Scaling Up Educational Interventions

Robert J. Sternberg

Damian Birney

Linda Jarvin

Alex Kirlik

Steven Stemler

Elena L. Grigorenko

Robert J. Sternberg

PACE Center

Yale University

Box 208358

New Haven, CT 06520-8358

Abstract

Some educational interventions successfully “scale up.” Others do not. Little—arguably, almost nothing—is known about the factors that lead to successful scaling up. The goal of this chapter is to identify a number of these factors through a disciplined and methodologically rigorous approach.

The difficulties associated with scaling up can broadly be summarized as falling into two classes: (1) difficulties associated with interventions (i.e., is a particular program suitable for upscaling?) and (2) difficulties associated with dissemination of the developed intervention (i.e., what geographical, economical, and human contexts are suitable for upscaling?).

There has not been a systematic review of the available knowledge, either at the level of theory or at the level of empirical evaluation of hypotheses and observations on the process of upscaling. This chapter attempts to carry out a systematic analysis of some of the factors contributing to scalability of a given educational program. The chapter does not claim to address all factors. However, it does attempt to capture and characterize at least some of the critical dimensions of scalability.
SCALING UP EDUCATIONAL INTERVENTIONS

What leads interventions successfully to scale up—to go from being small to being large? Some educational interventions successfully “scale up.” Others do not. Little—arguably, almost nothing— is known about the factors that lead to successful scaling up of educational interventions. The goal of this chapter is to identify a number of these factors through a disciplined and methodologically rigorous approach.

In this chapter, we use the terms scaling up and upscaling interchangeably. Correspondingly, the factors that facilitate or interfere with the process of upscaling are referred to as scalability factors.

This chapter is based upon the idea that Brunswik’s (1956) theory of probabilistic functionalism, and more specifically his notion of representativeness, provides a theoretical framework suited to meeting the goals of scaling up, and to addressing the problem of identifying the conditions under which educational interventions to improve pre-K–12 learning will succeed when applied on a broad scale. In order to understand scalability factors, it is necessary to conduct careful investigations, guided by the prescriptions of representativeness. These investigations should target both the characteristics of educational programs (i.e., the program aspects of scalability) and descriptions of environmental conditions (i.e., the dissemination factors of scalability) toward which educational interventions should be targeted.

Why is Scaling Up Hard to Do?

The problem of moving an educational intervention from one location to many locations is beginning to receive considerable attention in the research literature (e.g.,
Berends, Kirby, Naftel, & McKelvey, 2001; Elmore, 1996; Fullan, 2000; Ramey & Ramey, 1998; Slavin & Madden, 1996). In principle, it might seem that scaling up would be a simple, even trivial, task: One simply takes an intervention that has been used successfully on a small scale and applies it on a large scale. In practice, however, there are many difficulties associated with scaling up. These difficulties might be broadly summarized as falling into two classes: (1) difficulties associated with interventions (i.e., is a particular program suitable for upscaling?) and (2) difficulties associated with the social and human contexts in which interventions are implemented (i.e., what contexts in which the dissemination is to take place are suitable for upscaling?).

In the next sections, we review the literature supporting our distinction of the two types of scalability factors and provide a theoretical framework for interpreting issues of scaling up. In addition, we summarize some of our empirical observations related to issues of scalability collected in the course of life of an upscaling project funded by IERI. Thus, to provide the readers with the context for our empirical observations on upscaling, we start with the discussion of a theoretical framework for upscaling.

Theoretical Framework for Upscaling

One angle from which the scalability issue can be approached is that of scientific generalization (Brunswik, 1956). Reformulated in the context of educational interventions, the issue of scientific generalization can be stated as follows: What characteristics of a given educational program have a potential to transcend the context of discovery (i.e., the immediate context in which the program was developed) and to be validly implemented in contexts unlike the context of discovery? The hypothesis of the existence of such characteristics originates from practices of natural sciences, where there
are universal laws applicable to all (or almost all) contexts. However, a growing apprehension exists in the field of educational sciences that many cognitive and social phenomena may never adequately yield to a description or explanation in terms of “universal” characteristics or laws. For example, as a result of extensive research indicating the importance of phonological skills in mastery of reading, many reading-intervention programs based on teaching phonemic skills have been developed. However, not all children respond equally well to the direct-teaching phonics-based approach—some other interventions are needed to address the needs of a diverse group of children with reading difficulties. (For discussion of this issue, see, for example, Blachman, 1997.)

