Design for Defenestration: A Strategy for Scaling Up Promising Research-Based Innovations


Research as a Means to Scalability

This paper describes a research strategy for scaling up promising innovations from the fertile, greenhouse environments in which they were conceived to the barren contexts of schools with poor facilities, overwhelmed teachers, and struggling students. Adapting a locally successful innovation to a wide variety of settings—while retaining its effectiveness, affordability, and sustainability—is very challenging (Dede, Honan, & Peters, in press). In contrast to experiences in other sectors of society, scaling up successful programs has proven very difficult in education. For example, in the business sector, insights from changing operations at one fast-food location (e.g., implementing a faster system for processing customer orders) may easily transfer to every store in that franchise and perhaps to any comparable type of restaurant. However, in education a new type of teaching strategy that is successful with one practitioner often is difficult to generalize even to other instructors in the same school, let alone to a broad range of practitioners. When it comes to education, one size does not fit all, so strategies for adaptation and transfer are necessarily complex.

In making judgments about scalability of an intervention, differentiating the intervention’s design from its “conditions for success” is important (Dede, in press). For instance, the effective use of antibiotics illustrates the concept of “conditions for success”: Antibiotics are a powerful “design,” but worshiping the vial that holds them or rubbing the ground-up pills all over one’s body or taking all the pills at once are ineffective strategies for usage—only administering pills at specified intervals works as an implementation strategy. A
huge challenge educators face, and one of the reasons this field makes slower progress than venues like medicine, is the complexity of conditions for success—and the sophistication of the processes necessary to achieve these conditions—required in effective interventions. Nothing powerful in facilitating learning is as simple or as easily administered as an inoculation in medicine.

Research findings typically show substantial influence of contextual variables in shaping the desirability, practicality, and effectiveness of designs. For example, scholarly studies on innovations often describe “conditions for success” challenges related to teacher professional development. Resolving this implementation issue presents choices about various approaches to the iterative evolution of a design. Alternative strategies include changing the design so that the intervention is more “teacher-proof,” expanding the design so that extensive teacher professional development is now part of the “treatment,” or abandoning the design as unpromising because its effective use requires a level of knowledge and skill likely unattainable in the typical teaching population, due to current methods of recruitment and training. This is not an easy dilemma to resolve and illustrates a scalability issue of great interest to practitioners and policymakers.

This paper discusses strategies for scaling up educational innovations to relatively inhospitable settings characteristic of schools in general, as opposed to altering schools to make them more conducive contexts for implementing a particular design. This concept of scalability stands in contrast to a frequently used method of disseminating educational innovations: i.e., altering schools to make them more conducive contexts for implementing a particular design. For example, many scholars seek to involve teachers participating in the project as co-designers or co-researchers. However, expanding the size of the research team is not a strategy that can accomplish true scalability. Similarly, the scholarly literature on innovation shows frequent
“design creep” (e.g., a curriculum intervention escalating into a full-scale systemic reform initiative), with investigators responding to every implementation difficulty with increasingly more sweeping designs rather than providing bounded research on a particular type of potential advance. But, turning every design into a systemic reform so that the context is hospitable to an innovation is also not a feasible means to true scalability. True scalability requires developing interventions that retain substantial effectiveness in relatively inhospitable settings, such as urban schools, in which an innovation’s conditions for success (e.g., supportive administration, qualified and enthusiastic teachers, a well maintained technology infrastructure, a student population consistently present) may be absent or greatly attenuated.

Under circumstances without these criteria, major intended aspects of an innovation’s design may not be enacted as planned (Means & Penuel, in press); developers can expect parts of their design to be “defenestrated” (thrown out the window). Planning for successful implementation in such contexts involves design akin to the “egg drop” experiment that is part of many science curricula. Students are given raw eggs and a few basic materials, such as dry pasta or pipe cleaners. The learners are asked to construct some sort of “packaging” for an egg that will cushion it from breakage, even when dropped from a considerable height. Researchers similarly attempt to develop aspects of their design package that help its effectiveness to survive even when parts of its intended enactment are defenestrated. This is not an easy task; oftentimes, scholars fail to identify the key features that lead to small-scale successes, resulting in failure when their recommended adaptation-and-transfer strategies are implemented large-scale.

