Mark Dynarski
Pemberton Research

Presented at the National Conference on Scaling Up Effective Schools
Nashville, TN
October 2015

Scaling up effective education policies or programs is in everyone's interests. Who argues that education should not improve? And findings from research often underscore where improvements are possible. But scaling up these findings—using them on a larger scale, in more districts or schools—is a **huge** issue. The issue is partly the size of the public-education enterprise, with its 15,000 districts, 65,000 schools, 4 million teachers, and 55 million students. For an effective practice to find its way into even a fraction of this enterprise would be progress. The concern is that effective practices may not be finding their way into the enterprise at all.

In what follows, I compare effectiveness research, which strives to measure effects of programs, practices, or policies ('what works'), and improvement science, which strives to turn information about what works into real improvements in education outcomes at the level of districts and schools. I will explain how the approaches complement each other and how they can be more closely aligned to work together.

Background: Experiments, improvement science, and the scaling up problem

A defining feature of effectiveness research is the use of an experiment, which typically is structured as a randomized controlled trial. It also could be a regression-discontinuity or single-case design, but for purposes here, it is convenient to use 'experiments' as a term for a research approach that measures effects of programs, practices, or policies.

The logic of a randomized trial is startlingly simple. It needs only *two* pieces of data to estimate a causal effect: an indicator for whether an individual is in the treatment group or control

group, and an outcome. The indicator needs to be generated through a randomization mechanism, which any spreadsheet provides. Of course, real experiments are more complex, but their logic is the same. Researchers randomize individuals into groups and compare average outcomes of the groups. It is remarkable that this simple logic allows researchers to measure effects *caused* by programs or policies. By design of a trial, the program had to cause the difference (after allowing for variance arising because groups are sampled), because there is no other explanation for differences between outcomes of the two groups. Randomization eliminated all other causes.

Using experiments to study effects in education got a significant boost in 2002, when Congress created the Institute of Education Sciences. There were a few experiments before 2002 but no entity whose mission was to produce them. Since 2002, IES has funded hundreds of experiments and disseminates their results mostly through the web. It also disseminates syntheses of research findings through its 'What Works Clearinghouse,' also by way of the web. The information reaches a large audience. Practice guides produced by the What Works Clearinghouse are downloaded about 22,000 times a month. One of its most popular practice guides was downloaded nearly 90,000 times in its first *month* of release.[1]

This sounds like a lot of dissemination, but there is a blind spot. Whether educators are actually changing their practices because of the findings is not known. It is logically possible, though admittedly extreme, that educators are ignoring the findings and continuing to do what they are doing. The Government Accountability Office expressed its concerns about this possible disconnect in its recent review of IES.

Findings from experiments are information, but changing practices to do something with the findings is implementation. As Pfeffer and Sutton (2000) have written, knowing is a long way from doing.[2] Improvement science strives to close the gap between knowing and doing.[3] At a

[1]Full disclosure: I served as director of the What Works Clearinghouse from 2008 to 2010, when practice guides were first released, and chaired a panel that produced one of the first guides. I continue to be involved with the Clearinghouse in various roles.

[2]Jeffrey Pfeffer and Roberto Sutton. 'The Knowing-Doing Gap.' Harvard Business School Press: 2000.

risk of oversimplifying, improvement science poses a model in which researchers work with educators inside districts and schools, using research as a guide for the changes to effect. Notable partnerships between schools and researchers operate in Chicago, Boston, San Francisco, Oakland, Baltimore, San Diego, and New York City.

But improvement science also has a blind spot. It focuses intensively on a small number of districts. An improvement in one district is difficult if not impossible to replicate in another without undertaking the same investments to work with educators, organize staff, identify objectives, create consensus, and implement changes. But there are about four times more school districts than higher education institutions. Unless knowledge of how to do these efforts without researchers being involved can be disseminated, the limited number of researchers trained and available to support improvement efforts essentially precludes scaling them up.