Because we lack universal laws, the task of assessing, or attempting to ensure, the generalizability of educational-research products will have to be based instead on a logic of representativeness. An example of such a task would be determining whether a newly proposed educational intervention emerging from a psychological laboratory or a specific limited educational setting will scale up to the realities of the modern educational environment.

The term representativeness was introduced into discussions of psychological methodology by Brunswik (1956) to indicate the level of similarity between the context in which a scientific investigation is performed and the target class of contexts toward which research products are intended to generalize. As such, the concept of representativeness embraces the notion that outcomes of any learning process are highly dependent on the context in which this learning unfolds (e.g., Sternberg & Wagner, 1994). At the same time, however, the concept of representativeness provides a logical basis for assessing and assuring generalizability, and thus, it plays an important role both
in our general theoretical approach and in our specific research plan for contributing knowledge to better enable the scaling-up of educational interventions.

Thus, a functionalist program focuses on principles of adaptation to particular environments, rather than on context-free mechanisms. Consequently, the corresponding methodology needs to include characteristics of educational interventions, students to whom these interventions are administered, and the environmental conditions under which these programs are administered (Hammond & Stewart, 2001). Because the environments in which schooling and learning take place vary and because schooling is “applied” to all kinds of children at different levels of abilities and of different backgrounds, the generalizations should be based on statistical sampling of the environments, as well as on statistical sampling of participants in the process of schooling. The latter is common today: Educational studies routinely employ large numbers of participants to ensure reliable generalization of results beyond the particular participants studied to the wider population to which the investigator intends to generalize his or her results.

What is still rather rare is statistical sampling of the environmental conditions toward which generalization is intended. Just as in participant sampling, a crucial factor determining the success of a representative design is accurately specifying and describing the properties of the environmental (in the context of this chapter, “educational”) settings toward which the research results are intended to apply (or again, “to scale up”). As with participant sampling, the specification of the target environments toward which generalization is desired should be performed prior to the performing a study. Target environmental contexts can be extremely variable.
As evidenced by our framing of upscaling in terms of the more general issue of statistical generalization, it should not come as a surprise that we are not the first to note the need to sample diverse contexts in educational science. More than sixty years ago, for example, in his text on statistical methods for educational research, Lindquist suggested that interventions be evaluated using a random selection of schools (Lindquist, 1940). Nor is the upscaling problem specific to education, a fact opening doors for educational scientists to draw upon lessons learned by those performing practically-relevant research in neighboring psychological disciplines. Successful upscaling is a challenge any time theoretically-motivated interventions must be applied in specific contexts that were not themselves the primary empirical foundation for theory. Consider, for example, what the human factors pioneer Alphonse Chapanis had to say on this point:

There are two ways one can go about doing studies that will extrapolate to a wide variety of situations. The first is by deliberately building heterogeneity into studies, a tactic recommended by Brunswik (1956) over 30 years ago; the second is by replication. (Chapanis, 1988, 253-267).

As scientists striving for robustness and generality in our educational research, we are clearly not alone in embracing Brunswik’s method of representative design as a basis for the design of effective psychological interventions. We believe achieving the goal of generality or robustness is best achieved by iteratively creating, testing, modifying, and retesting theory in heterogeneous contexts.

Representative design solves the problem of generalization by requiring that the integrity and heterogeneity of contextual variables in the target, educational environment be preserved in one’s fundamental research. To the extent that our discussion and
examples persuade others to join us in this venture, we note that representative design places one additional requirement on research design and communication. This requirement concerns the need to treat the educational ecology as an object of scientific study on a par with internal, psychological activities or processes.

