districts in order to conduct an evaluation? Often curriculum adoption decisions come

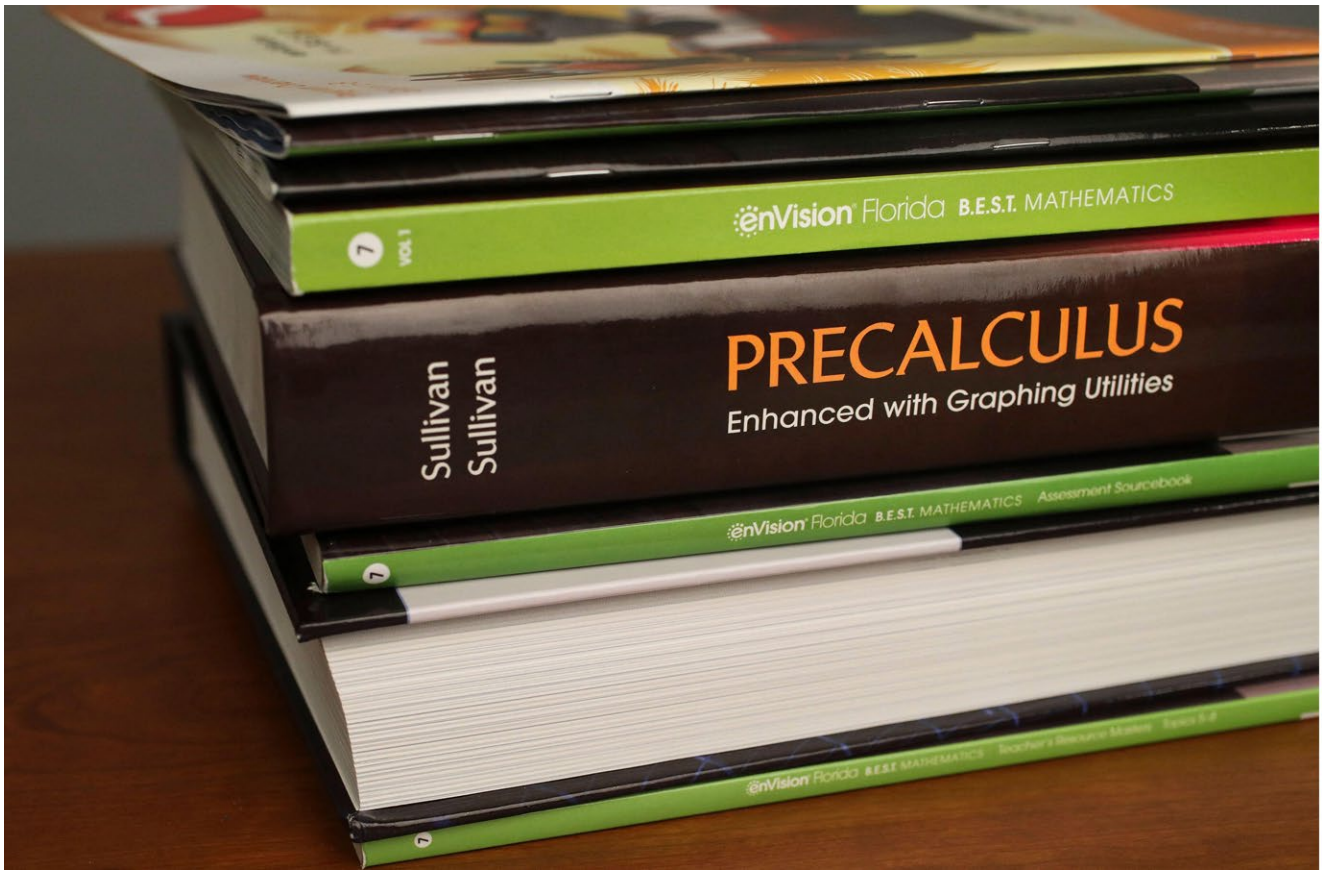
shortly after standards revisions, but companies need time to adapt their materials to new standards. Add in the time needed to do a serious evaluation, and the process can take several years. And that's if researchers are only looking at short-term outcomes like test score gains over a single year—outcomes that may be misleading about a curriculum's long-term effects.

