Multisite Studies and Scaling Up in Educational Research

Michael Harwell
University of Minnesota

A scale-up study in education typically expands the sample of students, schools, districts, and/or practices or materials used in smaller studies in ways that build in heterogeneity. Yet surprisingly little is known about the factors that promote successful scaling up efforts in education, in large part due to the absence of empirically supported theories of scaling up. A literature for scale-up studies in education is growing but is years away from providing research-supported practices in planning and conducting these studies. Following the suggestion of Schneider and McDonald (2006) to import relevant knowledge from other fields into the scale-up literature in education, this paper examines the multisite public health and nursing literature in search of a multidisciplinary knowledge base that can inform scaling up efforts in education. Five strategies and practices identified in these literatures as critical to scaling up success are described.

Scale-up studies are a relatively new phenomenon in education. These studies typically have their origins in one or more studies (called demonstration studies in some literatures) examining the same program (intervention, treatment). Demonstration studies often vary in resource constraints, scope, and breadth but share the trait of providing evidence of a program's promise for improving important educational outcomes. The Institute of Education Sciences (U.S. Department of Education, 2012) website offers the following description of scaling up under **Goal Four**:

"Goal Four: If interventions are able to produce positive effects in small efficacy evaluations, they may be ready to be

evaluated in a scale-up evaluation. Scale-up evaluations determine whether or not an intervention is effective when it is implemented under conditions that would be typical if the district were to implement it on its own (i.e., without special support from the developer or research team) across a variety of conditions (e.g., different student populations, different types of schools)."

A typical scale-up study examines a program's effects in ways that build in variation of educational populations (e.g., students, teachers, schools), practices (e.g., professional development), and materials, and usually involves large numbers of students, classrooms, and schools. Examples of scale-up studies in education can be found in Borman, Dowling, and Schneck (2008), Denton, Vaughn, and Fletcher (2003), and Stein et al. (2008).

Purpose and Research Questions

The absence of a detailed scaling up literature in education suggests the value of drawing on other literatures to inform scaling up. Denton et al. (2003) convincingly made this argument: "There is a knowledge base on the science of scaling in other disciplines that educators and policymakers should access and utilize. To facilitate research on the scaling of educational innovations, participation, with explicit government support, among researchers, state education agencies, and local education agencies is essential." (p. 209)

Schneider and McDonald (2006) made a similar argument and identified multisite studies in public health research as one area to draw on. The basic idea is that multisite studies often follow a single site (demonstration) study and thus represent a kind of scaling up.

The purpose of this paper is to identify successful models, strategies, and practices in the public health and nursing multisite literature, and to use this work to inform the planning and execution of scale-up studies in education

including clarifying areas needing research. This focus produced two research questions:

(1) Are there specific models/theories that guide the planning and development of multisite studies in public health and nursing research that enjoy empirical support? If so, can this knowledge base be used to inform scaling up studies in education?

(2) Are there general strategies and practices that guide the planning and development of multisite studies in public health and nursing research that enjoy empirical support? If so, can this knowledge base be used to inform scaling up studies in education?

Scaling Up in Education

Importance

Elmore (1996) has articulated the importance of scaling up programs that show empirical evidence of their effectiveness. As Elmore (1996), Resnick, Stein, and Coon (2008), and others have pointed out systemic problems in the U.S. educational system require systemic solutions, which creates a need to identify programs that are effective across different student populations and different types of schools. Evidence of the growing importance of this topic includes the appearance of scholarly papers (e.g., Denton et al., 2003; Hamilton et al., 2007; Schalock, Schalock, & Ayres, 2006), books such as *Scale-up in Education: Issues in Practice Volumes I and II* (edited by B. Schneider & S.K. McDonald, 2006), creation of the Data Research and Development Center at the University of Chicago and the National Center on Scaling Up Effective Schools at Vanderbilt University, and increases in the number of scale-up studies appearing in the education

literature (Hamilton et al., 2007; McMaster & Fuchs, 2011; Stein et al., 2008). Still, a core literature that researchers and policymakers interested in scaling up studies can turn to for detailed guidance is years away.

