Principles for Scaling Up: Choosing, Measuring Effects, and Promoting the Widespread Use of Educational Innovation

CSE Report 634

Eva L. Baker
National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
University of California, Los Angeles

July 2004

Center for the Study of Evaluation (CSE)
National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
310-206-1532

Eva Baker, Project Director, Project 2.1: Cognitively Based Models for Assessment and Instructional Design, Strand 1: Cognitively Based Models and Assessment Design

The work reported herein was partially supported by NORC and under the Educational Research and Development Centers Program, PR/Award Number R305B960002, as administered by the Institute of Education Sciences (IES), U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the positions or policies of NORC or the National Center for Education Research, the Institute of Education Sciences, or the U.S. Department of Education.

Excerpts of this paper have been taken from "Evidence-based interventions: What then and now what?" paper presented at the 2003 annual meeting of the American Educational Research Association. In M. Constas, session 48.010 PRES-59, *Advancing the Scientific Investigation of Large-Scale Interventions: The Interagency Education Research Initiative* (Presidential Invited Session), Chicago.

PRINCIPLES FOR SCALING UP: CHOOSING, MEASURING EFFECTS, AND PROMOTING THE WIDESPREAD USE OF EDUCATIONAL INNOVATION[1]

Eva L. Baker
CRESST/University of California, Los Angeles

Abstract

The goal of scaling up of educational innovation is to produce robust, effective, replicable outcomes. This report addresses requirements to support scale-up of scientifically vetted innovation (or new ideas that are built on the findings of quality research and development). In this report, a number of issues are considered: the context of evidence of effectiveness, definitions of scaling up, how to measure implementation and effects, and areas in need of continuing work, from the research point of view. The analyses grow in part from long-ago research training and in part recent R&D efforts. Recommendations are compressed into principles to guide action, elaborated with rationale and discussion, and then illustrated by a running example of CRESST's Model-Based Assessment program. The principles are: Principle 1: Understand the Appropriateness of the Evidence for the Innovation; Principle 2: Document the Scientific Evidence in Support of the Design of the Intervention; Principle 3: Embrace Design and Development; Principle 4: Measure a Broad Range of Variables in Formative Evaluation; Principle 5: Design the Best Study Possible; Principle 6: Support Flexibility and Local Innovation. The ultimate goal is the articulation of usable principles for scaling up and a contribution to the definition of the slippery concept of scaling up; Principle 7: Minimums for Scalability and Sustainability.

Introduction

This report will address requirements to support the scale-up of scientifically vetted innovation (or new ideas that are built on the findings of quality research and development). Scaling up, that is, adapting innovation for widespread and supported use, is a laudable goal, but one clearly not appropriate for all research-

[1] Baker, E. L. (forthcoming). Principles for scaling up: choosing, measuring effects, and promoting the widespread use of education innovation. In B. Schneider (Ed.), *Proceedings of the Data Research and Development Center Conference "Conceptualizing Scale-up: Multidisciplinary Perspectives."* Chicago: University of Chicago, NORC.

based innovations. Documented effectiveness and feasibility are especially important for innovations that involve children, who in schools have little or no option to agree to the services provided by the schools. Despite the enthusiasm of designers and early users, many innovations need an extended period of time to develop, in order to meet feasibility requirements, and to gain the credibility conferred by a strong base of evidence. In this report, a number of issues will be considered: first, the context of evidence of effectiveness; then, definitions of scaling up, how to measure implementation and effects, and areas in need of continuing work, from the research point of view. The ultimate goal is the articulation of usable principles for scaling up and a contribution to the definition of the slippery concept of scaling up.

The Scientific Context of Evidence for Innovation—a Brief Personal History

No discussion of scaling up research can stray far from the current frame of federal policy preferences for types of scientific evidence, when evidence is the prerequisite to broader implementation. To begin this analysis, I will draw on my own scientific training, the evolution of thinking in the field as I experienced it, and, where relevant, the work of my colleagues at the Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Experimental Design Beliefs and Realities

Let's start with me. Few of the readers will have any reason to know, that when I was in graduate school, I was mentored by an experimentally oriented learning psychologist and statistician (A. A. Lumsdaine). He worked with a very small number of students, and was both revered and feared by most students, mostly because he gave backbreaking assignments and detailed, public feedback. For instance, he required to be turned in each week an original, full, complete (long) proposal for a significant, instructional true experiment, including a pre-Google review of the literature, and a formal oral defense of the merit of the work. In addition, we were to select one of these proposals (with his guidance) and conduct at least one true experiment with a reasonable treatment time each term. Apart from garnering an early publication record, and steeled acceptance for public criticism, we were taught by him about programmatic efforts that should build and accumulate knowledge—knowledge from the work of other scholars and our own line of effort. His interests were in making a difference in the real world. My studies

with him lasted two years, including summers, prior to his assuming the chair of the Psychology Department at the University of Washington. In abbreviated form, let me share the following 10 precepts synthesized from this experience:

1. The gold standard for research design and analysis was randomization, which meant, in addition to randomly assigned treatments, randomized selection of subjects and order of dependent measures, else probability estimates were inappropriate.

2. The situations and learners to which one intended to generalize necessarily defined the subjects and domains from which one sampled.

3. Comparing treatment X with present practice was not research, but an imprecise approach to evaluation. Research required manipulable, independent variables compared at different intensities.

4. Satisfactory research findings required generalization and replication.

5. Fidelity of treatments delivered had to be adequately verified.

6. Treatments should be of sufficient duration to have an impact.

7. The unit of analysis should be based upon the unit of randomization and generalization—that is, students, classrooms, or schools.