Embracing this aspect of representative design requires an appreciation of Brunswik’s deep understanding of the inherently social nature of scientific conduct and progress. In this light, Brunswik (1956) noted that truly cumulative knowledge about the contextual influences on cognition and behavior will be achievable only if those conducting and reporting research provide descriptions or theories of their research contexts that are just as precise and detailed as the descriptions or theories psychologists have learned to provide of phenomena such as learning, intelligence, and the like. By requiring this type of “equal treatment” for theory of both the internal and contextual aspects of cognition and behavior, empirical findings can bear, not only on the truth or utility of theory of phenomena such as learning and intelligence, but also on the truth or utility of theory of the contextual influences on these activities.

Brunswik’s theory provides a framework for research on scaling up, but does not explicitly specify the factors involved in scaling up. What are these factors?

Scalability Factors

Characteristics of the Program as the Scalability Factors

With regard to the first class of factors, there are a number of relevant issues. The first factor is obvious—the intervention itself should be of high quality. In other words, the program should have been proven to address specific aspects of educational process in such a way that there are identifiable and measurable outcomes distinctly illustrating
the impact of the intervention as compared to other intervention. Moreover, Elmore (1996) highlights a number of other key features of programs suitable for successful upscaling; these key features are suitability of the program for the present levels of the personnel, personnel training, and scaffolding during initial implementation. Here we describe and expand Elmore’s ideas.

First and foremost, the program needs to be one that teachers can cope with, given their present competencies and skills, and that teachers can implement with an amount of in-service instruction that is reasonable, given the other demands made upon them. Second, the program must create realistic expectations about the amount of time required from teachers and principals related to mastering and delivering the program. Programs often work on a small scale due to adoption by highly motivated individuals. Moreover, in the context of small-scale implementations, a key factor is often the proximity of the creators of the program to its first adopters and implementers. This proximity implies not only the physical proximity, but also the proximity of ideas and beliefs—those educators who are willing to try new programs usually see a value in doing so and thus are more predisposed to raise the odds of the program to succeed. Therefore, programs may work on a small scale, but they fail when they are upscaled because the initial sample of the program’s deliverers was not representative of the larger population. To maximize the likelihood of the success of the program, the personnel training should include not only a skill-enhancement component, but also a motivational component. Moreover, the program must provide sufficient supplemental materials for use as scaffolding for teachers navigating and delivering the program for the first time.
What empirical factors make successful upscaling a challenge? Here we discuss some of them briefly.

**Heterogeneity of content and skills standards across states, districts, and schools.**

In the United States, there is a lot of variability in standards at the levels of states, districts, and schools. Therefore, prior to undertaking the task of scaling up a given intervention, it is important to determine the degree of correspondence between the objective of the program and a set of standards—where do the program’s pedagogical objectives fit the best, and at what level? Clearly, the better the fit, the more successful will be the dissemination. It is possible that a given program has been tailored for the educational standards promoted by a certain district; therefore, it might not be fruitful to try to disseminate the program more broadly than to the schools within that district. Alternatively, it is possible that the unit of dissemination, given the educational standards of the program, should be the state; in this case, unless the educational standards of the program are modified, its dissemination at the country level might not be successful.

In our own work, we have developed pedagogical materials targeting the National Standards, since we wanted to operate at the level of the country. For example, in our current project funded by IERI (Interagency Educational Research Initiative), which is an upscaling of our earlier work, we have created grade 4 curriculum materials in three subject areas—language arts (5 units: true wonders, wonders of nature “pourquoi, journeys, biographies, mystery), mathematics (5 units: number sense and place value, equivalent fractions, measurement, geometry, data analysis and representation), and science (4 units: nature of light, magnetism, electricity, ecology). In making the decisions about content, the curriculum development team researched state, national, and local
curriculum frameworks, but always focused on the national interpretations of the educational standards. For example, in language arts, each unit targeted one or more of *The International Reading Associate/National Council of Teachers of English National Standards for the English Language Arts*. In mathematics, unit content decisions were driven by the National Council of Teachers of Mathematics *Principles and Standards for School Mathematics* (NCTM, 2000). In science, each unit targeted one or more of the National Science Education Standards (1996) and Benchmarks for Science Literacy for grades 3-5.