Logistics, especially the challenge of recruiting participants, also present challenges. To conduct an impact evaluation that meets current Tier 1 standards, publishers would need to get some set of schools or districts to agree to random assignment of curriculum materials, with researchers deciding by chance which schools or districts would use a given curriculum and which would not. Realistically, this is hard: Districts aren't exactly champing at the bit to have their core instructional materials randomly assigned. Probably the best case is that publishers could convince an existing district customer to allow some subset of their schools to undergo random assignment. In that case, though, the evaluation would compare the company's new curriculum with its older one. But because companies want to sell both products, that study design is guaranteed to produce results that won't benefit the company. It's hardly good marketing practice to present a conclusion like "Our new curriculum is 0.25 standard deviations more effective than the one we've been selling for a decade."

Questions of validity. Studies relying on random assignment or analyzing early adoption of curriculum materials may well be sacrificing the generalizability of their results at the altar of causal surety. That is, the kinds of schools and districts likely to allow random assignment or to be early adopters may differ systematically from those that are not, so the results produced may not represent what would occur if the materials were more widely adopted. And random assignment studies may well entail ideal implementation conditions that don't mirror most school environments. For instance, the publisher may provide high-quality professional learning directly to educators in participating districts, in contrast to the more typical conditions in more naturalistic, quasi-experimental studies of districts that *choose* to adopt a particular set of materials.

Measuring student outcomes. Evaluation of curriculum materials requires some measure of student outcomes, usually the state test. But state tests often lag behind the development of new state learning standards, or, as has been true in the past, they may fail to accurately assess the full breadth of content prescribed by the standards. Beyond this, state tests infamously struggle with their sensitivity to instruction, and different tests correlate differently with measures of instructional quality. Why wouldn't we expect the same variability to show up in curriculum effectiveness studies? What's more, many children are performing far below grade level, but state tests are required to assess grade-level standards, potentially obscuring meaningful student growth. We certainly can't expect that districts participating in curriculum-evaluation studies will also allow for additional assessment using a neutral external test. If they were to do so, however, and if the results based on different tests were to disagree, how would we determine which set of results is right?

The thorny issue of implementation. Teacher implementation of curriculum materials varies widely, and some teachers even leave their materials unused on the shelf. Without careful measurement of implementation—which likely requires some combination of teacher surveys and careful analysis of metadata produced by digital curricula—the most we can say about a curriculum, under even the best-case scenario, is how its average effect compares to that of other curricula that might have been adopted. Building more robust infrastructure for curriculum-implementation research, though a tall order, is essential.



ZUMA PRESS INC / ALAMY LIVE NEWS

The biggest curriculum markets, including Florida, require evidence of learning impact, but states have not invested resources to collect that data, leaving evaluative work to the publishers trying to sell into those markets.

The replication crisis. Another obstacle is the so-called replication crisis, which refers to the phenomenon in which researchers seeking to reproduce a colleague's experiment often fail to obtain the same results. Researchers, regardless of who pays for their work, can knowingly or unknowingly shade their design and analysis in myriad ways that affect their results. Carrying out causal research involves many decision points, and it's easy to imagine curriculum researchers filing away null or negative effects or adjusting study designs, just as scientists might in other fields. The adoption of open-science principles and more careful design would help address these issues, but this would again require independent leadership and a more centralized structure.

Cost barriers. Investing in the development of new curriculum materials is expensive and time-consuming. If a publisher has no guarantee that such investments will result in a demonstrably effective product, why would the publisher take the risk? Less cynically, we've argued that companies do want to improve their products, but such improvements probably require kinds of research—qualitative and implementation focused—not recognized by ESSA standards. The ESSA evidence tier requirements can therefore stifle innovation and risk-taking.