8. Obtained effects should be robust over stringent or lax criteria and across multiple dependent measures and time.

9. Never say "proven" or "was significant," because we are operating in a probabilistic world and we have many unmeasured variables. The best you can do is "found" given results.

10. If you can't do it right, don't do it.

There were other preferences, like stay away from complex statistics if simple ones would do, most represented by his fixation on the sign test and other nonparametric statistics

The precepts taught by Lumsdaine in the form of intense project-based learning, were not original. We read R. A. Fisher's *The Design of Experiments* (1951), a seminal volume that made the point that the use of inferential statistics was logically legitimated by introducing chance into the experiment through randomization, a point that is no doubt still news to an occasional student. We replicated Fisher's findings in simulations. We studied the evaluations of the *Why We Fight Series* and

other experiments in instruction (*Experiments in Mass Communication,* Hovland, Lumsdaine & Sheffield, 1949). We analyzed the eloquent expansion of Fisher by Campbell and Stanley in a chapter in the *Handbook of Research on Teaching* (Gage, 1963), the details of which decades of my own graduate students can no doubt still recite. We also consumed *The Conduct of Inquiry* (Kaplan, 1964) to understand logic and evidence from the perspective of the philosophy of science. I was happily schooled in analysis of variance methods (my mentor preferred plots, with the goal to "see" whether differences were important rather than "significant"). We calculated effect sizes and plotted power curves. All were pathetically heady experiences for me—a literature graduate student with history and anthropology minors. I was sure I was objectively and technically prepared to fix American education, or at least to know when it happened. Cookbooks were only used for cooking dinner.

In fact, 30 years ago at the annual AERA meeting, as a fledgling professor I took my learning into the professional realm; I presented a paper on the topic of the use of true experiments during the development of instructional interventions, and modestly suggested that, in the absence of scientific knowledge on which to base design decisions, instructional developers should conduct applied experiments to choose the components most effective empirically for potential interventions. I was ripped by the discussant. Undaunted, I continued to apply experimental approaches in my early work in teacher education and in developing primary-age reading curricula, using true experiments to investigate theoretically based options or seemingly arbitrary choices. Early on I was concerned with the sensitivity of measures to detect change, so I also employed the newly formulated criteria for designing achievement tests suggested by Wells Hively and his colleagues (Hively, Patterson, & Page, 1968).

The Context: From Then to Now

Note that the gold standard I was trained on doesn't much exist today, partly because of methodologies that have been invented to compensate for experimental controls, partly because of unintended burdens placed on researchers by human subjects protection requirements, and last, and most obviously, because the complexities and options of educational delivery of instructional alternatives rely on agreements, acceptance, and actions by multiple players.

But 20 years ago, or more, back then, I was hardly alone in my beliefs. At that time, there was in the air the proposal that schools ought not to use interventions (curricula or instructional products) unless experimental studies supported the superiority of the proposed treatments against some criterion. EPIE (Educational Product Information Exchange), at Columbia University, had a proponent of this view in its director, Ken Komoski. Lumsdaine (1965, chap. 7) argued that effectiveness was a defining characteristic of the term "instructional program"; one didn't really have an educational program unless one could document that at least two criteria had been met: that the program **processes** could be *essentially reproduced* from occasion to occasion, and that the program's **effects** were also *reproducible or replicable.* In fact, California put on its books a law stating that before curricular materials could be used in public schools, commercial publishers were required to present experimental data documenting the intervention's effectiveness. The law was never enforced. Somewhat later, the National Diffusion Network, an arm of the U.S. Office of Education, required that experimental data be presented if federal support were to be used in delivering interventions to Title 1 schools. At that time, evaluation studies followed the paradigm of experimental research, a view advocated variously by Freeman and Sherwood (1970), Cook and Campbell (1979), Boruch (1976), and of course, Campbell and Stanley (1963).

For the most part, the procedures of experimental research on learning and instruction continued to be divorced from the large-scale implementation of educational curriculum. With some notable exceptions, including a few products of some of the Regional Educational Laboratories and a scattering of technology-based interventions, most of the curricula and instruction headed for schools was not evaluated convincingly. Written by one or more university professors or classroom teachers, these curricula (texts, workbooks, etc.) were published and distributed by commercial firms. The marketing cycles and profit margins associated with these enterprises were, respectively, too frequent and too thin for serious scholarly inquiry about the pre-market effects of their products. When advertisements referenced "demonstrated effectiveness," they more than likely referred to reviews and tryouts of lessons by a handful of teachers prior to the development of a teacher's manual in the curriculum of interest.

Methodology Expands

So what happened to true experiments? The utility of standard experimental features as essential to the evaluation of broad-based education programs was countered in the 1970s and 1980s with writings by Glass (1977), Cronbach et al. (1981), Bloom, Madaus, and Hastings (1981), Stake (1967), Scriven (1967), House (1974), Patton (1978), Weiss (1977), and Alkin (1985). Many argued for decision-theoretic studies—for studies of whether the program met identified needs, and whether comparisons, if any, were sensible. These writers variously identified multiple purposes for the use of such data. They also focused on the valuing of findings and the ability to understand what happened, including both intended and unintended consequences for all the participants in the studies. They were concerned with "merit," not simply numbers.

Their thoughts may have been influenced by a well-developed conceptual document by Cronbach and Suppes (1969) that identified criteria for distinguishing research from evaluation studies—between knowledge-producing and decision-related studies. This push-back on experimentation was bolstered by the creation of a large number of methodological innovations made to permit the study of real schools, that is, environments putatively unsuitable for true experimentation.