*Heterogeneity of students’ ability levels across and within schools.* This scalability factor is related to the issue of extreme variability in student achievement and needs encountered in the U.S. public-school system. Even within our relatively small home state of Connecticut, districts are extremely variable in terms of their levels of student achievement. Specifically, in Connecticut, the discrepancies in performance are great between the highest- and lowest-performing districts. Clearly, the gap is even wider if the country is considered as a whole. Such gaps stress the necessity of taking into account the heterogeneity of the student body when attempts are made to develop scalable intervention programs. In our experience, this issue can be addressed by supplementing teachers with differential pedagogical activities that address the needs of low-, average-, and high-achieving students (see [www.yale.edu/pace](http://www.yale.edu/pace) for specific illustrations).

*Heterogeneity of teachers’ skills.* Throughout various programs developed at and implemented through the Center for the Psychology of Abilities, Competencies, and Expertise (PACE Center) at Yale within the last 3 years, we have interacted with over
1,000 teachers in diverse regions of the United States. In our work, we have encountered a large degree of variability in teachers’ pedagogical skills and content knowledge. Therefore, we consider the development of teacher scaffolding materials and training to be an essential scalability factor.

Our approach to the development of the pedagogical materials is such that they are extremely detailed, but always offer the teacher a certain amount of freedom in their interpretation and choice of particular activities. Specifically, for our current IERI project, multi-unit materials that have been developed and distributed include a teacher guide containing instructional material for each unit (approximately 5-8 lessons spanning 50-70 pages), background information, resource materials reflecting print and non-print sources, and student workbooks (see www.yale.edu/pace for details). In addition to detailed pedagogical materials, a 12-hour in-service training program was developed and has been delivered to over 300 teachers across the nation. Day 1 of the teacher training focuses on the theoretical principles of teaching and instruction for each one of the three modes of instruction that serve as the conditions. The second day focuses on modeling the units in each subject area and provides teachers with an opportunity for hands-on experience with the unit format.

*Embedded accountability of student progress.* Finally, one more scalability factor that is relevant here is that of the capacity of the program to track the progress of the participating student on an ongoing basis. The ultimate goal of any successful pedagogical intervention is to deepen and broaden the knowledge base of the students. This knowledge is captured in part in state- or district-administered standardized achievement tests. Therefore, we try to track changes in test scores over time as a
function of our interventions. In addition to standardized tests, we also use homemade tests that are more specifically tailored to our interventions. In our work, the delivery of each pedagogical unit is preceded by the administration of a pre-test and followed by the administration of a post-test. In each one of the three subject areas, we developed content-specific pre- and post-tests to be administered before and after the implementation of each curricular unit. These assessments include multiple-choice and open-ended questions, the latter of which are scored by specifically trained raters. Students’ performance at the pre-test and post-test across the three design groups are compared using hierarchical linear modeling methods. Thus, student baseline achievement is assessed and student and teacher motivational data for the administration of a particular unit are collected. This approach, clearly, creates much data. For example, in the first year of our IERI program, 82 teachers were recruited from 36 schools; approximately 8,500 student assessments were collected from 1,770 students. In the second year, 82 teachers were recruited from 42 schools and approximately 12,000 student assessments were collected from 4,385 students. The third year of data collection is underway as we prepare this chapter and 118 teachers from 36 schools have so far participated. Although the amount of monitoring and data processing related to the introduction of program-internal assessment is large, we think that these efforts are necessary for understanding the impact of a particular pedagogical program on a country-wide scale.

*Characteristics of the Dissemination Context as Scalability Factors*

With regard to the second class of scalability factors (the dissemination context factors), there are also several groups of issues. The first issue has to do with getting
people to adopt the program (i.e., the diffusion of the innovation). The second issue relates to the degree to which the program is successfully implemented. The issue of adoption is relevant to the problem of upscaling because no intervention ever can be upscaled if educators do not “buy” into it. Of relevance here are Rogers’s (1995) Diffusion of Innovations and Ely’s (1976) Conditions of Change models of educational change. The Diffusion of Innovations framework emphasizes the impact of characteristics of the innovation on the rate of adoption, whereas the Conditions of Change model calls attention to circumstances that predispose an environment toward change.

However, even reasonably good educational interventions might fail the upscaling test if educators did not commit themselves to implementing them (Elmore, 1996). For example, there is some research suggesting that, even during peak reform periods, only about 25 percent of teachers are interested in experimenting with reform efforts (Cuban, 1990; Elmore, 1996). Thus, even a good program might fail to upscale successfully as a result of the resistance of those who are expected to carry out scaling up.

