William T. Grant
FOUNDATION

Evidence at the Crossroads

A BLOG SERIES FROM THE WILLIAM T. GRANT FOUNDATION

Read all posts online at: wtgrantfoundation.org/tag/evidence-at-the-crossroads

Contents

EVIDENCE AT THE CROSSROADS, PART 1

What Works, Tiered Evidence, and the Future of Evidence-based Policy

VIVIAN TSENG | OCTOBER 27, 2015 | HTTP://BIT.LY/1WLYKFR

As a lifelong science geek, I've always thought research was fascinating, but I never thought it would inspire much political interest. Current events suggest that I may have been wrong.

Research evidence is increasingly at the center of political and policy debates, and much of the federal focus on building and using research evidence has embraced a "What Works" agenda. Even as Congress and the Obama administration are debating various initiatives to use research evidence of what works in administering federal programs, the administration is encouraging executive departments and agencies to apply behavioral science insights in their work. Advocates are proposing legislative language for defining "evidence-based" in the reauthorization of the Elementary and Secondary Education Act. And, in a rare alliance, Representative Paul Ryan (R-WI) and Senator Patty Murray (D-WA) have partnered to introduce the Evidence-based Policymaking Commission Act, which, if approved by the Senate, would identify ways that data and research can improve public policy.

How did we get here?

Numerous forces over the past 15 years have shaped both the supply and demand for What Works evidence. On the supply side, agency leaders, advocates, and researchers sought increased rigor in evaluating the effectiveness of social interventions. In leading the Institute of Education Sciences from 2002-2006, for example, Russ Whitehurst promoted standards of evidence that privileged the use of randomized controlled trials to evaluate the effectiveness of programs and practices in education. Similar evidence standards were promoted in child welfare, mental health, criminal justice, and youth development. At the same time, Congressional set-asides and agency investments expanded the funding for rigorous evaluations, and What Works clearinghouses were launched to disseminate the evaluation evidence in education, criminal justice, child welfare, mental health, and substance abuse.

On the demand side, both the Bush and Obama administrations sought to bolster the use of What Works evidence in executive branch agencies. "Spending more on what works and less on what doesn't" became a catchphrase for policy initiatives that directed public dollars to programs with evidence of effectiveness, preferably generated from randomized-controlled trials. The financial crisis of 2008, and the period of austerity that followed, allowed policymakers to attach stronger incentives to What Works evidence. Since the recession, for instance, the federal government has invested over six billion dollars in "tiered evidence" grantmaking initiatives. In a tiered evidence design, interventions with more rigorous evidence of impact are eligible for the most funding, while interventions with less rigorous or emerging evidence are eligible for smaller grants, with the intention that interventions will ascend tiers as they demonstrate impacts. Federal tiered evidence initiatives include the Investing in Innovation Program, Social Innovation

Fund, Evidence-based Home Visitation Program, Evidence-based Teen Pregnancy Prevention Program, Workforce Innovation Fund, Trade Adjustment Assistance Community College and Career Training Grants Program, and the First in the World Initiative.

Where are we going?

Now is an opportune time to take stock of these tiered evidence initiatives and the broader What Works agenda. Much-anticipated evaluation findings from the initiatives will be released over the next year, and a new administration and Congress will need to decide whether to continue these initiatives and in what form. We are approaching a crossroads. In navigating through it, policymakers and researchers will need to confront several issues:

1. **What counts as evidence?** The most contested issue is what is considered “good” evidence. What Works evidence—defined as evaluation findings gleaned from randomized controlled trials—is currently at the top of the evidence hierarchy. Stronger research evidence, more confidence in the findings—what could be wrong with that? In research, as in life, things are more complicated. Evidence on What Works typically reflects the average impact of an intervention and its effects in the places where it was evaluated. Local decision makers say that that evidence is only somewhat useful. They don’t just want to know whether an intervention works on average, or somewhere else. They need to determine whether a program, curricula, or professional development strategy will work in their local context and for their communities. In essence, they want to know what will work for whom and under what conditions.They also want strong evidence that can inform their implementation decisions. How much staffing is required? How much training? How should resources be allocated? What alignment is needed with existing programs? While rigorous What Works evidence may inspire confidence among the federal policymakers and agencies who invest in programs, knowledge about context and implementation is critical for making What Works work in local communities. Yet building rigorous evidence on context and implementation has received remarkably little attention.
2. **What about improvement?** The federal evidence agenda has maintained a laser beam focus on using What Works evidence to make funding decisions. But with some analysts indicating that as many as 90% of RCTs generate weak or no positive effects, we face a dilemma. Will enough programs pass the bar? What happens to those that do not? And what should we make of evidence that is mixed or inconclusive? Rather than being used for a simple thumbs-up or thumbs-down decision, evidence could be harnessed to support improvement goals. To get there, however, we will need to determine what an improvement agenda would look like and how the federal government can support improvement.
3. **What about systems?** While the federal government has focused largely on evidence-based programs, the leaders of state and local human service and education agencies are more often concerned with systems. A narrow focus on evidence-based programs encourages people to run after silver bullet solutions that are not necessarily aligned with the myriad other interventions that they are running. Moreover, little support exists to ensure that the evidence-based programs that are adopted fit the key problems systems confront, and don’t create unintended negative consequences.
4. **What is the federal role?** What about states, localities, practitioners? Finally, while making remarkable strides, the What Works movement in the US has taken a largely top-down approach. The role for states, localities, and practitioners is not well articulated, and the existing approach has caused a not-insignificant amount of tension with those who feel at the mercy of decision makers in Washington DC. Teachers and service providers on the front lines often feel that evidence is something done “to” them, and here they have a point. The past 15 years have not created a

meaningful role for practitioners in building evidence agendas. Instead evidence agendas have been largely under the province of policymakers and researchers. As we look to the future, we will need to consider the federal government's role in building and using evidence. If there remains a significant federal role—which we believe there should—then we need to ensure that future initiatives better meet the needs of states, localities, and practitioners.

What is the future of evidence-based policy?

We are at a crossroads in evidence-based policy. Federal evidence initiatives can be strengthened, but doing so requires the will and the patience to learn from the work thus far. Otherwise, evidence-based policy will likely recede into the background as yet another policy fad that came and went. To move forward, let's take a good hard look at the current evidence initiatives and identify what can be learned from them. We will need to come to terms with outsized expectations, develop ways to improve programs and systems, and determine how the federal evidence agenda can better align with state, local, and practice needs.

EVIDENCE AT THE CROSSROADS, PART 2

Moneyball for Education

FREDERICK M. HESS AND BETHANY LITTLE | NOVEMBER 3, 2015 | HTTP://BIT.LY/20XWEPI

Earlier this year, we made the bipartisan case for why and how federal education policymakers need to start playing “Moneyball.” By adopting and adapting the Oakland Athletics’ pioneering approach in baseball of making decisions informed by data—rather than hunches, biases, and “the way we’ve always done things”—we can get better returns on our federal education investments and better outcomes for students.

Specifically, playing Moneyball for Education would mean:

- Collecting better, more useful data and building evidence about how well programs and policies work;
- using evidence to improve practice and inform policies; and
- shifting funds toward those things that deliver more promising results.

Sounds easy, right? Not so fast. Like baseball, Moneyball is actually a game of nuance, a concept that is rarely synonymous with federal policymaking. Too often policy is made absent the data that would allow it to be more effective.

Luckily some important groundwork has been laid to allow Moneyball for Education to begin to take hold. For example, Congress has funded several tiered evidence initiatives that offer a way to explore the possibility of rewarding evidence of performance and building the body of evidence in the field. These could help make some important shifts to better policymaking, if the initiatives themselves are used as sources of data, evidence, and models for continuous improvement.

As we grapple with the information emerging from these initiatives, it is important to keep in mind both the possibilities and the limits of the Moneyball approach.