Different Perspectives

While there is considerable agreement on the need for scale-up studies in education there is less agreement on what the focus of these studies should be. One perspective is that a scale-up is essentially a larger version of a demonstration study. McDonald, Keesler, Kauffman, and Schneider (2006) captured this view:

> We view scale-up as inherently about size, numbers, "doing more"—about extending the reach of an exemplary intervention to produce similarly positive effects in different settings and to help a greater number of students. Interventions that are not implemented with larger numbers (of students, teachers) are not "scaled-up"—they are local interventions with promising results.
> (p. 16) McDonald et al. (2006)

also argued that educational context is important (see also Raudenbush, 2006). Coburn (2003) provided a different perspective and discussed scaling up as involving something beyond simply "doing more" and proposed "… conceptualizing scale in four dimensions: "depth, sustainability, spread and shift in reform ownership" (p. 4). Spread refers to the implementation of a program at a larger number of sites or to more groups, depth represents an improvement in practice in deep and meaningful ways, and sustainability is putting the infrastructure and systems in place to support continued improvements in practice over time. Shift in ownership represents a transfer of the knowledge and

authority to sustain a program to the implementing sites themselves to allow for continued improvement over time. Most scale-up studies seem to focus on spread which is generally consistent with McDonald et al. (2006). Along these lines, Sternberg et al. (2006) argued that context is crucial and focused on building heterogeneity into a scale-up to assess the generalizibility of a program's effectiveness.

Difficulties

Discussions of scaling up in the education literature inevitably include descriptions of the difficulties of doing so. Examining these difficulties produces the not surprising result that most reappear albeit in different contexts, forms, and intensities. A common theme of these difficulties was captured by Dewa et al. (2002):

"The proposition of introducing the same study design in different settings and programs is deceptively straightforward. The difficulty is not in the conceptualization but in the implementation." (p. 173) The are many categorizations of the difficulties of scaling up (e.g., Cohen, Raudenbush, & Ball, 2003; Fletcher, Foorman, Denton, & Vaughn, 2006; Foorman, Santi, & Berger, 2007; Finnan & Levin, 2006; McDermott, 2000; Schoenfeld, 2006; Sternberg et al., 2006). Collectively, this literature suggests six overlapping themes (a) The nature of the program often increases or decreases the likelihood it will scale-up (b) Inadequate management of the scale-up can undermine training and communication among participants (c) Building in heterogeneity to assess the generalizability of a program's impact increases the complexity of the scale-up and strains financial resources (d) Assuring treatment fidelity often requires significant support during implementation (e) There is an ongoing need to build constituencies for change (e.g., among teachers and school leadership) (f) Methodological challenges.

To some extent these difficulties are a natural outcome of the early stages of the development of a literature (Fuchs & Fuchs, 1998). However two challenges to responding to these difficulties stand out. First, a scale-up literature to guide researchers is enhanced by research on scale-up studies (Constas & Brown, 2006), for example, identifying the most effective strategies for ensuring fidelity of implementation training for teachers in different schools. Researching facets of a scale-up study is a substantial undertaking requiring substantial resources and there is not much evidence that this support is present or forthcoming (Denton et al., 2003).

The second challenge is to develop comprehensive theories or models of scaling up that are empirically supported (Denton et al., 2003; Lee & Luyks, 2005; Schoenfeld, 2006). Denton et al. (2003) commented that comprehensive theories/models of scaling up are non-existent but several theories/models for particular aspects of a scale-up are available.

Theories/Models

Perhaps the best known work in this area is due to Elmore (1996) who proposed five models for replicating educational innovations (i.e., scaling up). One model proposed by Elmore provides teachers with professional development in a program under study each year. This strategy has the effect of incrementally increasing the total number of teachers trained in the program and thus represents a way to scale-up a program. A second model monitors the effects of the program on the actual practice of teachers who receive the professional development and provides continuing support for those who do not implement the program with satisfactory fidelity. Elmore also proposed a "trainer-of-trainers" model in which one group of teachers trained in the program provides training to subsequent teachers.

A fourth model described by Elmore places high performing teachers in selected schools (where lower performing teachers are also in place) who are given instructions to assist each other in the implementation of the program and to provide support to teachers less proficient in implementing the program. Elmore also proposed a fifth model in which a core group of schools nurture leaders in a program who later form another school and mentor new groups of teachers.