A Case Study of Design-Based Research Evolving Towards Scalability

To illustrate the challenges of designing for scalability, this paper discusses, as a case study, research my colleagues and I are doing to prepare an effective innovation for use in
educational settings lacking some of its conditions for success. With National Science Foundation funding, we are creating and studying graphical multi-user virtual environments (MUVEs) that enhance middle school students' motivation and learning about science and society (Dede, Nelson, Ketelhut, Clarke, & Bowman, 2004).

**The Design to Evolve towards Scalability**

Our “River City” MUVE is centered on higher order inquiry skills such as hypothesis formation and experimental design, as well as on content related to national standards and assessments in biology and ecology. We are documenting how students can gain this knowledge through immersive simulations, interaction with digitized museum artifacts, and "participatory" historical situations. Students learn to behave as scientists through collaboratively identifying problems through observation and inference, forming and testing hypotheses, and deducing evidence-based conclusions about underlying causes. Our goal is to promote learning for all students, particularly those unengaged or low performing.

*Figure 1: Talking with an Agent  Figure 2: Collecting Water Quality Data*

The River City virtual “world” consists of a city with a river running through it; different forms of terrain that influence water runoff; and various neighborhoods, industries, and institutions, such as a hospital and a university (http://muve.gse.harvard.edu/muvees2003/). The learners themselves populate the city, along with computer-based agents, digital objects that can include audio or video clips, and the avatars of instructors (Figure 1). River City is typical of the
United States in the late nineteenth century; the right hand window in Figure 1 depicts how we use museum artifacts to illustrate building exteriors and street scenes from that period in history. In addition, throughout the world, students encounter residents of River City and “overhear” their conversations with one another. These computer-based “agents” disclose information and provide indirect clues about what is going on in River City.

Content in the right-hand interface-window shifts based on what the participant encounters or activates in the virtual environment (Figure 2). In this case, the right hand window presents water quality data from one of eleven water-sampling stations in River City. Through data gathering, students observe the patterns that emerge and wrestle with questions such as “Why are many more poor people getting sick than rich people?” Multiple causal factors are involved, including polluted water runoff to low-lying areas, insect vectors in swampy areas, overcrowding, and the cost of access to medical care.

Multiple teams of students can access the MUVE simultaneously, each individual manipulating an avatar through his or her computer. In our implementations, the class is divided into teams of two to four students, who are "sent back in time" to this virtual environment. The lab notebook that the student teams use asks the class to help the city solve its environmental and health problems, which are directly related to middle school science content. To accomplish this, the students must collaborate to share the data each team collects. Beyond textual conversation, students can project to each other "snapshots" of their current individual point of view (when someone has discovered an item of general interest) and also can "teleport" to join anyone on their team for joint investigation.

Students work in teams to develop hypotheses regarding one of three strands of illness in River City (water-borne, air-borne, and insect-borne). These three disease strands are integrated
with historical, social and geographical content to allow students to experience the realities of disentangling multi-causal problems embedded within a complex environment. Each time a team reenters the world, several months of time have passed in River City, so that learners can track the dynamic evolution of local problems over time and through various seasons. During their time in the MUVE, students answer questions in a Lab Notebook, which starts with questions that guide exploration of the environment and develop mastery of the interface, building towards later investigations that are content specific and require completing a data table based on the water samples encountered in River City.

Finally, students write a letter to the mayor of River City describing the health and environmental problems they have encountered and making suggestions for improving the life of the inhabitants. The Lab Notebooks provide a familiar way for teachers to assess student progress through the unit and their ultimate learning of the science content. In our research on River City, we are studying usability, student motivation and self-efficacy, student learning, and classroom implementation issues.