First, we don’t yet have all the advanced metrics needed to play the game really well. By necessity, the coming evaluations of the tiered evidence initiatives in education will focus largely on what is readily measured: reading and math scores and graduation rates. Those results should be valued and used, but our reactions to them should be tempered by acknowledging that there are limits to the current state of measurement. Simply put, we don’t yet know how to measure some of our most sought-after student outcomes, like critical thinking and collaborative problem solving. That limits what can be known about the true effectiveness of some interventions. Accordingly, just as baseball now benefits from a wealth of recently-developed advanced metrics, education needs a broader set of indicators that lead to improved student outcomes. The federal government (via the Institute of Education Sciences) could help, but the entire educational ecosystem, including non-governmental entities, should play a role in identifying the right metrics and developing and refining them over time.

Second, it’s as important to learn from a strikeout as from a home run. In the rush to celebrate winners, important lessons to be learned from mistakes, missteps, and disappointments are often overlooked.

Accordingly, all evaluation results should be studied closely, including those finding no, mixed, and even negative effects. Learning from failure is one of the most important drivers of continuous improvement and an essential tool for creating a winning team—and better schools. Yet, too often the federal government doesn't make the investment in evaluation that is necessary to provide insight into how programs could be improved. A small but significant percentage of all program funds should be set aside for performing the high-quality program evaluations needed for us to even be in a position to learn from all of our at-bats.

Third, there are different expectations of veterans and rookies. While some slack should be given to the young rookie with unproven potential, veterans should be held to higher standards of performance. After all, the veterans are expected to continue producing results and often cost more given their track record of success. In other words, just as tiered evidence programs differentiate the available funding based on tiers of supporting evidence, so too should they differentiate the perspective through which they review the grantees' evaluations. Because it is vital that we encourage creative new solutions to persistent problems, we should accept as part of building the knowledge base the notion that we will spend some dollars on grants with promise that ultimately don't pan out. However, when we are investing to scale up a proven solution, we should hold it to higher expectations. Merely saying that a lot was learned from scale up grants is insufficient. If scale up grants do not produce the desired outcomes, that may have implications for policy design (e.g., perhaps the evidence bar was set too low to merit an investment at scale, or perhaps we weren't able to fully measure the benefits of the program given limited metrics). The ability to differentiate like this is one big reason that the tiered funding framework holds a lot of promise.

Fourth, learn to field the best team possible within the salary cap. A baseball general manager would be run out of town if he signed new players to contracts without considering the team's overall payroll and without analyzing whether a player's salary reflected his production. Yet in education, we typically focus just on program outcomes without paying much attention to the costs of producing them. It's time to get smarter about measuring the return on investment through cost-benefit analyses. To do so, more precise, transparent, and common cost accounting rules are ultimately needed, but in the meantime the available cost data should be considered in making sense of the upcoming evaluations.

Fifth, don't judge the whole season on the basis of one game. There is much we can learn from these evaluations, but we must resist the urge to make overly broad judgments on the basis of one study. This caution applies somewhat to how we view the effectiveness of discrete interventions, but more so to how we assess the impact of a big federal program that funds states and providers that vary in their activities and their quality. The complexity of federal programs means it is often difficult to determine whether the funding stream itself is effective—or to reach pat conclusions about which programs do or don't "work." In other words, as we prepare for the coming season, we should avoid the tendency to oversimplify, and instead seek to answer questions like, "What adjustments would bring about better results next time?"

Our paper explains more of our thoughts about how to play Moneyball in the context of federal education policy, including a set of seven specific policy recommendations for Congress and the executive branch. These recommendations, however, do not expect the federal government to play Moneyball by itself. Rather, they form a playbook for how the government can help nurture an ecology of information, institutions, and incentives that will make it easier for everyone involved in education to play this game well.

Now that's a winning formula.

EVIDENCE AT THE CROSSROADS, PART 3

Research-Practice Partnerships, the Future of the Evidence Movement

GORDON L. BERLIN AND REKHA BALU | NOVEMBER 10, 2015 | HTTP://BIT.LY/1RQHZ1C

One of the most unheralded aspects of the recent federal investments in scaling evidence-based programs is the novel partnerships formed between researchers and program developers. Both the U.S. Department of Education's Investing in Innovation (i3) grants and the Corporation for National and Community Service's Social Innovation Fund (SIF), for example, supported expansions that have more than doubled the participants served in funded programs, while also funding evaluations involving new partnership between researchers and program developers. These alliances have provided new ways for researchers and program staff to work together and have supported programs in their efforts to continuously improve and refine their interventions.

In both the SIF and i3, MDRC worked in concert with program leaders to think about how to use evidence to improve or even re-design interventions. We sometimes learned as much from disappointing findings as we did from programs that "worked."

BELL (Building Educated Leaders for Life), a recipient of a SIF grant, runs a successful elementary school program and wanted to test its application in a middle school setting. MDRC's evaluation of their summer program for middle-schoolers found encouraging results for math but not reading. Both organizations then launched a joint diagnostic process to identify possible programmatic solutions for middle school. As Tiffany Gueye, the Executive Director of BELL noted in an *Education Week* piece, BELL leaders "plugged findings from the study into our continuous-assessment process" and decided to focus enhancement efforts on staff training, curriculum, and student assessment, as well as a greater focus on social and emotional learning. MDRC and BELL are continuing to work together to learn how these efforts are playing out in the programs and to design a learning agenda for the future.

The story was a little different for the Center for Employment Opportunities (CEO), another SIF grantee. An earlier MDRC study of CEO's transitional employment program in New York City for recently released offenders re-entering the community found that, although the program did not yield long-term employment effects, it did reduce recidivism among program participants. Using what it learned from the evaluation, CEO restructured its employment programming to place more emphasis on post-program employment retention services as it expanded to other jurisdictions. MDRC is working with CEO in an initiative to pilot a cognitive behavioral education component with the hope that it will further improve employment outcomes, particularly for younger offenders.

This type of collaborative approach to building reliable evidence also can be applied to improving systems, not just specific programs or interventions. Outside of the SIF or i3 structure, MDRC is working with New Visions for Public Schools, which supports a network of 77 New York City public schools, to promote evidence-based practices while embedding ongoing evaluation and evidence-building into the system of support it offers schools. In a constantly changing operational environment, a support organization or system of schools finds as much benefit from using evidence to build systems for identifying and supporting students at risk as from introducing specific evidence-based interventions. This next generation of research–practice partnerships is employing a mix of advanced analytic techniques—including using predictive analytics, exploiting variation across schools, and conducting quick turnaround experiments—to inform decision making and improve school and student performance.

As the evidence movement matures, it is increasingly clear that we need to build on lessons not only from clear successes, but also from interventions that have not worked. Neither program developers nor researchers can tackle this task in isolation. Research–practice partnerships are a strategic way to ensure we learn from what does and does not work as we actively translate evidence into practice. These next generation alliances between social policy programs and evaluators also have the potential to tackle more complex challenges by capitalizing on the availability of the data housed in public systems and the opportunity to embed random assignment and other rigorous methods inside organizations' continuous improvement cycles, heralding a new age in evidence-based practice.

The Obama Behavioral Insights Team, an Important Addition to Evidence-based Policy

RON HASKINS | NOVEMBER 17, 2015 | HTTP://BIT.LY/21UZ76A

Although there are numerous signs at both the federal and state levels that evidence-based policy is influencing policymakers, a lurking danger in any movement is that it will fall short of its potential unless it continues to expand. Concern that the evidence-based policy movement is merely a flash in the pan should be at least somewhat assuaged by the breadth the movement has already achieved.

Perhaps the most far-reaching element of the movement are the six evidence-based initiatives started by the Obama administration, which are now supporting well over 1,400 programs at the local level in teen pregnancy prevention, home visiting, preschool and k-12 education, community-based programs, and employment and training. The Teen Pregnancy Prevention Initiative and the Maternal, Infant and Early Childhood Home Visiting Program are especially notable. These two initiatives demonstrate several of the characteristics that seem certain to become classic features of evidence-based policy. Notably, both illustrate the advantages of tiered funding initiatives in which most of the federal grant funds (75 percent in both cases) go toward programs with strong evidence of producing impacts on important outcomes, while smaller grants go toward promising and innovative programs that are comparatively untested. This arrangement ensures that most funds are spent on programs that maximize the chances of producing impacts, while simultaneously providing adequate funds to develop new programs that may prove their worth in the future. The pipeline to innovation must remain open because new and more effective programs are always needed.