Of course, efforts to change practices, based for example on evidence from a WWC practice guide, also need to identify objectives, create consensus, and so on. But these efforts do not require researchers to work directly with districts. To the extent that research findings spur actions in schools, the cost for an educator to acquire that information is nearly zero, mostly the value of their time. This is not to say the research was inexpensive, but disseminating it on the web is inexpensive. Schools can benefit from the information at very low cost.[4]

So we face a tradeoff. Experiments can be inexpensively disseminated but might not be used to improve education; improvement science focuses directly on improving education but is

---

[3]The underpinnings of randomized trials can be traced back a hundred years to groundbreaking work by R.A. Fisher and Jerzy Neyman. See Fisher's 'Design of Experiments'.' Improvement science is much newer. I am using 'Learning To Improve: How America's Schools Can Get Better At Getting Better,' by Anthony Bryk, Louis Gomez, Alicia Grumow, and Paul LeMahieu (2015) as a reference work for it, and also Bryk and Gomez's 2007 paper on improvement science and the National Academies monograph laying out the design of the Strategic Education Research Partnership. Cohen-Vogel and her colleagues provide a succinct recounting of the emergence of improvement science against the backdrop of experiments. See L. Cohen-Vogel *et al.*, "Implementing Educational Innovations At Scale: Transforming Researchers into Continuous Improvement Scientists." *Educational Policy*, vol. 29(1), 257-277.

[4]This ability for research findings to be widely and cheaply shared is an important reason the Federal government should play the leading role in education research, as I have written elsewhere.

expensive to disseminate and has capacity issues. The challenge is to narrow the gap between them. Each contains useful elements the other should do.

(II) How experiments can be more useful for practice

Experiments are an ideal tool for studying effectiveness of treatments, policies, or interventions. Indeed, studying effectiveness using a 'clinical trial' is the dominant approach in medicine, and its use in education is typically reported by saying a study used the 'gold standard,' an archaic term that means just about nothing but seems to resonate.[5]

But while education and medicine both use experiments to study effectiveness, they have notable differences that affect whether evidence is used. Medicine has 'standards of care' and a legal framework for malpractice that comes with it. Not applying a standard of care, which is informed by research on treatment effectiveness, opens a health provider to malpractice claims. Nobody would be surprised if a doctor who treats infections using leeches to bleed a patient is sued for not having used antibiotics. They are known to be effective and are the standard of care for routine bacterial infections.

In contrast, a school, district, or state is on its own to do what it wants with evidence, in a way physicians are not. A school district that uses a discredited or unsupported approach for teaching reading or math faces no consequences if its board and its parents are satisfied with it. The 'Drug Abuse Resistance Education' model has been widely discredited, for example, but the organization that administers the program claims it operates in 70 percent of the nation's school districts.

A second difference is that our understanding of human physiology is more developed than our understanding of how children learn. This is not to say human physiology is completely understood, but the mechanism by which a statin drug reduces cholesterol, for example, is about the same from one person to another. When physicians prescribe a statin, they can be

[5]The gold standard was the system used to convert paper currency into gold at the stated 'standard' rate. The U.S. abandoned it fifty years ago. Why it has come to be used as a synonym for valid research is a puzzle.

confident that their patients' cholesterol will go down. In contrast, a teacher faces 20 to 30 students, each with a different history, a different set of abilities and interests, and a different day-to-day energy level. The number of ways anything is learned in a typical classroom is some combinatorial of all these factors. An experiment might average outcomes of hundreds or thousands of students to find that 'reading scores are higher when this approach is used rather than that one.' It is not surprising if educators treat this finding as highly abstracted from their day-to-day experiences. Surely, they think, the approach worked for some students differently than for others. Any classroom teacher knows students learn differently.

That educators use evidence voluntarily provides a lens for thinking about what kind of properties the evidence would need to have to be adopted. Suppose an experiment reports that it studied an approach for reading or math or science or attendance, for example, and found that it improved outcomes. That improvement is likely to attract attention from educators. But then a number of steps need to happen. The number will vary depending on the intervention or policy being considered, but it might look something like this. A district or school:

1. Learns of an improvement
2. Also has the problem on which the experiment focused
3. Does not currently use that approach (if it did, there's nothing to do)
4. Needs to judge how much staff in the study differ from staff in their schools
5. Needs to judge how much prior history, skill levels, and characteristics of students played a role in the experiment's findings, because their students may differ
6. Needs to judge how implementation will fare in their context; the research probably does not describe how the program was rolled out and the kinds of obstacles it hit
7. Needs to judge cost, both in direct terms, such as purchasing the intervention, and in indirect terms, such as the time of teachers and other staff to be trained on how to carry it out and to have questions answered as issues arise; the research probably does not provide data on how much the intervention costs.

That is a lot of judgments, and together they can create a significant gap between how the researcher may view the results—'the study shows that the approach worked'—and how an educator might view the results—'it worked for somebody but I don't know if it will work in my context.' What a researcher viewed as evidence becomes what an educator views as risk.

In the practical world of education organizations, which essentially are large non-profit businesses, implementation and cost are crucial pieces of information. Imagine telling a business executive that a new approach will be effective but nobody knows how to implement it or how much it will cost. This will be a brief conversation. Considering ways to improve what is learned about implementation and cost will improve the likelihood that evidence from experiments is used, which I will return to below.