More recently, Cobb and Jackson (2011) offered a theory of action for scale-up studies in mathematics education predicated on the argument that instructional improvement is essentially a problem of organizational and teaching learning. These authors organized their theory of action around five themes: (a) A coherent instructional system for instruction that encompasses both formal and job-embedded teacher professional development (b) Teacher networks (c) Mathematics coaches' whose practices provide job-embedded support for teachers' learning (d) School leaders' practices as instructional leaders in mathematics (e) School leaders practices in supporting the development of school-level capacity for instructional improvement.

A different focus is evident in Sternberg et al.'s (2006) theory of contextual variation. Sternberg et al. offered a theory of scaling up based on Brunswik's (1956) notion of representativeness, which is the similarity between the context in which a program was found to be effective (demonstration study) and the class of contexts which the scale-up targets. According to Sternberg et al., representativeness provides a coherent basis for assessing and ensuring generalizability using statistical samples of the environments (conditions) that affect a program. Sternberg et al.'s model is composed of four features: (a) Structural features (policies, mandates, student abilities) (b) Training issues (e.g., logistics, training consistency within groups,

distinctiveness between groups (c) Intervention concerns (e.g., experimental controls, implementation fidelity) (d) Analytic issues (e.g., equating achievement scores, sampling bias).

In the Sternberg et al. theory heterogeneity is the foundation of the scale-up and can encompass content and skill standards across states, districts, and schools; students' ability levels across and within schools; district political environment and commitment to change; and teachers/administrators levels of experience. Data are collected reflecting a program's effectiveness across these conditions and in this sense generalizability is studied empirically.

Other theories/models for scaling up include Baker (2006), Coburn, (2003), Dunst, Trivette, Masiello, and McInerney (2006), Fishman, Marx, Blumenfeld, Krajcik, & Soloway (2004), Flamholtz and Randle (2006), and McDermott (2000). However, none of these theories/models enjoy strong empirical support.

Multisite Studies in Public Health and Nursing Research

Multisite studies have been especially prevalent in public health (e.g., Environmental Health Science, Epidemiology, Behavioral and Community Health Sciences, Health Policy and Management) and nursing research and are typically randomized control trials (RCTs). Not surprisingly there is variation in the definition of what constitutes a multisite study. For example, Meinert (1980) defined a multisite RCT as having three characteristics: (a) The study must involve two or more clinical sites and their separate staffs (b) All sites must follow a common treatment and data collection protocol (c) One site is charged with accruing, processing, and analyzing the data from all of the sites. On the other hand, Kraemer (2000) argued that having multiple sites with

different treatment protocols does not qualify as a multisite RCT but rather represents collaborating multiple single-site RCTs. Dewa et al. (2002) offered a definition that is generally consistent with educational scale-up studies:

> In a multisite study, participating sites may provide different services but share a common protocol. Operationally, this translates into measuring the same outcomes with the same instruments using the same timeframe across differing programs at multiple sites. The common protocol makes outcomes comparable. (p. 175)

An example of a multisite study in public health is Davidow, Katz, Reves, Bethel, and Ngong (2009).

Importance

Lindquist et al. (2002) provided a rationale of the importance of multisite studies which **"…** allow for larger sample size, broader sampling, faster accrual rates, and meaningful subgroup analyses. Successful multisite research requires more thorough planning, and deliberate steps are required to ensure its feasibility and acceptability. Multisite research protocols can be challenging regarding communication, reliability, and data integrity. However, defining and addressing these challenges and selecting subjects and settings appropriately can lead to results that are more generalizable and relevant to practice." (p. 270)(see also Flynn, 2009). Multisite studies are also important because they help resolve the most contentious conflicts in a field (Kraemer, 2000), and in some instances allow expertise to be employed that may not be present at a single site (Organization for Economic Cooperation and Development, 2002).

Different Perspectives

There appears to be substantial agreement in public health and nursing research that a multisite study is a larger version of a demonstration study. In reviewing more than forty multisite studies for this paper all appeared to adopt this perspective, which is consistent with that of Schneider and McDonald (2006).

Difficulties

Discussions of the difficulties linked to multisite studies in the public health and nursing research tend to overlap with those in scale-up studies in education. In general, the difficulties fall into one of four categories: (a) Inadequate management which can undermine building trust and collaboration among participants (i.e., subjects, researchers, staff, vendors, funders) (b) Failure to implement a treatment (program) following study protocol (quality control/fidelity of implementation) and/or a failure to respond to variation in implementation immediately (c) Methodological challenges such as collecting data using site-specific instruments that compromise the ability to assess a treatment's effectiveness (d) Lack of agreement regarding the dissemination of findings, authorships, and contributions to manuscripts (Baynes, 2010; Binswanger, 2000; Constantine & Cagampang, 2000; Davidow et al., 2009; Dewa et al., 2002; Flynn, 2009; Henry & Farrell, 2004; Oncology Nursing Society, 2008; Schene et al., 2000).