The difference between Lumsdaine's view and those positions more recently adopted resided in the difference between conceptions of the entity, *program*. For the most part, *programs* to be evaluated referred to state or federal policy interventions—far more likely to be complex goals and requirements rather than tightly conceived instructional treatments or products.

Considerable scholarship on the topic of the relationship of evidence to decisions of policymakers (for instance, Lindblom & Cohen, 1979; Weiss, 1977; and Wildavsky, 1993) has had a continuing life, across the realm of social science, rather than from an exclusively psychological perspective (e.g., Lumsdaine, excluding Fisher). Such analyses credited a wide range of approaches to obtaining legitimate evidence. Furthermore, the pieces broke away from a literal view of research, development, and evaluation following a clean left-to-right chronology and logic. Differing points of entry into the R&D process and the influence of political and social factors were provided as explanations for seeing the R&D situation as vastly more complex than choosing between two or more options.

A simple summary is that most policymakers use the best evidence they can obtain *to bolster their own preferences within their political context,* rather than using the evidence as the basis for forming their original opinion. The inference to be drawn, of course, is that there were more powerful incentives operating to influence decisions than documented effectiveness—like beliefs, political power, and cost.

Furthermore, studies of program implementation documented the enormous variation in the application of policy and program features, which resulted in great diversity of uses and effects. Nonetheless, the evaluation industry continued to the mid-1980s, during which time some applied true experimental designs to ill-defined and differently implemented educational programs, using generally insensitive outcome measures, resulting in an avalanche of no-difference findings, and the drum signaling that educators didn't know what they were doing in classrooms began its steady beat.

Another factor that shifted methods was the simultaneous redesign in Schools of Education of scholarly education faculties. Faculties diversified and expanded well beyond the dominance of psychologists and philosophers, to include scholars who were anthropologists, political scientists, historians, and sociologists. These scholars brought with them a wider set of methods, a penchant for focusing on close-up interpretative analyses, and clear goals to explore the *why* as well as the *what*. New faculty shifted the norms of research, so that what had previously been rewarded—for example, conducting lots of brief, treatment-true experiments (still the norm in some psychology departments)—was no longer of high value. Serious redefinitions of "scholarship" occurred at top-tier Schools of Education.

Why dwell on the past? And why these fascinating snippets of yore? They are directed to the young among the readers, who may believe that the present federal focus on scientific evidence is a wholly new concept. Now let's turn, as promised, to the crux of the presentation.

Scientific Evidence

Although I recommend a highly readable volume from the National Research Council (Shavelson & Towne, 2002) for a recent treatment of scientific evidence, let me make some brief remarks on the topic. Certainly scientific evidence is a matter of degree. We in social science understand that there is no single scientific method that produces sufficient evidence, for every purpose. We acknowledge, almost revel in

the reality that there is a range of legitimate forms of inquiry, and that quality criteria exist for each of the different methods and strategies.

Yet, for the most part, the lay public and most policymakers don't get the idea. They think that we do the best science when it is large-scale and objective (read *quantitative*), and that we can *prove* for all time (or for 4 years) that one alternative is better than all other choices. The public may now be encouraged to believe that one method should always be used to determine educational value. If we could help people understand the difference between the production of knowledge and the making of decisions (to reference Cronbach and Suppes, 1969, and Clark, 1985), we would be mightily advantaged. The differences between theory development, inductive logic, and *ad hoc* evaluation contrasts may well never be learned by our broadest constituencies, so let us focus our sights on the Interagency Education Research Initiative (IERI) and other programs at hand.

Principle 1: Understand the Appropriateness of the Evidence for the Innovation

Scientific evidence may be inferred by many to mean experimental and controlled studies of the effectiveness of a proposed intervention. Yet, the minimum requirements for true experiments are rarely met in educational situations, especially those in urban settings. For instance, volunteer populations of districts or teachers may be used because of bargaining agreements, extant policy, or penchants for collaboration. Thus, generalizing results to nonvolunteers is always problematic. Second, there is great volatility in the stability of some urban school sites, where students, teachers, and leadership may substantially change within a treatment period such as one year. Students are mobile. Substitute teachers may be a regular feature of instruction. There is also often a climate of distrust that has implications for methodology. For example, relying on self-report approaches, logs, surveys, notes, or diaries, in order to verify treatments is likely to yield overstatements of activities and result in inferences about large process differences where they may be very slight. Because affecting the course of education requires attention to a number of socially connected practices (as opposed to the difficult, but more simple matter of patient compliance, to regularly swallow a dose of Lipitor in the privacy of one's bathroom), in educational change social context and trust among players must matter. Furthermore, the embedding of schools themselves in broader political and economic reality suggests that influences far stronger than classroom events and the considered treatments may occur and vitiate or exaggerate impact of the treatment.

For example, changing state policy requirements for language of instruction, reducing or raising class size, the amount of money available for school spending, and the incentive structures implied by different accountability systems are intended to be "big ticket" interventions and may swamp other treatment effects. Contamination is also an issue. For instance, my colleagues and I valiantly tried to conduct true experiments over multiple years with real control groups and known interventions (see Baker, Herman, & Gearhart, 1996). We found, however, that when an intervention was perceived to be effective or even desirable by teachers, administrators, or parents, the leadership of the district or school was placed under pressure and ultimately found a way to compensate for the control groups' experience. In many ways, the administrators and teachers rapidly provided approximations of experimental treatment to the original control students, and naturally reduced the functional differences in teachers' and children's experience between control and treatment groups. It is true, however, that at least half the time (in two-group comparisons) students usually learn more.