(III) Improvement science needs to be more accountable and replicable

Improvement science separates knowing what works from doing what works. Its hallmark is being at the table with educators to improve education by doing what works. But just as educators might question whether they can really implement an approach that an experiment might highlight as effective, educators might question whether improvement science really results in improvements, and whether it is cost-effective.

Improvement science lacks the kind of evidence of its successes that experiments produce. It is a young field, of course, and using experimental or quasi-experimental methods to study its effects is a serious challenge. But it needs to be tackled. Bryk *et al.* (2015, pp. 206) argue that whether improvement-science efforts are effective needs to be judged by whether predicted improvements occur. This is close to being a tautology because it sidesteps how improvements are predicted in the first place. Just about any effort will meet its targets for improvement if targets are set low enough.

It is possible to specify targets that, if met, provide evidence that improvement efforts succeeded. For example, if after working with improvement scientists, a school district's math scores show the largest gains in the twenty years, the 'causal warrant' is stronger for believing

the gains resulted from the improvement efforts. Whether targets are being set this aggressively is unclear.

But having external targets—which is what experiments and quasi-experiments provide–definitely helps to unravel the argument's circularity. Using a counterfactual as an external benchmark means that the 'predicted' improvement is the improvement that districts or schools would have experienced without the improvement science effort. If the effort leads to larger improvements than districts or schools otherwise would have experienced, we can say it 'worked.' Funders are likely to find this argument more persuasive than claims by districts or schools that their outcomes exceeded their predictions. Combining these approaches also could be considered. For example, targets could be set based on recent trends of neighboring districts with similar achievement and demographic profiles. Essentially, this approach uses a counterfactual logic but does not need to measure contemporaneous improvements in comparison districts.

Analyzing simple trends shows the need for more valid evidence of improvement science's effectiveness. One of the most visible organizations using improvement science is the Chicago Consortium for School Research. It began working with the Chicago Public Schools in the early nineties. Data from the National Assessment of Educational Progress's 'Trial Urban District Assessment' shows that reading scores for Chicago's 4th graders rose 13 points between 2002 and 2013, from 193 to 206. For all large cities as a group during that time period, reading scores for 4th graders rose 10 points, from 202 to 2012. The difference between Chicago and other large cities is 3 points, which is not nothing but not quite something either. Reading scores for 4th graders in the District of Columbia rose 15 points in that period, from 191 to 206. The District invested in many improvements but it was not working with a consortium.[6]

The Austin school district has been participating in the Middle Schools Mathematics and Institutional Setting to Teaching, described by Cobb, *et al.* That effort concentrated on the

[6]A recent report from the National Academy on DC school reforms discusses those improvements.

years from 2007 to 2011. In those years, math scores for Austin's eighth graders rose 4 points, (283 to 287). In the same period, math scores for eighth-graders in urban districts rose *five* points (269 to 274). The difference is within the margin of sampling error, but it's fair to say Austin's gains in math were no larger and in fact numerically smaller than gains in 11 other cities.

With just these trends in hand, could improvement science argue that its efforts are worth funding? Trend analysis is open to criticisms: other events may happened, the demographics of student populations may have changed, scores may have been even worse if districts did not work with researchers. But these criticisms underscore the need for formal evaluations of improvement efforts.

And the lack of cost information adds to uncertainty about their value for money. How much effort was spent by the outside improvement team and by district staff? The Chicago Consortium is part of the Urban Education Institute at the University of Chicago, which has an annual operating budget of $47 million and employs more than 450 full and part-time staff. This cost figure overstates costs on the one hand because the Institute operates other programs. It understates costs on the other hand because it does not count the resources and time of district staff.[7] But if $20 million a year goes to the Consortium, then $220 million was spent to raise NAEP scores by a small amount, and to raise other outcomes too. It is not a lot of money relative to the $60 billion spent by the Chicago school district in the same period, but it is a lot of money for research.

Putting gains alongside costs will help to know whether simply disseminating research is the most cost-effective means of generating improvements. It's an open question, but comparative effectiveness research in health provides cautions about expensive and usually newer way of

[7]Organization budgets are clumsy tools for measuring cost, but I could find no information about costs of improvement efforts in particular places, as one might if a government funded the efforts though a contract or grant.

doing things being no more effective than cheaper and usually older ways of doing things. 'New' and 'better' are not the same.[8]

(IV) The two approaches can learn from each other

Accepting that experiments are weak on implementation, questions arise about how implementation research in the context of experiments can be more useful. Experiments have a template for measuring effects, based on underlying statistical methods, and a relatively narrow range of techniques for measuring effects. Implementation studies do not have a template and their methods encompass a variety of techniques in qualitative and quantitative research.