Providing the Evidence We Need

If policymakers and educators want good evidence of product effectiveness, the system needs to be

better calibrated to reality. This means many expectations we might naturally have simply don't make sense at this time. For instance:

- It's not reasonable to require strong causal evidence of effectiveness before individual districts or states make adoption decisions. Evaluations of instructional interventions, like curriculum materials and assessments, are not the same as clinical drug trials. Districts must base their adoption decisions primarily on the content and form of new materials and their alignment to whatever principles or frameworks (for instance, "knowledge richness") the state thinks should guide curriculum—not their impacts on student achievement. Evidence on the quality of instructional supports, like professional learning offerings, could also be relevant. Evaluating materials against a multitude of different frameworks is a tall order, so states should prioritize the most essential elements rather than create endless laundry lists each publisher must meet. As researchers generate well-conducted impact evaluations, those results can filter into and inform subsequent rounds of adoption decisions—but they can't be the sole driver.
- It's unreasonable to expect that curriculum evaluations be conducted by independent researchers unless states require it and create a robust infrastructure to fund large-scale independent evaluations and generate the data needed to do them well. Only states or philanthropic organizations can provide that infrastructure and financial support.
- It's not reasonable to privilege small, random assignment studies over larger, well-designed quasi-experimental studies. Nor is it reasonable to expect that all curriculum materials can meet Tier 1 standards through random assignment studies, given both the difficulty of obtaining participants for such studies and the concerns about generalizability that might result. Well-designed, large-scale evaluations that meet Tier 2 standards are more appropriate and feasible.

These caveats do not mean that curriculum evaluation can't be done—and done well. But they do mean we need a new model. What would such a model look like?

Any better approach must start with the collection of better data on what materials are being used where. Far too few states even attempt to track curriculum adoption (and none seek to understand curriculum use), and the data collected by states that do track it are woefully inadequate. Every state should require annual reporting on the core curriculum materials in use at each school. Even better, they should move beyond core materials to collect data on formally adopted interim assessment systems, supplemental curriculum materials, and other curriculum-related interventions.

Further, states need to go beyond data collection to ensuring the development of a robust infrastructure to conduct ongoing analyses of these data. These methods can include quasi-experimental investigations and simpler designs that compare value added to student achievement across schools adopting different curriculum materials.

States could also foster the development of robust systems for understanding *implementation* of curriculum materials at the classroom level. This would likely entail conducting teacher surveys or requiring teachers to log aspects of their instruction; these efforts should be carefully designed to minimize burden on teachers, maximize the quality of data obtained, and answer the most important questions about

curriculum use and supplementation. Much can also be learned about implementation through the analysis of curriculum-related metadata from apps, programs, and websites.

Finally, states should embrace other rigorous forms of evaluation, including studies emphasizing A/B testing within individual curriculum systems, which could be used to identify the impact of new features or approaches. Options also include larger-scale and more traditional impact evaluations of major new innovations, which would very likely have to be externally supported to be feasible. Because conducting these evaluations should not be the sole province of the curriculum companies, states or private foundations should allocate funds for this ongoing evaluation research.

We believe strongly that curriculum materials play a vital role in K–12 classrooms and that we need to amass a better evidence base on their learning impacts. We have been arguing this for a decade. To stay grounded in what's possible, we need to build an approach to evaluation that reflects the realities of schools and the incentives of the various actors in the ecosystem—from educators and leaders to researchers and publishers. Deploying some of the ideas proposed here, these groups, working together, can advance the agenda of getting better materials into every classroom.



ASSOCIATED PRESS

Morgan Polikoff is a professor at the USC Rossier School of Education and a faculty director of the USC Ed-Policy Hub. Jake Rowley is a PhD candidate concentrating in K–12 education policy at the USC Rossier School of Education.

(Disclosure: Morgan Polikoff serves on a research advisory committee for Curriculum Associates, for which he provides advice on the conduct and dissemination of the company's internal research. He also consults on similar advisory boards for other curriculum and assessment companies, and he conducts grant-funded research in partnership with curriculum companies.)

This article appeared at EducationNext.org on November 12, 2026.