But the distinct approaches that are sustaining the evidence-based movement do not end with the Obama initiatives. Joining the parade are the reforms of Head Start, based in large part on observational measures of teacher performance in the classroom; the Pew-MacArthur Results First initiative, in which 19 states have agreed to review the evidence on which their social programs rest and, based on this assessment, make recommendations to their respective state legislatures about shifting funds to programs that work; the recent focus on Pay for Success programs (also called "Social Impact Bonds); and others.

To this impressive list we can now add the recent work by the White House's Social and Behavioral Sciences Team (SBST). Founded just last year, and under the leadership of wunderkind Maya Shankar, the team has just published the results of 17 studies, all but one based on random-assignment designs, to influence people's choices or improve government efficiency. Based in large part on the work of Richard Thaler of the University of Chicago and Cass Sunstein of Harvard, SBST is teaching government to "nudge" people into making better decisions.

The team is using the results of behavioral research which, roughly speaking, shows subtle ways that various messages can influence what people do. If people are reminded of obligations in timely fashion, they are more likely to fulfill those obligations. When filling out a form, if people sign at the beginning stating they will provide honest and accurate information in their answers, they are likely to give more accurate answers than if they sign at the end when they have already provided their answers. When employees specifying withholdings are defaulted to depositing withheld funds in a retirement savings account or taking an action to opt out, rather than having to take an action to opt in, the enrollment rate will increase and people will save more for retirement.

There are several reasons the SBST behavioral science initiative is such a stellar entry into the flooding tributaries of the evidence-based movement. First, the energy of the evidence-based movement must be rekindled by new initiatives that strengthen the claim that evidence-based policy can improve program impacts. Second, the findings from all but one of the administration's studies are based on random assignment designs, the gold standard of program evaluation. It follows that their results are likely to be reliable and replicable. Third, all the studies have immediate application to government policy. We learn how to write letters that increase compliance, ways to increase contributions to retirement plans, how to increase the likelihood that government workers will increase use of two-sided copying, how to increase college entry, how to increase the call-in rate for programs that need information from account holders, and many other lessons that improve individual choices and promote government efficiency. Fourth, it seems reasonable to conclude from the entirety of the SBST's results that if scaled up, their behavioral innovations would save the government billions of dollars. A more subtle and difficult to measure outcome of scaling up the SBST discoveries is the effects on health and on well-being in old age. It may be difficult to measure the benefits of millions of Americans saving more for retirement, but who doubts that a savings account is a key to peace of mind and greater choices in spending by retirees who have more in their savings account, not to mention the indirect benefits generated from less dependence of retirees on their children and other relatives.

The field of application of nudges based on behavioral science is limited only by the imagination of program designers and government officials. Equally important, the early success of applying behavioral principles to government policy illustrates yet again the soundness of basing policy choice on evidence of outcomes. Evidence-based policy is expanding its reach, this time by showing new ways to influence behavior, improve the efficiency of government programs, and save money. Evidence-based policy is on a roll.

EVIDENCE AT THE CROSSROADS, PART 5

Improving Implementation Research

BARBARA D. GOODSON | DECEMBER 1, 2015 | HTTP://BIT.LY/1VHXLLW

Over the past two decades, the education research and policy landscape has been shaped by the concept of using evidence to make policies or adopt practices. And not just any type of evidence, but high-quality evidence generated by "scientifically valid research." The methods for assessing intervention effectiveness are becoming a matter of greater consensus, thanks to the development of agreed-upon standards of evidence for evaluating the impact of interventions. But, in an environment where effective programs are often adopted and implemented in new contexts, is it time to focus on implementation research that can support this adoption?

A Focus on Evidence of Effectiveness

Today's definitions and standards of high-quality research and evidence are far more specific because of well-articulated systems for assessing the evidence generated by studies of intervention effectiveness. For instance, the What Works Clearinghouse (WWC), developed in 2002, seeks to inform decision making by rating evidence from individual studies and from a body of research on particular interventions or outcomes. Similar evidence rating systems now exist for other policy fields. The U.S. Department of Labor's Clearinghouse for Labor Evaluation and Research (CLEAR) and the GRADE rating system in health care are two examples of systems that assess the strength of evidence gathered from studies of program or policy effectiveness.

This focus on effectiveness continues to grow and sustain itself as more evidence is produced. The Obama Administration, for example, has embedded a strong evaluation focus into many new funding initiatives, such as the Investing in Innovation Fund (i3), which uses a tiered evidence framework to provide greater funding to programs with scientifically valid evidence of impact. If one of the goals of initiatives like i3 is to generate more rigorous evaluation of program effectiveness in a given field, this is certainly being accomplished. Over 90 percent of the 150 programs funded by i3 between 2010 and 2013—an unprecedented majority—are conducting impact evaluations with the potential of generating valid and credible evidence on effectiveness. This includes all of the larger grants and even a majority of smaller grants that were not required to conduct rigorous studies.

What About Evidence on Implementation?

More recently, though, evidence on the implementation of effective practices has received increased attention.

Decisions about which practices to adopt depend on a variety of considerations, including questions of implementation. Stakeholders need evidence that tells them about the feasibility of implementing the new practice in similar contexts or with certain populations, as well as data on the costs associated with adoption. These potential adopters also need explicit instructions on how to implement the practice, and how to provide necessary supports, such as training, coaching, or infrastructure.

If impact evidence is believed to lead to the dissemination and adoption of improved educational practices, then we need to think more systematically about how to support this adoption. We need to think about the types and quality of evidence on implementation of practices.

Concern now centers on whether standards of what constitutes high-quality implementation evidence may be developed to parallel the standards now commonly accepted for evidence on effects.

The Investing in Innovation Fund is playing a key role in promoting a focus on implementation by setting expectations that grantees conduct high-quality implementation studies that are intended to provide data to support replication and broader dissemination of effective practices. But whereas i3 uses WWC standards as the metric against which to measure the strength of impact evidence generated by grantees, it does not have a comparable set of standards to gauge for strength of implementation evidence.

The absence of commonly accepted standards for implementation studies led i3 to ask its national evaluator, Abt Associates Inc. and its partners, to propose a set of evidence standards for measurement of implementation for programs funded through the initiative. The standards that have been developed are based on a comprehensive logic model, a system for measuring the fidelity of implementation of the key inputs identified in the logic model, and annual assessment and reporting of the extent to which fidelity of implementation of the key inputs has been achieved.

In light of the i3 experience, and the increasing expectations among funders that high-quality implementation evaluations be conducted in concert with rigorous impact studies, the field needs to take the next step toward systematic standards for research evidence on implementation. Even if it is too early to think about standards such as those delineated by the WWC, is it time to develop guidance for researchers that would, at a minimum, lay out types of information on implementation that they should collect in different evaluation contexts (formative, efficacy, effectiveness, or scale-up studies, for example), implementation measures, and ways to analyze and report data on implementation. As part of the evolution of standards for implementation studies, we can provide signals to researchers about best practices in measuring implementation.

Ultimately, if our search for effective reforms for educational practice is successful, having strong and reliable evidence on implementation will be crucial for enacting real reform in our schools.

EVIDENCE AT THE CROSSROADS, PART 6

Evidence is Only as Good as What You Do With It

MARK LIPSEY | DECEMBER 10, 2015 | HTTP://BIT.LY/2OXYPMP

We are living in a time of unprecedented systematic research on the effectiveness of interventions that are intended to produce better outcomes. This level of effort is producing a substantial volume of intervention research, but a critical question is what to do with these studies?

The federal government has made massive investments to stimulate and support research on intervention effectiveness. One form of this that should not be neglected is the grant programs administered through the National Institutes of Health, the Institute of Education Sciences, the National Institute of Justice, and other such agencies that support field initiated intervention research in their respective topic areas. And one of the most notable developments, as has been well described in previous posts in this series, is the launch of a number of federal initiatives that provide tiered funding to support programs differentially according to the strength of supporting evidence. Most important for present purposes is the common requirement that the funded programs be evaluated, adding still further research to our stock of evidence about what works.