A sketch of current practice helps to set the stage. A typical scenario is that a new approach, perhaps a reading curriculum or a dropout prevention program, is funded by a government agency or a philanthropy, which in turn commissions an evaluation of its effectiveness that includes studying implementation. Grants to university researchers differ in specifics but usually include the same elements: a new program, an entity that will implement the program, and a study of the program's effectiveness.

Within these studies, researchers looking at implementation might carry out time-intensive activities such as focus groups, one-on-one or group interviews with staff, on-site observations, and reviews of documents about planning and operating the program. Questions might include how the program developed its logic model, the kinds of hurdles it encountered in staffing, funding, and delivery, and its approaches for getting over the hurdles. In some studies, researchers measure how close programs came to their intended model, the 'fidelity'

[8]Comparative effectiveness studies (see here and here) report that different treatments to reduce blood pressure have about the same effects, which means the cheapest one is the comparatively most effective one. Exercise is about as effective for reducing back pain as surgery and far less expensive. Different physical therapies to reduce knee pain have about the same effect.

of implementation. Recent efforts measure the size of the 'contrast' between the program and what it is being compared to, such as business as usual or another program.[9]

The mishmash of questions and techniques creates wide variation in implementation studies. Several hundred pages of detailed information about implementation accompany some experiments; others describe implementation in a report chapter, and many published papers from experiments simply do not mention implementation.[10]

Implementation studies also do not have standards against which they can be judged. The What Works Clearinghouse and Best Evidence Encyclopedia, for example, do not have standards for implementation and their reports do not even mention implementation. The gap is understandable: most papers or reports provide too little information that would enable implementation to be assessed, or different studies of the same program or intervention report on different kinds of implementation findings. Standards follow from consistent approaches.

With respect to cost, Henry Levin and colleagues have fleshed out approaches for studying costs and compiled them into a reference book.[11] They stress that cost analysis should enumerate all the resources a program needs to operate, which can differ greatly from its out-of-pocket cost. The stated price of a new math textbook does not include the cost of the time teachers devote to attending workshops to learn about the curriculum and to developing new lesson plans. For districts considered using that textbook, having an estimate of how much teacher time is needed to use it provides a more realistic picture of costs. Districts should want to know this, or they might discover it the hard way when teachers begin to use the new textbook and find their nights and weekends have just been claimed.

---

[9]Cordray explores implementation fidelity here. Hulleman and Cordray propose methods for measuring the size of the contrast. See Hulleman, C. S., & Cordray, D. S. (2009). Moving from the lab to the field: The Role of fidelity and achieved relative intervention strength. *Journal of Research on Educational Effectiveness*, 2, 88-110.

[10]For examples, see the Reading First study reported here, the study of teacher induction programs reported here, and the study of supplemental services reported here.

[11]Henry Levin and Patrick McEwan. Cost-Effectiveness Analysis: Methods and Applications, 2nd ed. Sage Publications: 2002.

But even direct costs need to be known for educators to make informed decisions about whether to adopt an evidence-supported program or practice. Reporting that a program or practice is effective is less useful than reporting that it is cost-effective. Knowing the 'effect per dollar' is a central piece of information. Inexpensive programs that yield small effects might suit educator budgets more than expensive programs that yield large effects.

Improvement science focuses on developing improvements. In a sense, it's all about implementation, informed by evidence of effectiveness. But its gaps are worth noting too. One gap is that, like implementation studies, there are no standards for improvement science efforts. The What Works Clearinghouse or the Best Evidence Encyclopedia can rate an experiment and give readers a sense of whether the underlying research was sound. These standards certainly are argued about, as any scan of the literature will show, but the point is that they can be applied. There is no counterpart for improvement science, which means a third party cannot judge the quality of an effort.

The degree of interaction called for by improvement science also creates a premium on relationships. It is well known that educational leaders turn over frequently, which means an improvement effort (or any innovation or initiative) that starts under one leader might be ended or diverted by the next. So, improvement science works to create processes and structures that need to outlast leaders, but the leaders to be outlasted are being asked to support work that might be slow, might require a lot of staff effort, and might show improvements only after the leaders have moved on. Generating buy-in under these circumstances seems like a challenge affecting scaling up.