Principle 2: Document the Scientific Evidence in Support of the Design of the Intervention

There are two related aspects for the scaling up of an innovation: its design and implementation. Evidence is required in support of both of these features. Unfortunately, a trend has been to credit evidence for full implementations of innovation. As a result, there is a misconception that the last-in, or the most recent, study matters most. This view has led a few among some practitioners and policymakers to support the idea of the grand finale study, the one that, hmmm, **proves once and for all** the superiority of X instead of Y. A more legitimate and scientific idea is that we should be looking at a cumulative impact of scholarship, analysis, interpretative studies, and experimental verifications when appropriate. The quality of the whole stream of evidence should be judged rather than looking only at the findings of the last study. It is true that more recent work may incorporate remedies to earlier research flaws, but the full range of studies should be addressed. Although standard meta-analytic approaches may be crude, they give a starting point to identify studies that have comparable goals and quality dependent measures. Furthermore, the evidence sources cannot be parochial, and reference exclusively the work of a particular research team. These studies should include

those completed by the research team but must include a wide swath of scholarship, especially those studies conducted by intellectual competitors.

CRESST Assessment Models Scale-Up—Example 1. Let me give an extended example using work from CRESST in the area of assessment.

How did we argue the scientific base of our Model Assessment Approach (Baker, Freeman, & Clayton, 1991; Baker & Mayer, 1999; Wittrock & Baker, 1991)?

Our first effort was to consolidate what the research evidence was in the area of knowledge comprehension. There were two main approaches. One held that all knowledge domains were idiographic and required specialized domain knowledge for expertise. The other held that there were some general principles of knowledge comprehension (Bransford & Johnson, 1972; Wittrock, 1974) and of developing expertise (Chi, Glaser, & Farr, 1988; Larkin, McDermott, Simon, & Simon, 1980). We chose to combine both approaches, adopting the argument that generalization of learning did occur and that attributes of expertise had common elements.

The integration of the literature suggested that deep understanding of content or subject matter required the following components:

- Distinguishing the main principles or themes of the topic
- Understanding the relationship among principles
- Connecting concrete examples and non-examples to the structure of knowledge

After many false starts, we converted these features into a scoring rubric where we asked students to write about various primary sources of knowledge—letters, speeches, experiments, graphical representations—and to explain their meaning and importance (and, in those cases where contrasting material was presented, their differences). This scoring rubric was applied to writing of students in history at a number of different grade levels. We then asked history experts, that is, professors and advanced graduate students, to answer the identical history questions, presented with the same materials. We contrasted their performance with that of teachers and students. Using methods derived from the expert-novice literature, we were able to document that experts did write principled, concrete, related pieces as responses. Moreover, their work included references to their prior knowledge in the area. When we looked at the contrast groups, we found that students tended to miss

the principle, relational and prior knowledge, and rather used such concrete knowledge as was available in the testing materials. They also wrote flat, relatively undifferentiated prose, with numerous misconceptions (Baker, 1994). Building on the literature in learning and in expertise, as well as our own series of studies, we refined the scoring rubric so that it generated a two-factor solution: an expert factor—principles, prior knowledge—and a novice factor—dependence on the text presented and misconceptions. The expert factor repeatedly predicted overall scores. Furthermore, the use of this scoring scheme gave teachers guidance for teaching students ways to learn types of material (rather than a particular text). Our own studies contrasted students who were in ability-tracked classrooms, students with different language backgrounds, students given multiple readings of the materials, students given support in prompts, students given tests first in background knowledge. Much of our attention was spent on getting score reliable (agreement with each other) and content valid (agreement with the intended expert scoring scheme) measures, as well as experimenting with response mode. Our results showed *predicted* relationships with other extant measures, such as standardized tests of knowledge, standardized writing measures, grades, etc. Relationships varied, as we would have predicted.

We decided to scale up after 30-plus such studies and conducted studies on a sampling basis statewide in Hawaii with 4th-, 5th-, 7th-, and 10th-grade students. I'll pick up on this story as I go. Note that before we began a modest scale up in Hawaii (about 1,000 students a grade) and different curricula, we had assembled data from the literature, from multiple trials of teachers, from classrooms at Grades 7-12, and in studies ranging from 60 to 400 students (Baker et al., 1996).

Principle 3: Embrace Design and Development

One of the persistent biases of psychologically oriented research has been its dependence on short treatment times, artificial situations, and a limited, volunteer subject pool. Scale-up is not just bigger research. For research to become a usable innovation, it must go through a development process. This process (described in Baker, 1973) involves the statement of goals, creation of procedures (replicable, *a la* Lumsdaine), and measured results that are subject to improvement by modified design and testing. It should sound a lot like No Child Left Behind (NCLB), which has adopted a systems approach to improvement, with only little control over the system elements. Nonetheless, in the more tractable area of a single innovation, it is

possible to use this kind of engineering (horrors!). But engineering is theory made real. The resulting comparison of the innovation to an alternative (spare us the unverifiable "conventional practice" as the option) is not strictly research to create new knowledge, but is decision-oriented investigation, what is commonly known as evaluation. The label in this case is important as it distinguishes between what types of inferences can be legitimately drawn, or the validity of the findings of the study.

There are tested approaches for the design and implementation of large-scale, research-based interventions that focus on the formative evaluation, or results-redesign, improvement cycle, on determining which research elements can be "saved" in a complex, messy environment, and at what cost and risk.