These initiatives will report a range of findings on a diversity of outcomes—some positive, some inconclusive, some possibly negative. How do we extract from these findings insights that can contribute to our understanding of what works and what doesn't? And how do we determine which findings have general applicability and thus identify programs and practices that could be effectively scaled up, and which ones are specific only to the situations and circumstances in which a particular study was done?

Drawing Conclusions

It will be tempting to try to interpret each study individually. If such a study shows positive effects, we might then conclude that the intervention it tested must be effective and we should promote its wider use. That would be a mistake. Any individual study has idiosyncrasies of method, circumstances, participants, intervention particulars, local support, and the like, along with a dose of happenstance, that make it uncertain whether the same results would occur if the study was replicated, and even more uncertain that those results would generalize to other settings, participants, etc. This is not to say that each study does not make an important contribution to knowledge. It does, but we must be careful about overinterpreting the implications of the findings of any one study.

If we are to draw conclusions with practical implications from intervention research, multiple studies are needed to provide a sufficiently broad base of evidence to support generalization beyond the idiosyncrasies of a single study. There is much talk of replication these days, and replication of the results of an intervention study is informative. But if what we want to know is how confident we can be that the findings generalize to other contexts, then multiple studies with realistic variations will be more informative than attempts at exact replication. Relatively consistent findings across different participant groups, implementation contexts, and variants of the intervention itself provide evidence that the intervention is robust to the kinds of variation likely to occur in actual practice. Inconsistent findings, on the other hand, provide clues to where the boundaries of generalization are—the circumstances under which the intervention may not work so well.

With multiple studies, of course, extracting and interpreting the information useful for practice and policy presents an even bigger challenge. One very intuitive approach would be to categorize the studies according to whether they find positive, null, or negative effects and then try to figure out what differentiates these groups. That too would be a mistake. By what criteria would we categorize the findings of each study? Individual studies universally characterize their effects according to their statistical significance, and it is that indication that is typically relied upon as an indication of whether a positive effect was found. That makes sense for a single study—it provides some protection against claiming effects that are likely to have occurred only by chance. But statistical significance is a joint function of the size of the sample and the magnitude of the intervention effect. When reviewing multiple studies, it is necessary to focus on the magnitude of the effects so that smaller studies that find large effects are not overlooked, and very large studies that find small effects are not over-interpreted.

Integrating Findings

As many readers will recognize, we have a well-developed procedure for integrating the findings of multiple intervention studies with a focus on the size of the effects rather than their statistical significance, and for exploring the characteristics of those studies that are associated with the effects they find. It is meta-analysis.

In meta-analysis, statistical metrics for the magnitude of the effects on each outcome (effect sizes) are systematically extracted from each study along with a profile of study characteristics related to methods, participant samples, intervention characteristics, implementation, setting, and the like. Analysis is then conducted to describe the distributions of effect sizes on different outcomes and, most important, to explore the relationships between the study and intervention characteristics and the nature and magnitude of the effects. The results, then, characterize the findings of a body of research, not just individual studies, and does so in a differentiated way that attempts to identify the factors associated with differential outcomes.

Meta-analysis of broader and narrower bodies of intervention research is quite common, usually undertaken by academic researchers and promoted by such organizations as the Campbell Collaboration and the Cochrane Collaboration. However, it has not been integrated into the plans of many of the government agencies sponsoring the various tiered-evidence initiatives that are underway as a method for integrating and interpreting the findings of the many studies those initiatives are producing. There are exceptions, and some indications of movement in that direction. The Office of Adolescent Health, for instance, is sponsoring a meta-analysis of the studies of teen pregnancy prevention it, and some of its sister agencies, have initiated. Similarly the Centers for Medicare and Medicaid Services is undertaking meta-analysis of the research done under its Health Care Innovation Awards initiative. Other agencies may be making similar efforts or contemplating them, but this perspective is not widespread.

Initiating a meta-analysis of the studies developed under any of these initiatives, however, is something best planned from the beginning, not decided after the studies are underway. To support the most informative meta-analysis, there are some advantageous features that can be built into the initiative at the start. For example, the distribution of studies is important. It will be more informative to have multiple studies on practical variations of each intervention in a selected set of interventions than one study on each of a set of distinctly different interventions.

Most important, however, is the opportunity to specify in advance the kind of detailed information that each local evaluator should collect and report in order to provide a rich set of variables for the meta-analysis to explore. These should include many particulars related to the implementation of the intervention, the participants who receive it, the organizational and service delivery context, and other such factors that are often not reported in sufficient detail to support the most informative meta-analysis. Agencies sponsoring (and paying for) multiple studies under a single initiative can require a level of consistency in the way those details are reported that is rarely attained when study authors are left on their own to decide what to report.

Guiding Practice

The promise of this approach is not simply that the most effective interventions will be identified in a systematic and methodologically credible way, though that will be one result. That form of knowledge allows dissemination of program and practice models expected to be effective if they are implemented with fidelity, that is, in the same way they were in the supporting research. Having a repertoire of such models is indeed a big step forward in the quest to find and scale up effective programs and practices. However, despite the evidence, there will be many reasons why providers and practitioners will not adopt those models and, if adopted, will do so with adaptations that change them from the original evidence-based version.

The larger promise from sufficient bodies of evidence and differentiated meta-analysis is identification of the principles that make the respective programs and practices effective. Knowing why something works, and not just that it works, provides an explanation—a theory if you will—that can guide effective practice in flexible ways amenable to local adaptations and practical constraints, so long as those variants preserve the underlying change mechanism that makes the intervention work.

We need a cookbook full of recipes for effective practice, but even better is knowing how to create recipes for effective practice from the ingredients on hand in the local kitchen.

EVIDENCE AT THE CROSSROADS, PART 7

Using Evidence To Do the Most Good, Even When it Reveals an “Inconvenient Truth”

PATRICK T. MCCARTHY | JANUARY 6, 2016 | HTTP://BIT.LY/1TBSOUP

Our supply of evidence about What Works, though still too small, is beginning to grow. I want to focus here on how we use evidence—how we put it to work to help more of our nation’s children and families achieve their full potential. Let’s focus on two important aspects of this question: building the bridge from programmatic evidence to population-level impact and reacting wisely to evidence that reveals an “inconvenient truth.”

Most evidence-building research zeroes in on a particular programmatic intervention and tells us whether it achieved desired results with a specific group at a given place and time. When we are lucky, we see those results replicated in different settings, giving us even firmer ground for saying that this intervention works. Such evidence can help policymakers and program managers make wise decisions about where to invest their scarce dollars so they will do the most good. That is very, very good, but if we stop there we leave unrealized a lot of the potential of good evidence to drive great outcomes for all of our nation’s children and families.

To get to population-level scale, we need to embed the best of What Works into the large public systems that serve not a few hundred children or families, but thousands upon thousands. In part, successfully embedding evidence-based programs in large systems requires them to be adapted to the way the systems work so that referral pathways are smooth, contracting provisions are appropriate and the program has secure access to a dedicated funding stream. A greater and ultimately more important challenge is to draw lessons from the evidence and apply those insights to what public systems do and how they do it. When that happens, those systems can achieve better outcomes for all of the children and families they serve.

To do this, we have to look across multiple evaluations to see patterns, clusters and commonalities. If, for instance, you look at the evidence from the multiple evaluations of Multi-Systemic Therapy, Functional Family Therapy, the Oregon Treatment Foster Care model, and similar programs, common success factors begin to emerge. They all place a lot of emphasis on training and supporting their staff to effectively engage families who often are in crisis and hostile to outsiders at the very point that staff arrive on the scene. There is no moment more critical to the outcome of an intervention than these early steps to create the conditions in which therapeutic work can succeed. And there is no one for whom getting those first relationship steps right is more critical—and more challenging—than public system workers who arrive on a family’s doorstep talking about taking one of their children away or locking up one of their teenagers. If we can isolate the critical success elements common to evidence-based programs and build those into the training and supervision of frontline workers, we should begin to see every family that encounters one of these systems having much better odds for success.