**Selecting a Design Promising for Scalability**

In spring, 2004, we conducted large-scale implementations with more than 1000 students in Boston- and Milwaukee-area classrooms, with high proportions of ESL and free-and-reduced-lunch students. The findings thus far from this set of implementations are promising, but our data analysis is far from complete. However, preliminary analysis suggests that students and teachers were highly engaged. For example, during the curriculum unit students attendance (typically around 50% in some of our sites) improved significantly, and disruptive behavior in classrooms dropped.

Quantitative analysis of pre/post measures shows that that--in comparison to a similar,
but non-computer-based curriculum—the River City treatment was successful in teaching students biology content related to epidemiology, significantly raising improvement scores from 16% for the control curriculum to 32-35% for the River City treatments. Also, case study analysis of student logfiles indicates that learners developed 21st century skills in virtual communication, information analysis, and problem solving (Partnership for 21st Century Skills, 2003).

Moreover, after designing and conducting their experiments, teams of students in both the control condition and the River City treatment were asked to write letters to the Mayor of River City, in which they discussed their hypothesis, experiments, and recommendations for solving the city’s health problem. Preliminary analysis of these letters suggests that many students demonstrated an understanding of the process of the scientific method beyond what was captured in the science inquiry post-test measures. For example, some learners who scored low on the science inquiry post-test wrote letters that were of similar quality to those written by students who scored substantially higher on the post-test. We are finding that the complexity of the MUVE treatment creates intricate patterns of learning possibly more appropriately measured with an authentic activity, such as writing an experimental report, than with simpler, more easily quantifiable assessment measures.

Analysis of affective changes in our student population also shows interesting patterns. Preliminary results indicate that self-efficacy in science inquiry improves as much as 15% over the two week implementation for students who see themselves as less competent in this area. This is important, because self-efficacy changes can be a vicious cycle for those with low self-efficacy, as they are less likely to push for the academic success needed to affect their self-
efficacy positively (Bandura, 1977; Ketelhut, draft). Other interesting patterns are emerging about different types of students who did well under various pedagogical conditions.

We are continuing large-scale studies to assess the strengths and limits of this educational approach and to evolve our design in ways that will make it more effective and scalable. Ultimately, to make this innovation worth scaling up, our findings must show evidence of addressing an important educational problem with an intervention that has a large “effect size” (Thalheimer & Cook, 2002) and is affordable and sustainable (Dede, op. cit.). In other types of research, scholars sometimes value statistical validation over sizable effect. Investigators report findings that reach the 0.05 or better level of statistical “significance,” meaning their results have a less than a one in twenty chance of being due to random effects. However, at times the findings themselves are trivial, showing only a small impact on educational effectiveness from an intervention that consumes much time and resources. Practitioners and policymakers, in contrast, have greater interest in findings that reveal large effect sizes, backed by plausible evidence of likely causation -- even if the statistical significance of these results might be difficult to measure or below the typical standard of scholarly proof.

**Understanding How Conditions for Success Influence a Design’s Scalability**

As we evolve our design towards scalability, some conditions for success are essential given the nature of the innovation. For example, to use our curriculum teachers must have access to one reasonably modern, Internet-connected Windows computer per student for six forty-five minute class sessions over an approximately two week period. They must be able to implement our curriculum in the order provided (a specific mix of on-computer and in-classroom days) and to track student absenteeism. Teachers must be comfortable in their classroom pedagogy supporting individual and small team student investigation without providing direct
answers. The school administration and technology coordinator must support the use of this innovation as a deviation from typical practice. Enumerating these essential conditions for success is important because, the fewer and more attainable such conditions are, the wider the range of contexts to which an innovation can scale without losing its effectiveness.