Researchers themselves are rarely prepared to anticipate users' needs in a volatile policy and social environment. Let me continue with the CRESST example to drive the point home.

In our Hawaii studies, we were interested in a series of research questions: Did the models and templates we used for performance assessment generalize across grade levels (4th, 5th, 7th and 10th)? Did the approach to measurement work across topics within subject areas (10 topics within each of the elementary, middle school, and high school environments), and were similar patterns found among subject matters (Hawaiian history—where the research team had paltry levels of expertise, and high school American history—from colonial times to WWII)? In addition, we were interested in multilevel models that might predict performance (e.g., students' reading comprehension; and at the classroom level, teachers' self-reported topic-by-topic subject matter expertise), and we also investigated opportunity to learn and group differences in performance. All of these questions were related to the general validity and instructional utility of the assessments. For scale-up issues, we were interested in teacher reactions, complexity of administration directions, time, and central office efficiency in distributing, collecting and transferring data. For scoring, our concerns again were validity with the expert scale, level of scorer agreement, time and cost per student paper scored, training and retraining requirements for raters, and methods of clear reporting. In the second year, our interest focused on change in performance, but also on teachers' reactions and initiatives to share instructional knowledge, and most importantly, whether the system could train raters by themselves and obtain usable results (it could).

All of this effort was based on the assumption that CRESST would "transfer its technology" to an entity external to the university, such as a state, school district, or commercial enterprise and get out of the continued scale-up business. Why? Because our organization is focused on the R&D needed for quality models and tools, and our expertise does not lie with packaging, marketing, distribution, and massive technical support. It is clear that, under any transition, expert advice is during the scale-up phase to sustain and adapt the interventions.

Principle 4: Measure a Broad Range of Variables in Formative Evaluation

The credibility and ultimate success of any innovation will depend upon its motivated and practical use on the ground. We recommend that any intervention or indicator be subject to empirical tests using a broad base of dependent measures and the fastest cycle time possible, so that multiple iterations of trial and revision are possible.

Multiple measures need to be selected to address the range of potential outcomes or value of the innovation. Here are some precepts for multiple measures in the formative evaluation of scale-up.

- Multiple measures of the desired outcome, divided into those that measure the construct in ways similar to the intervention's goals, and those that measure it through different formats and item pools.
- Criteria for success, tested under different conditions of stringency—that is, does it work when criteria are high and rigid as well as when they are flexible and variable.
- Measures of short-term effect and retention.
- Measures of transfer—performance on similar tasks and formats; performance with varying prompts (much vs. little information; timed and untimed; changed stimulus situations).
- Measures of implementation—observations, self-reports, student data, records, technology evidence.
- Measures of user reaction.
- Cost and efficiency data.

CRESST Assessment Models Scale-Up—Example 2. Although to be reported in depth elsewhere, CRESST's venture with the Los Angeles Unified School District (LAUSD) scale-up was extraordinary and replete with opportunities for us to learn. LAUSD educates approximately 780,000 students, more than the majority of states, in an area of approximately 500 square miles. During our involvement, California moved from frameworks, to no standards-no test, to standardized test-no standards, to variations of standards and test development, and now, to a steady state of standards and an evolution to quality standards-based assessment. In this period, the laws about bilingual education and testing in multiple languages changed, five superintendents were in office, class size reduction required massive day-to-day changes, and NCLB came into being. LAUSD went from believing it needed to have its own standards and measures to comply with the Improving America's Schools Act (1994) to full compliance with the state. LAUSD also was keen at the outset to develop assessments that could be used in classrooms in order to communicate clearly with teachers and the public the expectations that they should hold for their students. The highlights of the scale-up involved the following:

- Obtaining and maintaining political support for multiyear trials, involving school board, union, legislature, and administrative leadership.
- Developing standards in conjunction with the district and the Council for Basic Education.
- Mediating ideological differences in curriculum and content.
- Developing content-focused assessments and trying them out in multiple classrooms at three grade levels (elementary, middle, and high school) in history, science, language arts, and mathematics in Spanish and in English.
- Reducing focus to math and English, dropping Spanish versions and adding grade levels, and retaining performance or open-ended focus.
- Changing the purpose of the assessments from classroom improvement and system monitoring to promotion.
- Changing the number of test occasions per child and the total number of minutes allowed.
- Moving to literacy testing for 450,000 students a year.

- Working with 6,000 teachers to help them design and score assessments according to our models, conducting appropriate studies of training effectiveness.

- Transitioning from CRESST-trained to trainer-of-trainers scoring models, with varying levels of monitoring.

- Interacting with six different points of contact, many of whom were shifted as a result of changing superintendents.

- Preparing camera-ready versions of the tests and overseeing scoring and data entry.

- Conducting series of technical quality studies, involving modifications for different subgroups of learners, multiple formats, and experimental contrasts among instructional variations.

- Monitoring cost per student.

- Maintaining support of the teachers union and board, with the help of our "idea champion" in the district.

- Briefing new superintendents and their staffs on progress.

- Being involved in presentations that the district wanted, providing advice on other technical matters.

- Developing validity evidence using external examinations; for example, pass-rates on the California High School Exit Examination, generalizability studies, and evidence that domain-independent components could be embedded in different subject matter. This finding is of great importance as it reduces the cost of renewal and also supports the ability of teachers and students to engage in transfer.