The second issue related to using the evidence we are accumulating is how we act on evidence that challenges prevailing paradigms, especially when such programs don't just not work, they actually do harm. The most egregious example we have as a country are youth prisons. We have known for decades and without a shadow of a doubt that these facilities, which contradict everything we know about adolescent development, cause harm to young people and create more rather than less crime in communities. Contrary to all available evidence, we continue to lock up two to three times more young people than any other nation. We are beginning to hear calls for change from more and more quarters, and we see political will and leadership emerging. To encourage these efforts and to spark even more, in a recent TEDx talk I committed the Casey Foundation to work with any state that makes the decision to do the reform work and make the investments necessary to successfully and safely close these factories of failure and replace them with programs and facilities that help young people get back on track.

Evidence doesn't turn itself into policy, especially when it contradicts prevailing paradigms or entrenched funding streams. If we are serious about a What Works movement, we can't allow ourselves or other decision makers to pick and choose which results we want to act upon. Philanthropy can help by investing in developing and testing alternatives, supporting evidence-driven advocacy, and underwriting technical assistance to evaluators and decision makers.

The conversation about What Works continues, thanks in part to other contributors in this series, including Ron Haskins and Gordon Berlin, as well as many others in the field. I hope this ongoing conversation includes attention to what happens as a result of the evidence. The futures of our nation's young people depend on it.

Building an Improvement Infrastructure

DONALD J. PEURACH | JANUARY 12, 2016 | HTTP://BIT.LY/1THFMGW

With month's congressional budget deal preserving level funding for the Investing in Innovation Fund (i3) and with the passage of the Every Student Succeeds Act (ESSA), through which i3 will be given new life as the Education Innovation and Research program, we are witnessing renewed investments in federal efforts to build and use evidence of What Works in education. Even so, funding for the Social Innovation Fund (the evidence standards of which were modeled on those of i3) will likely be reduced—not incidentally following reports on largely mixed evaluation results of grantee programs.

Now that we know these federal grantmaking initiatives will continue, how can we ensure that investments in innovation are maximized? What can the next generation of tiered evidence initiatives learn from the investments thus far? Several writers in this series have focused on program impacts. In this entry, I argue for building a complementary infrastructure focused on program improvement.

The Challenges and Learning Needs of Practicing Innovators

From 2011–2014, the William T. Grant and Spencer Foundations sponsored a learning community of program developers and practitioners funded through i3, a group I refer to as "practicing innovators." The i3 learning community offered grantees a venue to share their experiences and insights with each other. It also brought their perspectives to bear on policy conversations about incubating educational interventions. Because of my work studying the scale up of educational programs, I was invited to participate as a consultant in these meetings and was able to witness educational innovation from the perspectives of both policy and practice.

All told, the i3 program stands as a powerful example of the possibility of using policy to shape the work of practicing innovators. It also stands as an important example of the ways in which policy can introduce new challenges for practicing innovators. By design, i3 introduced structure by requiring that practitioners develop clear intervention designs, work timelines, and budgets. They also had to establish clear goals and evaluation criteria to discipline their work. At the same time, the i3 program stretched practicing innovators to develop a much broader array of capacities and skills. This is no surprise. Such challenges are endemic in large-scale educational innovation. For i3 grantees, the challenges arose in three key areas: 1) building new collaborations, 2) working with evaluators, and 3) leading and managing complex organizations.

1. Building new collaborations and negotiating fidelity-adaptation tensions

The i3 program has pressed grantees to collaborate with more (and more varied) districts and schools than they had in the past, often using interventions that had to be revised and extended in new ways. Scaling required managing issues of recruitment and retention while, at the same time, building open, trustful relationships that supported productive collaboration. For example, the IDEA Public Schools were awarded an i3 development grant to expand professional development to 600 new teachers, 400 instructional leaders, 24 new principals, and 160 aspiring teacher leaders to support the human-capital pipeline in the Rio Grande Valley. ASSET STEM Education, a validation grantee, proposed providing comprehensive professional development to teachers across Pennsylvania in K-6 standards-aligned STEM instruction. Reading Recovery, a scale-up project, proposed training 15 new Teacher Leaders and 3,750 new Reading Recovery teachers while also establishing new training sites in rural areas to target low performing schools and high needs students.

Scaling required balancing a deep tension between local adaptation and fidelity of implementation. On one hand, i3 grantees recognized a strong need to support local adaptation in order to manage uncertainty and increase effectiveness in fielding complex interventions across increasingly diverse arrays of schools. On the other hand, the i3 program placed a premium on fidelity of implementation in order to produce evidence of program impact, and local adaptation risked corrupting their evaluation designs.

The result was a breakdown in a fidelity-adaptation synergy that has driven continuous improvement among leading educational innovators. In the past, organizations such as Success for All and Reading Recovery (both i3 scale up grantees) navigated challenges arising through new collaborations using an approach to learning and improvement that had schools faithfully implementing practices that had been validated by research and experience while also adapting and extending these practices to address local needs and, then, propagating the new and promising practices in other sites. The i3 program design left grantees with a dilemma. Adapting their programs risked compromising their evaluations. However, not adapting them risked weak evidence of program impact as a result of continuing to implement dimensions of their programs that they recognized as problematic.

2. Working with evaluators

The i3 program also has brought innovators and evaluators into closer working relationships than has been typical in educational policy and reform. But because the two groups bring different priorities to projects, striking positive working relationships is no simple matter. For instance, innovators aim to produce replicable, effective interventions that are also responsive to local needs, opportunities, and problems. Doing so requires maintaining flexibility and adaptability in their programs. Evaluators aim to complete successful studies that yield unbiased and rigorous assessments of program effectiveness. Yet that, in turn, requires program stability and fidelity.

Crafting trustful, collaborative, and productive relationships required managing these competing priorities. Doing so was especially important in managing the tension between fidelity of implementation and local adaptation. As one grantee with a validation project explained:

> *“One challenge for us was teacher attrition. We had instances of teachers leaving their schools after the first year of the program. From our perspective, we needed to provide training to replacement teachers to maintain high-quality implementation in the school and to make sure that students continued to benefit. But that meant that a second year or third year school would have teachers who were only in the first year of implementation. Also, we were adapting our training every year as we learned from experience, including moving some of our online training to face-to-face coaching. So, these new, first year teachers would have slightly different training from the other teachers. But we knew that we were a validation project, and we had to make sure that these changes would be okay with our evaluators before making the final decision about how to proceed.”*

Managing the tensions required that innovators and evaluators learn to work together in new ways. Logic models and measurement models became the media through which they negotiated shared understandings and struck agreements on the nature and scope of their work. Early on, some evaluators worked in pseudo-coaching roles, as they guided innovators in clarifying the logic of their program designs. Practicing innovators were then free to adapt and improve their programs in ways that did not compromise the logic and measurement models around which their evaluations were structured. These logic and measurement models did not exist at the outset of the i3 program. Rather, they emerged and evolved over the course of the program as a product of the collaborative relationships between innovators and evaluators.

3. Leading and managing complex organizations

While most of the i3 grantees had some prior experience managing educational innovation, few, if any, were formally trained or fully prepared to manage the entire scope of work and the uncertainty they encountered while working within the i3 program and scaling their interventions. Rather, all of the members of the learning community found themselves managing networks, programs, organizations, relationships, and trade-offs that were in some way new to them. For example, extending and modifying intervention designs required increasing their development staffs. Further, working with more (and more varied) districts and schools required increasing their training/coaching staffs. Working with external evaluators and with U.S. Department of Education personnel required expanding their executive and managerial staffs. While critical to the success of their programs, much of this organization-building occurred “behind the scenes” and sometimes without support.

Absent established traditions of research and professional development focused on the practice of educational innovation, most leaders found themselves addressing their learning needs either through reflective practice or through participation in communities of practice. As one participant with a development grant explained of the i3LC experience:

> *“Especially in the early years of the grant, the i3 Learning Community was our most important resource for innovation implementation support. The opportunity to meet with like-minded and like-challenged colleagues in a place where there was no judgment was invaluable. We knew we could speak honestly, and there would always be thoughtful people who would understand and help us work through our decision making process and then reflect on the results.”*

Building an Improvement Infrastructure

While the i3 program has introduced more structure and discipline into the practice of educational innovation, so too has it introduced new challenges that will inhibit the program’s ultimate success. Developing system-level “improvement infrastructure” that supports practicing innovators in addressing these challenges could be a game changer.