Theories/Models

Bossert, Evans, Van Cleve, and Savedra (2002) described a systems approach to planning and conducting multisite studies which treats the parts as more than the whole. The general stages in a systems approach are: (a) Structuring the multisite study in ways that clearly communicate the need for the project to a funder or potential site and its benefits

(b) Examining the feasibility of the study (c) Conducting the project such that needed expertise is applied to problem solving and is mediated by communication among sites and study personnel that in turn is impacted by collegial interaction (d) Closing down the study and disseminating the results while simultaneously discussing the findings and their implications (see also Minnick et al., 1996).

Dewa et al. (2002) offered a model that focused on collaboration as central to the planning, execution, and success of a multisite study. These authors adapted Lancaster's (1985) six "C's" model to guide multisite studies in responding to challenges to the collaborative process (a) Contribution which is the expertise each collaborator brings to the project and is enhanced through multisite meetings, the development of scientific papers, and the dissemination of research results (b) Communication including listserves and email, written, conference calls, website, and in-person meetings (c) Compatibility represents the ability to function as a team, to appreciate strengths, and to blend approaches to create an atmosphere of respect and collegiality (d) Consensus which is a process involving compromise, negotiation, and respect and is closely tied to opportunities for collaborator contribution and communication (e) Credit especially in relation to authorship (f) Commitment both physical (e.g., time, energy, resources) and emotional.

While not a model per se Cooley and Kohl's (2006) "*Scaling up—From vision to large-scale change: A management framework for practitioners*" represents a comprehensive list of activities and facets of a multisite study. This document provides an extensive list of factors to guide planning a multisite study and also describes the results of two public health multisite studies.

In sum, the multisite literature in public health and nursing research appears to lack empirically supported theories/models that can guide the planning and conduct of

these studies. Available literature emphasizes the importance of providing a compelling rationale of the need for a multisite study and of strategies to ensure positive participant collaboration, communication, and fidelity of treatment implementation.

Results

Examination of a sample of multisite studies in the public health and nursing literature provides a response to the research questions posed earlier:

(1) Are there specific models/theories that guide the planning and development of multisite
studies in public health and nursing research that enjoy empirical support? If so, can this knowledge base be used to inform scaling up educational studies? The answer appears to be no and no.
(2) Are there general strategies and practices that guide the planning and development of multisite studies in public health and nursing research that enjoy empirical support? If so, can this knowledge base be used to inform scaling up educational studies?

On the one hand, the answer appears to be no in the sense that commonly used strategies and practices in multisite studies have not been rigorously studied in an empirical sense. For example, a significant number of these studies describe the important role of a central site with the authority to coordinate the remaining sites (Bossert et al., 2002; Lebowitz, 2003; Schene et al., 2010), but there does not appear to be empirical evidence documenting the superiority of a central site model over a decentralized authority model.

On the other hand, the answer appears to be yes in the sense that several strategies and practices have been successfully used in these studies that are similar to those in

educational scale-ups. In particular, the importance of building trust and collaboration among participants, having open lines of communication, and following study protocol in implementing a treatment (program) and/ responding immediately to variation in implementation immediately in multisite studies reinforces their key role in educational scale-up studies. Moreover, the focus of the multisite literature on the critical role of a central site with the authority to manage the remaining sites can inform educational scale-up studies where central sites appear to be uncommon.

Below the result of combining strategies and practices often identified in the public health, nursing, and education literatures as critical to success in multisite studies are described. These are not new except in the sense that they draw on successful strategies and current best practices in multiple literatures, and thus represent multidisciplinary guidelines for educational researchers interested in planning and conducting a scale-up study. Nor are they always the most important strategies and practices in a scale-up, although it is likely they are almost always worth consideration.

Five Strategies and Practices Important in a Scale-up Study in Education

1. Provide a thorough justification for a scale-up study

A scale-up is justified when additional demonstration studies of a program's effectiveness are unlikely to add to existing evidence. Justification for a scale-up study would typically include key demonstration studies documenting the effectiveness of a program, and meta-analyses in which adding additional demonstration studies to the sample of studies has little or no impact on estimates of program effect size, their variability, or moderators of program effect size.