(V) Two proposals to move the approaches closer

Both approaches add to knowledge and both approaches can be more useful. Starting from a view that everything should scale up and that the two approaches are linking within a tiered model is appealing.

(A) Scaling up should be the starting point in designing experiments or improvement science efforts.

As discussed above, current efforts do not convey all the information needed to make sound decisions that contribute to scaling up. Perhaps the long history of null effects in education has led researchers to conclude that the program or policy probably will not show effects scaling up is moot, so designing a study from the perspective of scaling up is not needed. But that thinking contributes to effective programs not scaling up. A researcher that finds positive effects but did not study implementation and cost probably will find it is too late to go back to study them.

If studies were designed from the view of scaling up, implementation analyses could use 'how to do it' as a conceptual framework. Rather than studying how a program was implemented, the perspective would shift to what has been learned from this instance about how others could implement it. This knowledge would cumulate as more studies are done.

Improvement science likewise can follow scaling-up logic by documenting its implementation, costs, and effects. Nothing about documenting these is counter to improvement science's tenets. Other districts can learn from this information and possibly replicate it. Some might argue that scaling up improvement science misses its point: every school district is a complex and unique organization and needs to be treated uniquely. But until improvement science can show it leads to improvements, is cost-effective, and can be scaled up, it is hard to imagine that policymakers will fund more improvement efforts.

(B) A tiered model can combine experiments and improvement science

A tiered model such as the one used for response to intervention helps to put experiments and improvement science into a sensible balance. In the first tier, districts and schools can learn about recent innovations, where the field is trending, and how to implement promising interventions. All states, districts, and schools can examine this evidence and elect to put it into practice or not. The point is that this first tier is about disseminating *useful* evidence—

effects, implementation, and cost—on a large scale. It needs to be broad and able to reach many users for reasonable cost. Currently there is no organization doing what is envisioned here. In fact, dissemination is done by loosely connected organizations and agencies. And, as noted in the introduction, how the information being disseminated is affecting practices is not well understood.[12]

The second tier of the model is improvement science. Districts that use evidence from the first tier but are not satisfied with the results or do not meet targets can work with improvement scientists to identify ways to improve. The second tier needs only enough improvement-science capacity to work with districts that are in it. This may still be too many districts and not enough capacity, but starting from the total number of districts certainly overwhelms capacity, as noted above. And if improvement science generates tools or approaches that help educators use evidence more effectively and supports local improvements, these can be disseminated in the first tier.

The first tier also can include findings on education strategies or policies that improvement science does not study. For example, the effectiveness of charter schools is important information, and educators can 'implement' those findings without efforts of an improvement science team. Teacher incentive pay is another policy that districts can implement directly.

The Bush Institute's 'Middle School Matters' collaboration with the University of Texas uses a three-tiered model. The first tier provides curated resources that any middle school can use. The second tier provides workshops, training, and follow-up closely tied to a school's self-assessment of where they need support. The third tier has researchers visiting schools to present approaches and model them with teachers. The level of direct interaction with schools rises in the higher-level tiers.

Who is responsible for operating the model? NCLB stressed that states and districts should identify effective programs and policies to raise student proficiency. But little evidence was available at the time of its passage. Almost 15 years later, we now have more evidence and a

[12]IES recently funded two research centers to explore these topics but it will be several years before findings are known.

stronger basis for both tiers. The 'Every Student Succeeds Act' put the onus of developing accountability structures on states. Using the tiered approach—with districts and schools moving to the second tier if improvement targets are not met—is a sensible balance of evidence and practice. Its third tier could be to require implementing school turnaround models, but evidence of the effectiveness of turnarounds is mixed at best. A more intensive approach that amps up activities in the second tier may be a more appealing approach.

(VI) Looking Ahead

Education research has yet to reach the scale of health research, and it may never get there. The National Institutes of Health spend more than 60 times what the Institute for Educational Sciences spends, a big difference that is not likely to disappear soon. Evidence that education research generates improvements will help break out of the cycle of not spending money on education research and being disappointed that there is not enough education research. The field needs some winners, like health research with the polio vaccine trials in the fifties, but winners in the last two decades have been scarce.

In this resource-starved context, effectiveness research and improvement science need to work together. The tiered model discussed in section V is one way they could work together. Information about implementation and cost is essential for both approaches, and developing templates for reporting on these will be useful, similar to templates such as the structured abstract or the CONSORT diagram. The field will benefit from consistent reporting of information that educators need to weigh when they are deciding whether to scale up.

Acknowledgements: I thank Lora Cohen-Vogel and Kirsten Kainz for their helpful comments on the previous draft.