It is often assumed that a good program or a great idea will sell itself. Perhaps one of the major breakdowns in going to scale comes from the fact that good programs have failed to address the critical need to disseminate their findings in a way that communicates effectively with educators. In other words, innovators need to know what educators, who make the critical decisions with regard to bringing innovations to school, consider while these decisions are made. These aspects of the “consumer mentality” of educators in adopting specific interventions have not been well studied. The second issue, the issue of implementation, is also relevant to the problem of upscaling. Even when a
decision to adopt successful interventions is made, adopting institutions have to consider all kinds of factors (e.g., institutional and individual competence of the program implementers; social and structural conditions of implementations) that influence upscaling (Fullan, 1982; Fullan & Miles, 1992).

In addition, there are a number of themes developed in the organizational-psychology literature that link adoption and implementation (Armenakis & Bedeian, 1999; Brink et al., 1995; Goodman & Stecker, 1989; Goodman, Stecker, Hoover, & Schwartz, 1993). These four research themes are (1) content, which pertains to the substance of change; (2) context, which involves existing forces or conditions in an organization’s external and internal environments; (3) process, which addresses actions undertaken during enactment of intended change; and (4) criterion, which deals with outcomes or markers for tracking the likelihood that necessary behaviors are enacted to achieve the desired changes. A number of ideas from the organizational-change literature have been successfully utilized in studies of scaling-up in schools (e.g., Blumenfeld et al., 2000; McDougal et al., 2000).

Although fragmented and infrequent, there are empirical studies addressing the significance of context scalability factors. Specifically, Veir (1990) used stepwise regression analysis to identify the key context variables for implementing staff-development programs in rural schools. Eight predictor variables were found to be important in explaining whether the program would be successfully implemented in schools. The variables were training time; socioeconomic profile of the student body; administrative participation; proximity to an institution of higher education; provision of
incentives; number of high-school-level teachers; number of administrators in the district; and the presence of a trainer from a higher educational institution.

Needless to say, we have also experienced the importance of the context scalability factors in our own work. We define an educational district as a unit of adoption of our programs. Over the period of four years of our current IERI project, we have worked with many districts in the United States. We have observed that the degree of success of our programs varied in different districts depending, broadly speaking, on four factors: available resources, working environments in the districts, commitment of the district leadership to the innovation, readiness to change, and the level of organization experience among both teachers and administrators.

Available resources. US educational districts differ widely in terms of the resources available for their everyday functioning and adoption of new programs. Thus, in preparing a program for scaling up, a developer should detail all expenses that will be encountered by a district and inform the district of the costs and the demand on the resources associated with the innovation.

District working environments. In our experience, the success of program adoption is extremely sensitive to characteristics of working environments within a district. For example, we have worked with districts in which there were open confrontations between the administration and the teacher unions. Clearly, these environments are not ideally suited for adoption of innovation. Moreover, we have found it necessary to work with district administrations at multiple levels: superintendents or assistant superintendents, curriculum specialists, principals, teachers, and so forth. The point here is that, since the average duration on a job for superintendents in the United
States is short, so innovators need to be sure that the program is “bought into” at multiple levels in the district, so that when the superintendent is transferred to another job or fired by the Board of Education, there is still a commitment to a particular program within a district.

Types of commitment. We have observed that districts agree to adopt educational programs for a variety of different reasons. This variability in motivation produces different types of district attitudes toward an innovation. We distinguish four types of these attitudes: circumstantial, permitting, promoting, and committed.

Districts with circumstantial attitudes toward innovations simply allow the program to be “tried out” by their teachers, but do not motivate the teachers or do not publicize the intervention. A circumstantial attitude toward intervention usually results from a combination of factors that lead a district to adopt a particular program when, in fact, this program could have been some other program.

Districts with permitting attitudes allow an intervention program to enter the district with a certain amount of caution and consider the program as a possibility among other potential alternatives. This attitude, from our experience, arises when a district is “shopping” for a new pedagogical program, but cannot specify the detailed characteristics of a program it is interested in.