2. Build heterogeneity into the scale-up

Heterogeneity should be built into all facets of a scale-up to enhance generalizibility. According to Sternberg et al. (2006), this process should be at the core of a scale-up effort and typically include heterogeneity in content and skills standards across states, districts, and schools, students' ability levels across and within schools, teachers' skills, and accountability of student progress, as well as the traditional focus on heterogeneous populations. Including multiple sites which are expected to do a better (or worse) job on key facets of a scale-up study, such as implementing the program properly, may also be part of building heterogeneity into a scale-up.

3. Establishing treatment fidelity

Ensuring a program is implemented faithfully is critical in a scale-up study. Delivering a clinical treatment in a demonstration study in public health or nursing (e.g., pain medication dosage) to multiple sites that vary in ways that support generalizability arguably offers modest challenges compared to scaling up many educational programs (see Foorman et al. (2006) and Mostow & Beck (2006) for a description of many of these challenges in education). Still, promoting fidelity of implementation in both settings typically requires standardized training of those implementing the program, standardizing documents and protocols associated with implementing the program, and establishing and evaluating support mechanisms for maintaining fidelity of implementation that likely includes teachers and school leaders.

4. Management of the scale-up

A significant percentage of scale-up and multisite literature focuses on management issues that include ordinary but important tasks. Among these are (a) installing and maintaining lines of communication (b) training project staff,

generating and maintaining support mechanisms for fidelity of implementation (c) managing IRB requirements (d) managing time and costs (e) securing cooperative agreements and outsourcing (f) providing evaluations of program development (g) facilitating community building (h) providing regular empirical feedback (i) conducting data analyses that enhance generalizibility arguments (j) disseminating results. The availability of a single site with authority to manage other sites may be a particularly effective way to manage a scale-up study, although such arrangements do not appear to be common in education.

5. Research design and statistical analysis of scale-up data

This is probably the best researched and most uniformly applicable facet of scale-up studies. A number of authors have provided guidance on planning a multisite study to permit strong causal inferences and substantial statistical power (e.g., Raudenbush, 2006; Raudenbush & Liu, 2000). The role of measurement is also important to ensure that outcomes matched to specific program features are used and that differences in instruments among sites still allow a program's effectiveness to be assessed.

Conclusions

Following the suggestion of Schneider and McDonald (2006) and others to import relevant knowledge from other fields into the scale-up literature in education, this paper examined the multisite public health and nursing literature for theories/models, strategies, and practices that can inform scaling up efforts in education. One important finding is that empirically supported theories/models or strategies and practices linked to successful multisite/scale-up studies are not available in either literature and thus there is a clear need for this work. A second important finding is that the overlap of strategies and practices in these literatures linked to

successful multisite/scale-up studies provides a multidisciplinary perspective that reinforces their prominence in planning and executing a scale-up study in education. The results also suggest that the use of a central site in multisite studies to manage the remaining sites may enhance several facets of a scale-up study.

References

Baker, E.L. (2006). Principles for scaling up: Choosing, measuring effects, and promoting the widespread use of educational innovation. In B. Schnedier & S.K. McDonald (Eds.), *Scale-up in education: Ideas in principle. (pp.*37-54). Lanham, UK: Rowman and Littlefield Publishers.

Baynes, M. (2010). Multisite study issues: International compliance. Presented at a meeting of the British Medicines and Healthcare products Regulatory Agency. Retrieved from http://www.mhra.gov.uk/home/groups/comms-con/documents/website resources/con071117.pdf

Binswanger, H. P. (2000). Scaling up HIV/AIDS programs to national coverage. *Science, 288*(5474), 2173-2176.

Borman, G.D., Dowling, N.M., & Schneck, C. (2008). A multisite cluster randomized field Trial of Open Court Reading. *Educational Evaluation and Policy Analysis, 30* (4) , 389-407.

Bossert, E.A., Evans, S., Van Cleve, L., & Savedra, M.C. (2002). Multisite research: A Systems Approach. *Journal of Pediatric Nursing 17* (1), 38-48.