Other conditions for success of our design are important, but not absolutely necessary. For example, teacher participation in the eight hours of online professional development we have prepared is very useful. Through experience from our implementations, we know that educators require only a little training in order to use the MUVE we have developed, and that most teachers need just a quick review of the science concepts and skills our curriculum conveys. However, a teacher’s ability to facilitate interpretive small group and whole class discussions is important to the effectiveness of our design. We find many teachers, particularly in urban settings, have few skills in this because they are accustomed to spending all their time in motivating learners, presenting foundational information, making up learning deficits, and managing behavior problems (due largely to students who are not interested in traditional “school” and are overwhelmed by conditions in the rest of their lives).

Our MUVE accomplishes these objectives for teachers, making our innovation very conducive to scaling up because our intervention greatly reduces problems that hamstring many other pedagogical and curricular approaches. On the other hand, our innovation asks teachers to utilize a type of higher-order pedagogy important to its effectiveness that few instructors have experience enacting. Participating in our online professional development helps teachers to develop this skill and so is an important, but not essential, condition for success.

While requiring large implementations to generate substantial statistical power, using multi-level modeling as an analytic methodology enables us to understand the effect on our
design of attenuating non-essential conditions for success, such as teacher professional
development, which occur at the classroom or the school level. In our analyses, we account
separately for the clustering of children within classrooms using generalized least-squares (GLS)
regression analysis (a.k.a., multilevel modeling) in which measures of children’s learning are
regressed on a system of dichotomous “question” predictors that represent children’s
membership in various settings.

In these GLS models, since the number of classrooms is much smaller than the number of
students, we account for the clustering of learners within classrooms by the method of fixed
effects, in which a system of dichotomous (dummy) classroom predictors are introduced into the
GLS models as control variables to account for all possible unobserved classroom-level
differences. Naturalistic differences across the many implementation sites intrinsically lead to
variations in our conditions for success (e.g., some teachers do the online professional
development, others do not). Through analysis of the dichotomous classroom predictors, we can
assess the impact of attenuating a particular condition for success on the effectiveness of our
design intervention.

Evolving the Design for Scalability

Evolving a design for scalability to contexts in which its non-essential conditions for
success are attenuated or lacking requires developing aspects of the design that enhance the
robustness of its effectiveness even when parts of its intended enactment are defenestrated. For
example, in some of our implementations, we found that a few teachers ignored all or most of the
professional development we made available online. These teachers then typically encountered
the following problems in implementation:
• They did not understand in depth the purpose and process of the curricular intervention and so did not provide a good induction for students beginning the learning experience.

• They lacked some knowledge about the higher order inquiry skills and standards-based scientific content the intervention helps students to learn.

• They were confused about how or whether to help students who were unsure what to do next in their learning experiences.

• They lacked skills in leading the small group and whole class interpretive discussions important for students’ understanding of both their MUVE experiences and the data collected.

• They were uncertain how or whether to assess the types of student products generated in the learning experience.

• They were uncertain how and to what extent to support students as they explored the virtual environment.

• They were uncertain how to ask students probing questions in order to help learners clarify their independent variable.

Although this list sounds quite grim, in practice the curricular intervention worked fairly well in these situations because we have designed for scalability, creating curricular interventions so compelling for students and with sufficient internal guidance so that they have a fulfilling, self-directed learning experience—albeit with reduced educational outcomes—even with a confused teacher. Case studies (including logfile analysis of actions and communications) of relatively unguided students illustrate this; for example, one such learner, “Princess,” said in a post-experience interview:

“It is interesting how you have to go around and talk to people and stuff like that. You don’t know what they are going to say. You don’t know if you can figure it out, it’s just like a
mystery. Because you don’t really pay attention in science class and a lot of people don’t pay attention in science class but when they have something to like they think is kind of fun and animated, and they get to do things they learn more and they want to do it more.”