- Bureaucratic issues, such as span of control, review of materials by the public and various committees, planning far out into the future, coping with real scale. I was asked to meet with the lead teachers to explain the progress of our work, and as an afterthought, I asked how many I would be seeing, imagining about 75 people at the outside. In fact, 1,000 lead teachers came to my presentation, a fact that rapidly changed my plans for an informal discussion and information-gathering session.

What happened, in a nutshell, is that for a small amount of money, we tested the scale-up of our research and Hawaii experience under conditions far more volatile and with far less control than we expected. It was hard, frustrating, with times of great satisfaction, and dismay. We learned that the assessment models are

robust across varying levels of design and scoring. We learned, from an independent study, that our assessments did in fact change teacher understanding of goals and promoted collaboration and the development of relevant performance tasks. We developed strong support among teachers and staff, some of whom were more invested than we were in the efforts. When the latest superintendent was appointed, he began the process anew with a different provider. So we scratched our heads and moved on. The timing for us was excellent as we were ready to commercialize and take the next step with technology.

Principle 5: Design the Best Study Possible

So what might count as evidence of an effective intervention for IERI? Let's consider just three options and some caveats.

1. The single, high-quality evaluation study. The most obvious is to conduct an evaluation study of an intervention, sampling settings and teachers and learners to which the findings should generalize. Here the unit of analysis should be the classroom average (of subgroups).

But—consider the tradeoff between sufficient treatment exposure and the need to disseminate something useful rapidly. How long should an evaluation last to show immediate and long-term impact?

At minimum, one should expect data to show that the intervention led to desired results compared to a criterion, or to well-defined alternatives, using dependent measures that are likely to be *sensitive* to change over the period of the study. A big flaw in the current application of experiments is the relative lack of credible and technically adequate appropriate dependent measure(s). Testing treatment and controls on poorly executed outcome measures yields no information. If appropriate measures are available, then niceties such as treatment verification, randomization, etc. should be observed, and mixed methods, including intensive substudies of groups or individuals, are desirable. Thus, this kind of study serves as an existence-proof—the intervention "worked" somewhere.

2. Extrapolating from cumulative results of scientific knowledge. One might analyze the *design* of the intervention to note the degree to which it is comprised of elements that exemplify best knowledge in a field (see Principle 2). Were teachers prepared with a range of examples to use? Were opportunities to engage in

significant intellectual activity provided in the program? Was useful feedback given? In this instance, one is using the history of stable scientific findings from series of studies to generalize to the new setting.

But—the analysis of instruction would have to be carefully conducted, by collecting information on use, as well as structure of materials. The view should be from the learner's eye. It is possible that the particular combination or the actual instantiation of heretofore effective variables might yield weaker-than-expected results. Nonetheless, this example emphasizes interventions built on application of accumulated knowledge rather than those dependent on a single, go/no-go study.

3. A good, one-shot case study with internal variations. A carefully implemented study that shows the differential effects of implementation conditions of the treatment (program) on multiple dependent measures. First, a range of implementations must be used so internal comparisons can be made (conditional differences, or even contrasts between strong and weak implementers) against the dependent measures of choice. Dependent measures (to harp on a theme) must minimally document that (a) they are sensitive to the highest strength instruction; (b) effects systematically show up on direct, transfer, and retention measures; and (c) the implementation supports (or has no negative impact on) measures of social capital (e.g., sense of efficacy of teachers and students, effort, teamwork, trust), as well as intellectual and skill outcomes. Longitudinal findings are always desired.

But—the comparisons on dependent measures assume that previous performance increments or trajectories (e.g., 2% a year) are known. Measures of instructional process (still not well formulated after all these years) would need to be obtained. This is not real theory building but rule-of-thumb guidance for practice.

In any case, the cost of such studies is an issue, as many researchers do not have the funds to translate their ideas into practical interventions, and many commercial entities have neither the inclination nor funds to defer marketing until a high-quality, perhaps multiyear study can be conducted. Our studies must recognize that effectiveness differences are necessary but not sufficient. The stability of impact and the acceptability of acquisition and maintenance costs (for material, required updates, and continuous training) will make a difference in the long-term impact of an intervention.

Principle 6: Support Flexibility and Local Innovation

The interesting tension in the adoption of innovation comes between the prestige, credibility, and cost of branded product lines and the "not invented here" (NIH, no federal relation) idea. While prestige, ease of use, and policy may push adoption of a particular approach; tradition and cost constraints operate to maintain current practices. My one time mentor, Susan Meyer Markle, shared a rule of implementation; just at the time people were first trying to design programmed instructional materials that were "teacher-proof," a time fast approaching from some directions again. Dr. Markle's advice was to encourage participants to "add their egg" to the mix, referring to cake mixes that required the cook to add a fresh egg (despite the fact that better effects were maintained by using a known amount of powdered egg). The reason was to help the cook feel more like a cook, by offering up an essential component. Similarly, in interventions, it is important to preserve avenues for local contribution and innovation. For the most part, these will not undermine the already effective programs. And more often than one realizes, the "egg addition" turns into a terrific recipe augmentation, much like the Pillsbury Bake-Off competition. With small effort and expense, through Web sites and Internet boards, it is possible to share interventions among users, and for the developer or marketer to identify potential variations for the next version.

Holding a place for local innovation and contribution shows more than tolerance of teachers "messing" with the intervention. It sets an ambience of welcomed exploration.