Blumenfeld, P., Fishman, B.J., Krajcik, J., Marx, R.W., & Soloway, E. (2010). Creating usable innovations in systemic reform: Scaling up technology-embedded project-based science in urban schools. *Educational Psychologist, 35* (3), 149-164.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments* (2nd ed.). Berkeley, CA: University of California Press.

Cobb, P., & Jackson, K. (2011, April). Towards an empirically grounded theory of action for improving the quality of mathematics teaching at scale. National Council of Teachers of Mathematics Research Presession Plenary Address, Indianapolis, IN.

Coburn, C. (2003). Rethinking scale: Moving beyond numbers to deep and lasting change. *Educational Researcher, 32*(6), 3–12.

Constantine, N., & Cagampang, H.H. (1998, August). Improving local evaluation utility within multisite prevention intervention programs. Paper presented at the 106th Annual Conference of the American Psychological Association, San Francisco, CA.

Constas, M.A., & Brown, K,L. (2006). Toward a program of research on scale-up: Analytical requirements and theoretical possibilities. In B. Schnedier & S.K. McDonald,(Eds.),*Scale-up in education: Ideas in principle* (pp.247-257). Lanham, UK: Rowman and Littlefield Publishers.

Cooley, L. & Kohl, R. (2006). *Scaling up—From vision to large-scale change. A management framework for practitioners.* Management Systems International. Retrieved from http://www.msiworldwide.com/files/scalingup-framework.pdf.

Davidow, A.L., Katz, D., Reves, R. Bethel, J., & Ngong, L. (2009). The challenge of multisite epidemiologic studies in diverse populations: Design and implementation of a 22-site study of tuberculosis in foreign-born people. *Public Health Reports, 124*, 391-399.

Denton, C.A., Vaughn, S., & Fletcher, J.M. (2003). Bringing research-based practice in reading intervention to

scale. *Learning Disabilities Research & Practice*, *18* (3), 201–211

Dewa, C.S., Durbin, J., Wasylenki, D. Ochocka, J., Eastabrook, S., Boydell, K.M., & Goering, P. 2002). Considering a multisite study? How to take the leap and have a soft landing. *Journal of Community Psychology, 30*, 173-187.

Dunst, C. J., Trivette, C. M., Masiello, T., & McInerney, M. (2006). Scaling up early childhood intervention literacy learning practices. Center for Early Literacy Learning (CELL) Papers, 1(2), 1-10.

Elmore, R. F. (1996). Getting to scale with good educational practice. *Harvard Educational Review, 66,* 1–26.

Finnan, C, & Levin, H.M. (2006). Accelerated schools and the obstacles to school reform. In M.A. Constas & R.J. Sternberg (Eds), *Translating theory and research into educational practice: Developments in content domains, large scale reform, and intellectual capacity*, (pp.127-150). Mahwah, NJ: Lawrence Erlbaum.

Fishman, B., Marx, R.W., Blumenfeld, P., Krajcik, J., & Soloway, E. (2009). Creating a framework for research on systemic technology innovations. *Journal of the Learning Sciences, 13 (1), 43-76*.

Flamholtz, E., & Randle, Y. (2006). Measuring and managing successful organizational scale-up. In B. Schnedier & S.K. McDonald *(*Eds.) *Scale-up in education: Ideas in principle (pp.* 217-243). Lanham, UK: Rowman and Littlefield Publishers.

Flynn, L. (2009). The benefits and challenges of multisite studies: Lessons learned. *AACN Advanced Critical Care, 20 (4)*, 388-391.

Foorman, B.R., Santi, K.L., & Berger, L. (2006). Scaling assessment-driven instruction using the internet and handheld computers. In B. Schnedier & S.K. McDonald (Eds.) *Scale-up in education: Ideas in practice*

(pp. 69-90). Lanham, UK: Rowman and Littlefield Publishers.

Fuchs, D., & Fuchs, L. S. (1998). Researchers and teachers working together to adapt instruction for diverse learners. *Learning Disabilities Research & Practice, 13,* 126–137.

Henry, D.B., Farrell, A.D. (2004). The study designed by a committee: Design of the multisite violence prevention project. *American Journal of Preventive Medicine, 26*, 12-19.

Kraemer, H.C. (2000). Pitfalls of multisite randomized clinical trials of efficacy and effectiveness. *Schizophrenia Bulletin, 26 (3),* 533-541.