In response to attenuation of the teacher-preparation condition for success, we are evolving the professional development portion of our design to increase its scalability. For example, we are altering the online professional development to produce a just-in-time, “light” version that an overwhelmed teacher can skim for ten minutes per day during the unit, providing essential information needed to guide students for that stage of the learning experience. As another illustration, we are working more closely with district science coordinators so that they provide additional onsite motivation and oversight for teachers.

However, some aspects of any educational innovation are difficult to design for defenestration. For example, earlier an essential condition for our design’s success was described: To use our curriculum teachers must have access to one reasonably modern, Internet-connected Windows computer per student for six forty-five minute class sessions over an approximately two week period. This tacitly presupposes essential conditions for success that have nothing to do with technology:

- Students regularly attend class during the two-week learning experience.
- The teacher is regularly present during the two-week learning experience or at least briefs a substitute to use the design.
- The teacher and school are willing to provide two weeks of classroom time for implementing the design.
- The computer labs are available for use during this time and other teachers are not inconvenienced by losing lab-time, thereby producing a hostile work environment.

All of these are problematic in many schools.
For example, in our urban sites, student attendance rates for class averaged about 50% (although this improved during the implementation of our learning experience, an encouraging measure of its effectiveness). One of our eleven teachers was repeatedly absent from class and did not prepare the substitute, who was frequently found reading a newspaper. In the shadow of high stakes testing and accountability measures mandated by the federal No Child Left Behind legislation, persuading schools to make available two weeks of curricular time is difficult for any design that does not use traditional pedagogy to inculcate students with basic skills and factual content.

These are not factors that can be resolved by DBR and present challenges difficult to overcome by even the best design for defenestration. However, innovators can still attempt to get leverage on these factors. For example, our design is very engaging for students and teachers, uses standards-based content and skills linked to the high stakes tests, and shows strong outcomes with sub-populations of concern to many schools worried about making adequate yearly progress across all their types of students.

**Developing a Generalizable Metric for Assessing a Design’s Scalability**

Researchers, policymakers, and practitioners would all benefit from the creation of a generalizable metric for assessing the scalability of an educational intervention or design. Such an index would summarize the degree to which the educational effectiveness of the design is robust to potential attenuation by its conditions for success. By identifying factors within the intervention’s context that represent important conditions for success and by knowing the extent to which the effect of the intervention is sensitive to variation in each, prospective adopters of the innovation would have a better sense of what its likely effectiveness would be in their own particular circumstances.
An initial step essential to the creation of a viable scalability index is the careful specification of a sensible framework of contextual factors that represent possible general conditions for success of educational innovations. Review of the literature could develop a limited taxonomy of important contextual factors that can serve as viable conditions for success across many types of educational intervention. For example, Russell (2003) has studied a variety of factors thought to influence the conditions for success of the implementation of instructional technology in school districts (see Table 1).

Notice that the listed potential “conditions for success” exist at several nested levels of the educational hierarchy. This nesting not only complicates the creation of a scalability index itself, but also increases the difficulty of estimating with equal precision at each level. Fortunately, for many types of innovations, a relatively small set of contextual factors are often very influential in determining effectiveness. Potential influential factors to be included in the subset include teachers’ knowledge of content and pedagogy, students’ socioeconomic and linguistic backgrounds, students’ mobility and absenteeism, and (for technology-based innovations) the extent and reliability of the computer/networking infrastructure. Examining scalability in the context of this subset of powerful conditions for success may still yield a workable index.

At its core, the evaluation of the sensitivity of an intervention’s impact to select contextual conditions is a question of statistical interactions. In evaluating the sensitivity to the conditions for success, one asks: Is the effect of the intervention dependent on the selected contextual conditions? Is the intervention more effective for children of lower SES, or higher? Does the impact of the intervention depend on important teacher qualities or on features of the classroom and school infrastructure? In a single study, such questions are usually addressed by
the inclusion in the statistical models of interactions between the treatment and its conditions for success.