The promoting attitude is characteristic of districts that work with the program developers and teachers to ensure the maximally efficient use of effort, and, correspondingly, success on both sides. These districts truly wish to ensure the success of whatever program they adopt. The districts are not necessarily committed to any particular program, however.
Finally, the *committed* attitude characterizes the district in which the administration and the teachers are committed to a particular program and seek out its further development in their district. Such districts not only adopt a program, but also contribute to the program’s adaptation in the context of particular needs of the district.

*Readiness to change.* In our work, we have found that the district’s attitude toward an innovation does not guarantee the success of the innovation in the district. In other words, we have found that the district’s readiness to change in implementing an innovation does not necessarily correspond to its attitude toward the program per se. We have worked with districts that demonstrated promoting attitudes towards our program, but were not ready to change by meeting the requirements imposed by the intervention (i.e., in allocating enough teaching time and district’s resources for the program to be implemented as designed). There are many different kinds and levels of readiness to change (Sternberg, 2000). In particular, what seem to matter are not only a school’s readiness to change, but also its readiness to *appear* to change and its perception of itself as modifiable.

To summarize, our experience in upscaling an educational program allowed us to identify at least some scalability factors that appear to be generalizable to attempts of other educators. These factors need to be considered before a researcher or district attempts to upscale.

Our own current IERI-funded project upscales our earlier work (Grigorenko, Jarvin, & Sternberg, 2002; Sternberg, Grigorenko, Ferrari, & Clinkenbeard, 1999; Sternberg, Torff, & Grigorenko, 1998). As mentioned earlier, we have created grade 4 curriculum materials in language arts, mathematics, and science. The materials were
designed for presentation in one of three instructional modes: creative-analytical-practical (CAP—based on the theory of successful intelligence—Sternberg, 1997, 1999) critical thinking (CT), and memory (M). Over the period of the project, we intend to work with approximately 15,000 students in diverse locations of the United States.

In terms of scaling up an intervention within an empirical research program, we have shown that triarchic teaching can easily be implemented in classrooms across subject matter areas, community types, socio-economic levels, and for a variety of student populations. We have developed an easily manageable Microsoft ACCESS-based database and a procedure manual for managing large-scale data collection. All the data collected from students and teachers arrives at the PACE Center in paper format and must be entered in electronic databases and then filed. The flexibility of the electronic database that we have developed allows easy monitoring of the large volume of data as it goes through the various stages of processing from initial logging, to the assignment of tests to raters for scoring, to data entry and analyses. The design permits easy access to the data to everyone who is interested in secondary data analyses and data processing. The key personnel working on this project have fostered contacts with numerous school districts. A total of 3,270 school districts across the United States have been contacted about the program. In order to maintain contact and sustain participation with a large number of geographically distant participants, we have also created a password-protected website specifically for this program that enables teachers to upload and download materials, chat with other participants and curriculum specialists, and contact our subject-area experts.

Conclusion
By making contextual descriptions and theories explicit, the educational science community as a whole can participate in what Brunswik (1956) called the large, concerted, group project that is necessary to realize the promised fruit of representative design. For educational science, this fruit would be a detailed understanding of what contextual factors matter in upscaling and applying interventions, what factors tend not to matter, and what factors matter in what ways. Naturally, these factors may be intervention-specific, with the result that what is learned through research is which types of interventions work in which types of contexts.

When the contextual theories and descriptions underlying educational research are explicitly stated and communicated, and findings shared, questions about the robustness and generality of those findings become empirically decidable, and thereby scientific. One researcher, for example, may not find a particular effect in an educational environment in which a theory’s contextual components suggest she should find one. This evidence can then be used to revise the contextual theory, growing our collective understanding of the contextual variables that matter. We encourage others to join us in this large, concerted, group project, as we believe it to be the most promising path toward improving the scalability of educational interventions.
Author Notes

Preparation of this article was supported by Grant REC-9979843 from the National Science Foundation. Grantees undertaking such projects are encouraged to express freely their professional judgment. We thank Cynthia Matthew for review of the relevant organizational psychology literature. This article, therefore, does not necessarily represent the positions or the policies of the U.S. government, and no official endorsement should be inferred.
References