After all, the structure and discipline introduced through the i3 program is very much a positive artifact of highly developed “impact infrastructure”: a system of political and policy supports emphasizing evidence as both an input to and output of the practice of developing effective, scalable educational innovations. The i3 program is one component of this impact infrastructure: a competitive grant program that structures requirements, resources, and incentives to support the use of evidence in innovation. The i3 program is supported by a web of interdependent federal policy initiatives supporting the use of evidence in innovation, including the establishment of the Institute for Education Sciences, the creation of the What Works Clearinghouse, and investment in the advancement of statistical methods and in early career professional development for researchers. It is further support by philanthropists advancing their own evidence-driven competitive grant programs, a population of private firms with capabilities for research and evaluation, and a powerful professional organization—the Society for Research on Educational Effectiveness—committed to advancing the cause.

This impact infrastructure, in turn, could provide a blueprint for building a complementary infrastructure focused on improvement.

Policy supports

A parallel improvement infrastructure could begin with a complementary policy web that promotes and supports continuous improvement. This policy web could include adapting the i3 program to support a sort of "novice–intermediate–expert" progression in creating capabilities for continuous improvement in sponsored projects. Support at the development level could focus on using design-based research to test and refine key practices and components. Further, support at the validation level could focus on developing capabilities in schools to enact evidence-driven Plan–Do–Study–Act cycles to adapt programs to address local needs. Support at the scale up level could focus on developing infrastructure linking the enterprise into a coherent, evolving learning system. This policy web could also include agencies to champion and monitor the work of continuous improvement, as well as initiatives aimed at investing in the development of formal methods of continuous improvement and in the development of researchers able to support the use of these methods.

Philanthropic, private, and professional supports

Already, elements of a complementary, private sector web are forming as key components of an improvement infrastructure: for example, the emergence of the Carnegie Foundation for the Advancement of Teaching and the SERP Institute as organizations championing the push for continuous improvement in educational innovation, matched by a community of researchers in universities and private organizations who are advancing and popularizing methods of design-based research. The growth and maturation of this private sector web could potentially be accelerated through federal resources and incentives that would draw in additional organizations, for instance by establishing requirements in competitive grant programs to incorporate formal methods of continuous learning and improvement. But it could also go further, sponsoring competitive grants programs that engage these organizations directly and provide resources and incentives to craft partnerships with practicing innovators.

Political supports

The bedrock of the impact infrastructure is political support anchored squarely in what Harvard University professor Jal Mehta aptly describes as the "allure of order": longstanding faith among policymakers in the potential to use principles of rational management to discipline otherwise "soft" educational practices. The allure of order has deep roots in norms of rationality that have long dominated educational politics and policymaking in the U.S., and it is reinforced by longstanding appeal to both business and medicine as sources of ideas and legitimacy to support educational reform.

Perhaps the biggest challenge to building an improvement infrastructure lies in extending political discourse to include an understanding of improvement that parallels the allure of order. Indeed, ideas are emerging in both business and medicine for doing so. Possibilities include the introduction of language emphasizing “infrastructure building” (rather than “turnaround”) as an approach to improving weak schools, the use of the tech sector’s notion of “perpetual beta” as characterizing the work of educational innovation (rather than the pursuit of “What Works”), and the use of Atul Gawande’s notion of “better” (rather than “scientific knowledge”) as an outcome of improvement-focused educational innovation.

The trick will lie in understanding how these notions have been used to shape and influence broader discourse and understanding in other sectors, and ultimately adapting these strategies to education and other social sectors.

Looking forward

There is no doubt that cultivating a highly developed, coordinated improvement infrastructure will be difficult. However, the emergence of the impact infrastructure provides evidence that it is possible.

Indeed, the development of an improvement infrastructure can actually be interpreted as the essence of innovation: working iteratively to make incremental improvements to well-reasoned, promising, yet inevitably imperfect strategies and plans. In this case, such incremental improvement would center on maintaining the positive benefits that have followed from a keen, sustained focus on impact while, at the same time, developing and sustaining an equally keen focus on continuous improvement.

Viewed from this perspective, balancing impact and improvement is not a matter of doing the impossible. Rather, it is a matter of duplicating success.

EVIDENCE AT THE CROSSROADS, PART 9

Promoting Evidence-based Teacher Preparation

EMERSON J. ELLIOTT | FEBRUARY 4, 2016 | HTTP://BIT.LY/1TBUDBW

In many of the posts in this series, evidence-based approaches to education are portrayed in the contexts of government agencies and research organizations. As Vivian Tseng has said, "the past 15 years have not created a meaningful role for practitioners in building evidence agendas."

Teacher education is one piece of the education landscape that illustrates how applying evidence-based strategies can make a powerful difference to the effectiveness of practitioners and educational services organizations.

Both data and research evidence can inform these strategies. For instance, data can help practitioners identify their level of performance and track changes over time. Research evidence can reveal factors associated with teaching quality and strategies to improve teaching, thereby helping practitioners modify teacher preparation in ways that are likely to improve instructional quality.

The Council for the Accreditation of Educator Preparation (CAEP), the nation's new education accreditator, provides an example of how aspirations for focused use of data, research, and continuous improvement can foster purposeful and effective preparation of teachers. As its first official act, in August 2013, the CAEP Board of Directors adopted standards for accreditation of teacher preparation. These were to employ the powerful leverage that an accreditation function can exert through challenging standards backed up by the rigor of strong and relevant evidence. CAEP accreditation actions, that is, would be "evidence informed."

A culture of evidence in the CAEP accreditation context

Since the standards were adopted, CAEP and the educator preparation providers (EPPs) have been transitioning to the new evidence-based framework, preparing for 2016, when all EPPs coming up for accreditation must employ it.

Three sources had particular influence on CAEP's approach to evidence-informed accreditation. The first is a regional accreditor, Western Association of Schools and Colleges, which calls on colleges and universities to use evidence in assessment, decision making, planning, resource allocation, and other institutional processes. A second is the Baldrige Education Criteria for Performance Excellence, criteria structured to help any educational institution achieve its goals and to improve its effectiveness through use of data. The criteria describe key attributes of performance and provide scoring rules to evaluate processes and results. And a third source is the compelling advocacy of the Carnegie Foundation for the Advancement of Teaching for improvement science: learning quickly, at low cost, and systematically using evidence from practice to improve it.

Here are some features of CAEP's interpretation of a culture of evidence in accreditation:

Data emphasize results

Accreditation data inform a diverse array of preparation practices and results. Some are familiar, such as GPA, licensure test scores, and clinical observation evaluations of candidates (i.e., college students preparing to teach). Some have not previously been expected as part of accreditation: for example evidence that clinical experience partnerships with schools and school districts are collaborative and mutually beneficial.

Three examples illustrate the new rigor in CAEP accreditation evidence:

1. One is a requirement to document effects that candidates have on P-12 student learning and development—both during pre-service clinical experiences and again after completers are on the job as teachers.

2. The second is annual progress monitoring. CAEP and individual EPPs will make data available annually on eight dashboard indicators of accomplishment.

- Four of these represent preparation outcomes: licensure rate, employment rate, employment in the field of preparation, and consumer information, such as initial salaries or places of employment
- Four more describe the effects of preparation after completers are employed: evidence that teachers have a positive impact on P-12 student learning, teacher instruction evaluations through observations and student perception surveys, employer satisfaction and teacher retention, and completer satisfaction with preparation.

These annual measures are one example of a shift in accreditation to a continuing process for gathering, interpreting, and using data—not just a procedure undertaken once each seven years and then set aside.

3. The third is data from admissions criteria and recruitment. Providers prepare recruitment plans, moving toward alignment of fields of preparation with changing employment opportunities for their completers (e.g., more science and math teachers, fewer elementary school teachers), and toward evidence that each year's class of candidates is academically able and diverse. EPPs monitor progress and modify plans as needed to meet candidate quality and employment goals.