One place where local flexibility may best be shown is the manner in which the users move to "hands-off" practice by the R&D team. In the assumption that we intend to promote independence and accountability, it is important to learn from users how, in what order, and with what schedule of support they can begin to take on the responsibility for renewing the intervention on their own. In the CRESST Assessment Models area, we began this work in two places—at the administrative level, cooperating with the leadership to internalize the policies needed to use the system, how to train and vet raters, and what data flow should look like, and at the teacher level, working so that teams of teachers could design new assessments that met our specifications and could share with one another positive instructional experience and help for those having difficulty.

Principle 7: Minimums for Scalability and Sustainability

As an enterprise emphasizing multidisciplinary approaches and technology, IERI more recently has focused on the scalability and sustainability of the R&D theme. There is a series of questions that can be directed to scalable and sustainable education innovations (and who should implement them), but let me start by defining scalable and sustainable in my own terms.

Scalable:

- Multiple sites (districts) with full implementation
- Predicted effects across diverse settings at known cost
- Cost-sensitive and fully available training and technical support

Sustainable:

- Robust
- Survives without halo effect or incentives
- Needs minimal cheerleading
- Good results without the oversight of the developer
- Pliant and adaptive while maintaining key attributes and effects

But who should be responsible for such functions and for documenting their empirical studies? The school user who wishes to adopt the intervention? The researcher who wishes to explore, refine, and promote it? The commercial entity that wishes to do good, and to make money from it?

On a train to Lyon, a venture capitalist illustrated to me a now obvious truth—that the surest way to fail with the scale-up of an intervention is to let the inventor or research team continue to guide the implementation and marketing of the enterprise. Simplification and robustness are not always compatible with the particular vision a researcher brings to an application. What if we posited that commercial enterprises or specially conceived organizations should be responsible for evaluation, scale-up, and sustenance, and not the originating developer? If we believed that such organizations have distribution capacity, help, training,

marketing capacity, and general savvy not usually shared by researchers, how would that change our view of IERI's program direction?

Here are some propositions:

- It is not possible for IERI to stimulate sensible between-project comparisons of ways to scale up.

- Developing a theory of scale-up is best done by studying like enterprises in other fields.

- Theories of dissemination, too, come and go, and come and go again (e.g., remember agricultural extension agent), and their adequate testing conflicts with demand for solutions.

- IERI projects have not adopted outcome measures that can be aggregated to enable one to make between-project comparisons.

Furthermore, cycles of change—political, ideological, substantive and, particularly, technological—set the limits for scale-up and sustainability. An important question is, what is the shelf-life of the intervention? We are in an era of very rapid change. Good things, even the glimmer of good things, need to get out into the market.

If IERI wishes to support the development of private enterprise versions of funded work, it need only fund Small Business Innovation Research (SBIR) projects with existing procurement mechanisms. Another option is to review the policies of other agencies, for instance, the Office of Naval Research (ONR). ONR requires for the funding of its applied research projects that a sponsor (a particular part of the Navy, like the submarine service) will agree to take on and pay for the intervention once the research period is over and effects are demonstrated. ONR researchers and managers must spend time lining up these potential, serious, ultimately paying users or else the funding priority for project continuation is lowered. In education, we might think of commercial companies or states or districts as the target partner users.

In such cases, no doubt, the relationship between researchers and the private sector needs consideration, to avoid much of the messiness attendant in the science or medical areas about ownership, research credibility, and conflict of interest. But it is clear that if broad-scale sustainability is desired, such concerns must be designed in early. We can't grow research projects forward, in tiny steps, from pilot to larger

trials to big trials as we usually complete transfer from the laboratory to practice. If practice is the goal, then the improvement of practice should be the aim. What implication does this backward-chaining view have for research? Theoretically oriented research must continue, and a good proportion of it needs to occur in settings of potential use. But scale-up and sustainability are criteria that are applicable not to most research, but to development and engineering. And they need continued expertise to survive. Research dollars cannot be justified exclusively by whether current applications are scalable and effective. Such an approach is not future-oriented and does not take into account the real intellectual basis of research—exploring, hypothesizing, and developing potential for the future.

One approach for IERI is to support both theoretically oriented studies and integrated applications. It could explore how expertise is used to sustain and expand potential successes. Greater emphasis in the research/evaluation phases should be given to studying differences on various dependent measures, with different types of students and with volunteer and nonvolunteer educators. Scale-up and sustainability should be built in to those projects with the appropriate "track record" using partnerships with existing or developing commercial entities. Scale-up, to be serious, must involve the early participation of both users and potential commercial partners. In conclusion, the Principles recommended for consideration are repeated below. If there are some that are not on target, let's have future conversations on strengthening them.

Principle 1: Understand the Appropriateness of the Evidence for the Innovation

Principle 2: Document the Scientific Evidence in Support of the Design of the Intervention

Principle 3: Embrace Design and Development

Principle 4: Measure a Broad Range of Variables in Formative Evaluation

Principle 5: Design the Best Study Possible

Principle 6: Support Flexibility and Local Innovation

Principle 7: Minimums for Scalability and Sustainability

References

Alkin, M. C. (1985). *A guide for evaluation decision makers*. Beverly Hills, CA: Sage.

Baker, E. L. (1994). Learning-based assessments of history understanding. Special Issue *Educational Psychologist*, *29*(2), 97-106.

Baker, E. L. (1973). The technology of instructional development. In R. M. W. Travers (Ed.), *Second handbook of research on teaching* (pp. 245-285). Chicago: Rand McNally.

Baker, E. L., Freeman, M., & Clayton, S. (1991). Cognitive assessment of history for large-scale testing. In M. C. Wittrock & E. L. Baker (Eds.), *Testing and cognition* (pp. 131-153). Englewood Cliffs, NJ: Prentice-Hall.