One approach, then, to the creation of a true scalability index is to ensure that such interactions are included in the statistical models that underpin the data-analyses conducted to assess the implementation of educational interventions. If the interactions have a statistically significant effect, then the effect of the treatment is sensitive to the conditions that participated in the interaction. Having successfully tested for the presence of such an interaction – with child SES, teacher quality or educational infrastructure – one can then estimate the several effect sizes that can be anticipated for the intervention under each of the interacting conditions and pool them into a global index of scalability that captures the extent to which the intervention’s effect size is sensitive to variation in the conditions for success.

However, there are a number of important technical challenges to implementing this approach in practice. First, it is not clear exactly how one should pool the several effect sizes, representing variation in the intervention’s impact across levels of a particular contextual factor, into a single index of sensitivity or scalability. Under the assumption of linear interactions between the treatment and the conditions for success, it is possible that an index might be forged to represent the difference in effect size per standard deviation difference in the underlying and identified condition for success. It may also be possible, depending on the mathematical form of the index proposed, to specify the distributional and statistical properties of the index. My colleague John Willett (who aided with writing this section of this paper) and I plan to investigate this approach, and others-- in both real and simulated data--in order to more fully understand the strengths and weakness of different kinds of indices.
Second, as conditions for success are drawn from higher levels of the organizational hierarchy (classrooms, schools, districts), it gets increasingly difficult to muster the statistical power necessary to detect interactions between these conditions and the intervention being studied. Through the use of multilevel modeling and statistical power analysis, we plan to investigate the extent to which it is feasible to anticipate that a viable scalability index can be developed whose component data would be readily collected empirically in school settings in a single investigation. It is entirely possible that, for conditions of success lying at higher levels of the organizational hierarchy, it may not be possible to estimate the sensitivity of the treatment effect to these conditions in a single study, but instead one would have to rely on a synthesis of findings across many studies in the manner of meta-analysis. We also plan to assess the viability of such meta-analytic approaches to estimating scalability.

Conclusion

Ultimately, design for defenestration involves snatching scalability from the jaws of chaos. What DBR should strive to achieve is the educational equivalent of the kudzu vine, which has largely overrun parts of the southern United States. Kudzu is the ultimate weed: it can grow up to a foot per day under ideal circumstances, thrives under almost any conditions, and is so resistant to herbicides that some chemicals actually escalate rather than retard its invasion of an ecology. In children’s lives outside of school, streaming entertainment videos across broadband connections, cellphones that send text messages and pictures, and videogames such as Pokemon have expansive properties like kudzu. For better or for worse, these interactive media are pervading children’s lives, shifting their patterns of engagement and learning in ways that make typical teaching less compelling and effective. As educators, we need to design “kudzu for the mind”: learning experiences based on these new interactive media to bridge kids’ learning in
and out of school. Our design-based research should focus on creating learning experiences so attractive, powerful, and scalable that they resist defenestration and help to transform education.

References


Dede, C. (in press). Why design-based research is both important and difficult. *Educational Technology*


| District | Community Attitudes about Educational Technology  
|          | District Vision for Technology  
|          | Leadership of Technology Initiatives  
|          | Resources for Technology Initiatives  
|          | Support Services for Technology Initiatives  
|          | Infrastructure of Computers and Telecommunications  
|          | Professional Development Related to Technology  
|          | Relationship Between Technology and Equity  
|          | Technology-Related Policies and Standards |
| School  | Leadership of Technology Initiatives  
|          | Principal’s Pedagogical Beliefs  
|          | Principal’s Technology Beliefs  
|          | Principal’s Technology Preparedness  
|          | School Culture |
| Classroom | Teacher’s Pedagogical Beliefs  
|           | Teacher’s Technology Beliefs  
|           | Teacher’s Technology Preparedness  
|           | Teacher Demographic Characteristics  
|           | Technology Resources  
|           | Students’ Home Access  
|           | Students’ Home Usage  
|           | Students’ Comfort with Technology  
|           | Students’ Demographic Characteristics |