Research probes more deeply

The new rigor in accreditation comes, in part, from evidence in the form of research and case studies. For instance, one of the CAEP accreditation pathways is structured so that an EPP conducts research on some major challenge in educator preparation, such as different models for clinical practice or recruitment and admissions policies. EPPs that choose this pathway follow standard research protocols to ensure validity that encompass literature reviews, appropriate study designs, data analysis, and interpretations of results. The work is to be of publishable quality.

Case studies will also be a frequent source of accreditation evidence. One example would be developing and testing new assessments, such as ones in which candidates demonstrate their abilities to teach through problem solving and critical thinking skills, or judging the effects of grit, perseverance, leadership or communication skills on candidate's instructional success with P-12 students.

Continuous improvement is the focus

These uses of data and research require that providers have capacity to gather, store, access, and analyze data and to interpret results. An EPP must maintain a quality assurance system that comprises valid data from multiple measures and builds capacity to disaggregate and inter-relate data so that analyses and interpretations of the findings can be conducted. Most important, the analyses and interpretations are to serve as the basis for continuous improvement—informing EPP judgments about how the courses and experiences offered to candidates can be made more effective.

CAEP has taken on responsibilities both to use data for improvement in its own efficacy and to make data better for the field. Unfortunately, the state of data in teacher preparation has been notoriously poor, characterized by very few common measures (primarily state licensure tests) and by an array of others that are developed uniquely in each EPP (e.g., clinical observation evaluations). These data characteristics make valid comparisons virtually impossible, so that there are no accepted norms or benchmarks for EPP performance. They are a significant challenge for evidence-based accreditation decisions. CAEP is seeking, in collaboration with states and EPPs, common data definitions and data gathering procedures for the 8 dashboard indicators as well as other aspects of preparation such as the characteristics of clinical experiences. And it has created a director of research and strategic data initiatives to focus these efforts and oversee an evaluation of the impact of CAEP's standards and data emphases as they unfold in the coming five to ten years.

Next steps

A culture of evidence that shapes the accreditation of educator preparation programs can have an enormous influence over the education landscape. Now comes the big test—will it work?

Success depends, ultimately, on the response of educator preparation providers. Will they perceive that evidence-based accreditation makes use of tools and capacities that are valuable for them? Will the investment to build and maintain these tools and capacities be undertaken and maintained?

CAEP's perspective is that evidence-based accreditation will encourage EPPs to build efficient oversight capabilities as well as capacity to modify preparation courses and experiences. These tools can ensure that those who graduate go on to become knowledgeable and skilled teachers, helping America's increasingly diverse P-12 students master challenging school curricula. And this is where evidence-based decision making can make a great difference in education outcomes.

The Need For Cost-Effectiveness Evidence in Education

HENRY M. LEVIN | FEBRUARY 11, 2016 | HTTP://BIT.LY/21UEJ05

Researchers are increasingly producing rigorous evidence on the effectiveness of specific educational reforms in order to improve teaching and learning. This research, which provides decision makers more accurate information on the probable impacts of education interventions, is at the heart of evidence-based decisions. But even the most rigorous studies of educational outcomes are incomplete if they do not consider the costs that must be covered to obtain those outcomes.

Issues of cost are always present in allocating resources efficiently, but they are particularly pressing when economic challenges arise. Consider that at least 31 states provided less state funding per student in the school year ending in 2014, several years after economic recovery, than they did in 2008. And local government funding for education fell in at least 18 states over the same period.

Many educational decision makers are challenged by severe economic constraints that restrict choices among alternatives for improving educational outcomes. For this reason, decisions must be premised not only on the promise of positive outcomes, but also on the costs for obtaining those outcomes. Simply being provided with effectiveness information, while not being informed of the costs for obtaining those effects, can be highly misleading and wasteful.

The costs of effectiveness

How serious are these concerns? Do costs vary enough across different interventions to have a powerful impact on cost efficiency? To answer these questions, the Institute of Education Sciences (IES) of the Department of Education provided funding to the Center for Benefit-Cost Studies in Education (CBCSE) at Teachers College, Columbia University, to compare the cost-effectiveness of dropout prevention studies and early literacy studies, using findings from the What Works Clearinghouse (WWC) of the U.S. Department of Education. The WWC uses systematic criteria to evaluate evidence on effectiveness, but does not consider costs.

Cost-effectiveness analysis enables comparison of the costs and effectiveness of the educational actions being considered, informing the decision maker of those alternatives that promise the lowest cost per unit of effectiveness. Employing a rigorous cost accounting method applied to the interventions in each group, the CBCSE obtained comparative cost-effectiveness results for the different dropout prevention strategies and, in a separate study, the early literacy strategies. Among dropout prevention approaches, there was a range of six to one in cost-effectiveness results, meaning that the most cost-effective approach could provide six times as many new graduates as the least. Among early literacy studies, the differences in costs per given increase in reading skills were typically three or four to one. By choosing among the most cost-effective programs for addressing dropout prevention or early literacy, schools could presumably save considerable resources to allocate to other important educational endeavors. Since no comparable cost information on alternatives was provided by WWC, decision makers were not informed about cost-effectiveness and had only the purported effectiveness results (irrespective of their costs).

Considering the true costs of outcomes

In most educational evaluations, discussion of costs is completely absent from consideration. The tacit message is that they don't matter, or that differences are likely to be trivial. But even in rare cases where some cost information is reported, it is usually erroneous because it is not the product of an acceptable cost method. To paraphrase Lee Shulman's famous phrase on comparative measurement, effectiveness is measured with calipers, and costs are measured with a witching rod. Most evaluators lack background in cost measurement and rely on easily accessed information, such as crude budgetary reporting or an "estimate" from a contract administrator. School budgetary documents simply list the allocation of financial resources among different spending categories—they were not designed to provide cost accounting for specific programs or interventions.

The true costs of an intervention represent the value of all of the resources required to obtain the outcomes found in a valid effectiveness study. Some resources may be reflected in a specific project budget, but others may be financed from other sources, such as reallocations from other school programs, or resources in kind, such as the time of volunteers. The overall cost of an intervention is determined by the required personnel, facilities, equipment, technical assistance, and other ingredients regardless of their source. The value of all of the resources used for the intervention must be acknowledged and included in costs. The most widely accepted method of cost-accounting for educational interventions is the ingredients method, which, as its name implies, identifies the ingredients that were used to obtain a particular evaluation outcome. The value of the ingredients is determined by market prices or some equivalent.

Strengthening the field

The ingredients method has been used in educational cost research for four decades and has been continually refined over that period.[1] Examples of educational studies that document the method and its application include cost-effectiveness studies of class size reduction, computer-assisted instruction, increases in length of the school day, peer tutoring, high school completion, early literacy, and programs to increase graduation rates in post-secondary education. However, reliable cost-effectiveness information is still rare in the overall evaluation literature, and cost-effectiveness studies merit much wider availability and dissemination to assist decision makers in making more efficient use of resources.[2]

Where cost analysis has been used to evaluate educational reforms, the ingredients method has been favored because of its reliance on the economic concept of opportunity cost and its validity as a cost-accounting method. Unfortunately, a major obstacle to its use is that few educational evaluators have familiarity with cost concepts and cost-accounting. To increase accessibility of these tools to evaluators and provide guidance on its use, the procedures have been incorporated into a free tool kit called COSTOUT, developed by CBCSE, which provides an expert application of accounting procedures for estimating costs, using a method that enables valid comparisons among alternative interventions. It can be used for studies of cost feasibility or comparisons of the cost-effectiveness of different educational alternatives that pursue similar educational goals.

In order to accommodate the needs of decision makers with a fuller and more useful set of findings on educational interventions, including economic consequences, evaluators should include measures of both costs and effectiveness in their evaluations. These should be based upon appropriate methods for providing valid and reliable costs in parallel with efforts to provide valid and reliable estimates of effectiveness. Cost-effectiveness comparisons can help decision makers take economic constraints into account when choosing educational reforms, ultimately improving evidence-based policy decisions and strengthening education systems.[3]

NOTES:

1. The ingredients method is described in Levin & McEwan (2001). However, it was first published in Levin, H. (1975) Cost-effectiveness in evaluation research, in M. Guttentag & Gruening, E. (Eds.), Handbook of Evaluation Research (Vol. 2, pp. 89-122) Beverly Hills, CA: Sage.
2. These studies are all found among the publications of CBCSE.
3. While this posts focuses on cost effectiveness, The Center for Benefit Cost Studies in Education also has a considerable bibliography of benefit-cost studies in education, extending to topics such as reducing high school dropouts, social and emotional learning, academic and social support services, college graduation, and other subjects. Benefit-Cost studies are devoted to comparing the monetary costs and benefits of educational interventions to calculate the monetary value of returns on such investments.