Baker, E. L., Herman, J. L., & Gearhart, M. (1996). Does technology work in schools? Why evaluation cannot tell the full story. In C. Fisher, D. C. Dwyer, & K. Yocam (Eds.), *Education and technology: Reflections on computing in classrooms* (pp. 185-202). San Francisco: Jossey-Bass.

Baker, E. L., & Mayer, R. E. (1999, May/July). Computer-based assessment of problem solving. *Computers in Human Behavior*, *15*(3/4), 269-282.

Baker, E. L., Niemi, D., Herl, H., Aguirre-Muñoz, A., Staley, L., & Linn, R. L. (1996). *Report on the content area performance assessments (CAPA): A collaboration among the Hawaii Department of Education, the Center for Research on Evaluation, Standards, and Student Testing (CRESST) and the teachers and children of Hawaii* (Final Deliverable). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Bloom, B. S., Madaus, G. F., & Hastings, J. T. (1981). *Evaluation to improve learning*. New York: McGraw-Hill.

Boruch, R. F. (1976). On common contentions about randomized field experiments. In G. V. Glass (Ed.), *Evaluation studies review annual* (Vol. 1, pp. 158-194). Beverly Hills, CA: Sage.

Bransford, J. D., & Johnson, M. K. (1972). Contextual prerequisites for understanding: Some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior, 11,* 717-726.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research. In N. L. Gage (Ed.), *Handbook of research on teaching* (pp. 171-246). Chicago: Rand McNally.

Chi, M. T. H., Glaser, R., & Farr, M. (Eds.). (1988). *The nature of expertise.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Clark, D. L. (1985). Emerging paradigms in organizational theory and research. In Y. S. Lincoln (ed.), *Organizational theory and inquiry: The paradigm revolution* (pp. 43-78). Newbury Park, CA: Sage Publications.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston: Houghton Mifflin.

Cronbach, L. J., Ambron, S. R., Dornbusch, S. M., Hess, R. D., Hornik, R. C., Phillips, D. C., et al. (1981). *Toward reform of program evaluation.* San Francisco: Jossey-Bass.

Cronbach, L. J., & Suppes, P. (Eds.). (1969). *Research for tomorrow's schools: Disciplined inquiry for education.* Stanford, CA: National Academy of Education, Committee on Educational Research. New York: Macmillan.

Fisher, R. A., Sir (1951). *The design of experiments* (6th ed.). Edinburgh, UK: Oliver and Boyd.

Freeman, H. E., & Sherwood, C. C. (1970). *Social research and social policy*. Englewood Cliffs, NJ: Prentice-Hall.

Gage, N. L. (Ed.). (1963). *Handbook of research on teaching*. Chicago: Rand McNally.

Glass, G. V. (1977). Downtime. *Outlook*, *25*, 3-6.

Hively, W., Patterson, H. L., & Page, S. H. (1968). A "universe-defined" system of arithmetic achievement tests. *Journal of Educational Measurement*, *5*, 275-290.

House, E. R. (1974). *The politics of educational innovation*. Berkeley, CA: McCutchan.

Hovland, C. I., Lumsdaine, A. A., & Sheffield, F. D. (1949). *Experiments on mass communication.* Princeton, NJ: Princeton University Press.

Improving America's Schools Act of 1994, Pub. L. No. 103-382, 108 Stat. 3518 (1994).

Kaplan, A. (1964). *The conduct of inquiry. Methodology for behavioral science.* San Francisco: Chandler Publishing Co.

Larkin, J. H., McDermott, J., Simon, D. P., & Simon, H. A. (1980). Expert and novice performance in solving physics problems. *Science*, *208*, 1335-1342.

Lindblom, C. E., & Cohen, D. K. (1979). *Usable knowledge: Social science and social problem solving*. New Haven, CT: Yale University Press.

Lumsdaine, A. A. (1965). Assessing the effectiveness of instructional programs. In R. Glaser (Ed.), *Teaching machines and programmed learning. II: Data and directions* (pp. 267-320). Washington, DC: National Education Association of the United States.

Patton, M. Q. (1978). *Utilization-focused evaluation*. Beverly Hills, CA: Sage.

Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagne, & M. Scriven (Eds.), *Perspectives of curriculum evaluation* (American Educational Research Association Monograph Series on Curriculum Evaluation, No. 1, pp. 39-83). Chicago: Rand McNally.

Shavelson, R., & Towne, L. (Eds.). (2002). *Scientific research in education*. Committee on Scientific Principles for Education Research, Center for Education, Division of Behavioral and Social Sciences, National Research Council. Washington, DC: National Academy Press.

Stake, R. E. (1967). Toward a technology for the evaluation of educational programs. In R. W. Tyler, R. M. Gagne, & M. Scriven (Eds.), *Perspectives of curriculum evaluation* (American Educational Research Association Monograph Series on Curriculum Evaluation, No. 1, pp. 1-12). Chicago: Rand McNally.

Weiss, C. H. (1977). Research for policy's sake: The enlightenment function of social research. *Policy Analysis*, *3*, 531-545.

Wildavsky, A. (1993). *Speaking truth to power. The art and craft of policy analysis*. New Brunswick, NJ: Transaction Publishers.

Wittrock, M. C. (1974). Learning as a generative process. *Educational Psychologist, 11,* 87-95.

Wittrock, M. C., & Baker, E. L. (Eds.). (1991). *Testing and cognition.* Englewood Cliffs, NJ: Prentice-Hall.