EVIDENCE AT THE CROSSROADS, PART 11

The Next Generation of Evidence-based Policy

VIVIAN TSENG | MARCH 23, 2016 | HTTP://BIT.LY/1WUCN7V

When we began this blog series, we posited that evidence-based policymaking was at a crossroads. In the past six months—despite rancorous partisan debates and a fierce presidential primary season—Congress surprised everyone and passed the long overdue re-authorization of the Elementary and Secondary Education Act, with strong support from both parties.

The Every Student Succeeds Act (ESSA) includes over 80 mentions of "evidence" and "evidence-based," and a devolution of power to states and districts to implement those provisions. And just last week, the Evidence-Based Policymaking Commission Act, sponsored by Representative Paul Ryan and Senator Patty Murray, was approved by the Senate and the House in another display of cooperation.

It is promising that at a time of heightened political rancor, evidence-based policy is finding bipartisan support. But the road ahead is still tenuous, and much will depend on whether the evidence movement can evolve. Here, I draw on the terrific ideas and insights from the authors of this series to suggest three steps for moving forward: focus on improvement, attend to bodies of evidence, and build state and local capacity for evidence use.

Focus on improvement

It's time to position evidence-based policy as a learning endeavor. Implementing and scaling interventions in different contexts with diverse groups is notoriously challenging. Promising results are emerging, but not all are home runs. The history of evaluation research shows that most evaluations yield mixed or null results, and this generation of studies will produce the same. Interventions work in some places for some people, but not others. Even new studies of established interventions turn up findings that are inconsistent with prior studies. What should we make of these results?

One direction we should not take is to obscure these findings or pretend they don't exist. I fear that already happens too often. The rhetoric of the What Works agenda—funding more of what works and less of what doesn't—has created an environment that pressures program developers to portray home run results, communications engines to spin findings, and evaluation reports to become more convoluted and harder to interpret.

Improvement could be the North Star for the next generation of the evidence movement. The idea of building and using evidence simply to sift through what works and what doesn't is wasteful and leaves us disappointed. We need to find ways to improve programs, practices, and systems in order to achieve better outcomes at scale. Let's not be too hasty in abandoning approaches that do not instantly pay off, and instead learn from the investments that have been made. After all, many established interventions had years to gestate, learn from evidence, and improve. Let's not cut short this process for new innovations that are just starting out.

This is not to say that anything goes. Patrick McCarthy reminds us that when research evidence consistently shows that a policy or program doesn't work—or even produces harm—it should be discontinued. Indeed, the next generation of evidence-based policy will need to aim toward improvement while keeping an eye on whether progress is being made.

Attend to bodies of evidence

If evidence-based policy is to realize its potential to improve the systems in which young people learn, grow, and receive care, we need to rely on bodies of research evidence. Too often, public systems are pressured to seek silver bullet solutions. A focus on single studies of program effectiveness encourages this way of thinking. But, as Mark Lipsey writes, "multiple studies are needed to support generalization beyond the idiosyncrasies of a single study." Just as a narrow aperture can exclude the important context of an image, so too does focusing on a narrow set of findings exclude the larger body of knowledge that can inform efforts to improve outcomes at scale.

State and local leaders need to draw on bodies of research evidence. This includes not only studies of what works, but of what works for whom, under what conditions, and at what cost. What Works evidence typically reflects the average impact of an intervention in the places where it was evaluated. For decision makers in other localities, that evidence is only somewhat useful. States and localities ultimately need to know whether the intervention will work in their communities, under their operating conditions, and given their resources. Evidence-based policy needs to address those questions.

To meet decision makers' varied evidence needs, the evidence movement also needs to focus greater and more nuanced attention to implementation research. Real-world implementation creates tension between strict adherence to program models and the need to adapt them to local systems. To address this tension, we need to build a more robust evidence base on key implementation issues, such as how much staffing or training is required, how resources should be allocated, and how to align new interventions with existing programs and systems. As Barbara Goodson and Don Peurach argue, we have built a powerful infrastructure for building evidence of program impacts, but we need to match it with equally robust structures for implementation evidence.

And finally, the evidence-based policy movement needs to recognize the importance of descriptive and measurement research that helps local decision makers better understand the particular challenges they are facing and better judge whether existing interventions are well suited to address those problems. For those needs assessments, descriptive and measurement studies can be critical.

Build state and local capacity

As decision making devolves to states and localities, the way the federal government defines its role will also change. In the wake of ESSA, officials in Congress and the U.S. Department of Education are aiming to move beyond top-down compliance. But to do so they will need to identify new means to support states, districts, and practitioners in the evidence agenda. States and localities are not mere implementers of federal policies, nor are they simply sites of experimentation. A key way to foster the success of the evidence movement is to support the capacity of state and local decision makers to build and use evidence to improve their systems and outcomes.

Technical assistance is one way that the federal government can support capacity, and it'll be important to direct technical assistance to state and local decision makers and grantees in productive ways. While tiered evidence initiatives such as i3 have provided grantees with technical assistance to conduct rigorous impact evaluations, assistance has focused less on other key issues: helping grantees apply continuous improvement principles and practices, vet and partner with external evaluators, and build productive collaborations with districts and other local agencies to implement programs.

Providing technical assistance in these areas would increase the ultimate success of these evidence-based initiatives.

Research-practice partnerships (RPPs) are another way to support state and local agencies. In education, these long-terms partnerships can provide the research infrastructure that is lacking in many states and districts as they seek to implement the evidence provisions in the Every Student Succeeds Act. RPPs can help districts and schools interpret the existing evidence base and discern which interventions are best aligned with their needs. In instances where the evidence base is lacking, RPPs are poised to conduct ongoing research to evaluate the interventions that are put into place. Similarly, in child welfare, research-practice partnerships could provide states with additional capacity as they develop Title IV-E Waiver Demonstration Projects to test new approaches for delivering and financing services in order to improve child and family outcomes.

The federal government is perhaps uniquely situated to build and harness research evidence, so that what is learned in one place need not be reinvented in another and the lessons accumulate. Mark Lipsey suggests that federally funded research require the collection and reporting of common data elements so that individual studies can be synthesized. Don Peurach imagines ways the federal government can support an "improvement infrastructure." We should consider these ideas and others as we move forward.

Foundations also have a role. Private funders are able to support learning in ways that are harder for the federal government to do. The William T. Grant and Spencer Foundations' i3 learning community, for example, provided a venue for program developers to share the challenges they faced in scaling their programs and to problem solve with one another. In another learning community, our foundation supported a network of federal research and evaluation staff across various agencies and offices to learn from each other. A learning community requires candor, and can provide a safe and open environment to identify challenges and generate solutions. Foundations can also produce tools and share models that states and localities can draw upon in using evidence. With fewer bureaucratic hurdles, we can often do this with greater speed than the federal government.

Realizing the potential of evidence in policymaking

The ascendance of research evidence in policy in the past two decades gave way to investments in innovation, experimentation, and evaluation that signaled great progress in the way our nation responds to its challenges. But for all the progress we've made in building and using evidence of What Works, we've also been left with blind spots. As a researcher, I did not enter my line of work expecting simple answers. Quite the opposite, in fact. Researchers, policymakers, and practitioners know that there is always more to learn than yes or no; more at stake than thumbs up or thumbs down. We build and use research evidence not just to identify what works, but to strengthen and improve programs and systems—to build knowledge that can improve kids' lives and better their chances to get ahead.

As we approach the next generation of evidence-based policy, it's essential that we take steps to ensure that practitioners and decision makers at the state and local level have the support they need.