Lancaster, J. (1985). The perils and joys of collaborative research. What to look for and what to avoid when putting together a team. *Nursing Outlook, 33 (5*), 231–238.

Lee, O., & Luykx, A. (2005). Dilemmas in scaling up innovations in elementary science instruction with nonmainstream students. *American Educational Research Journal, 42 (3),* 411-438.

Lindquist, R., Treat-Jacobson, D., & Watanuki, S. (2000). A case for multisite studies in critical care. *Heart & Lung: The Journal of Acute and Critical Care, 29* (4), 269-277.

McDonald, S.K., Keesler, V.A., Kauffman, N.J., & Schneider, B. (2006). Scaling-up exemplary Interventions. *Educational Researcher, 35*, 15-24.

McDermott, K.A. (2000). Notes: Barriers to large-scale success of models for urban school reform. *Educational Evaluation and Policy Analysis, 22,* 83-89.

McMaster, K.L., & Fuchs, D. (2011, September). Scaling up an evidence-based practice: Importance of fidelity and flexibility. Paper presented at the fall meeting of the

Society for Research on Educational Effectiveness, Washington, D.C.

Meinert, C.L. Toward more definitive clinical trials. *Controlled Clinical Trials*, 1:249-261, 1980.

Minnick, A., Kleinpell, R.M., Micek, W., & Dudley, D. (1996). The management of a multisite Study. *Journal of Professional Nursing, 12 (1),* 7-15.

Mostow, J., & Beck, J. (2006). When the rubber meets the road: Lessons from the in-school adventures of an automated reading tutor that listens. . In B. Schnedier & S.K. McDonald (Eds.), *Scale-up in education: Ideas in practice* (pp. 183-200). Lanham, UK: Rowman and Littlefield Publishers.

Oncology Nursing Society. (2008). Multisite research strategic plan progress report. Retrieved from http://www.ons.org/Research/Multisite/

Raudenbush, S.W. (2006). Designing field trials of educational innovations. In B. Schneider & S.K. McDonald (Eds.), *Scale-up in education: Issues in practice* (pp. 23-40). Lanham, UK: Rowmann & Littlefield Publishers.

Raudenbush, S.R., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods, 5,* 199-213.

Resnick, L. B., Stein, M. K., & Coon, S. (2008). Standards-based reform: A powerful idea unmoored. In R.D. Kahlenberg (Ed.). *Improving On No Child Left Behind: Getting Education Reform Back on Track*. New York: The Century Foundation Press.

Schalock, H.D., Schalock, M.D., & Ayres, R. (2006). Scaling up research in teacher education: New demands on theory, measurement, and design. *Journal of Teacher Education, 57,* 102-119.

Schene, A.H., Koeter, M., Wijngaarden, B.V., Knudsen, H.C., Leese, M., Ruggeri, M., White, I.R., & Vazquez-Barquero, J.L. (2000). Methodology of a multi-site

reliability study: EPSILON Study 3. *British Journal of Psychiatry, 177*, 15-20.

Schneider, B., & McDonald, S.K. (2006). Introduction. In B. Schnedier & S.K. McDonald (Eds.), *Scale-up in education: Ideas in principle* (pp. 1-18). Lanham, UK: Rowman and Littlefield Publishers.

Schoenfeld, A. (2006). Notes on the educational steeplechase: Hurdles and jumps in the development of research-based mathematics instruction. In M.A. Constas & R.J. Sternberg (Eds.), *Translating theory and research into practice into educational practice: Developments in content domains, large-scale reform, and intellectual capacity* (pp. 9-30). Mahwah, NJ: Lawrence Erlbaum.

Stein, M.L., Berends, M., Fuchs, D., McMaster, M., Sáenz, L., Yen, L., Fuchs, L.S., & Compton, D.L. (2008). Scaling up an early reading program: Relationships among teacher support, fidelity of implementation, and student performance across different sites and years. *Educational Evaluation and Policy Analysis, 30 (4*), 368-388.

Sternberg, R., Birney, D., Jarvin, L., Kirlik, A., Stemler, S., & Grigorenko, E.L. (2006). From molehill to mountain: The process of scaling up educational interventions (Firsthand experience upscaling the theory of successful intelligence. In M.A. Constas & R.J. Sternberg (Eds), *Translating theory and research into educational practice: Developments in content domains, large scale reform, and intellectual capacity* (pp.05-222). Mahwah, NJ: Erlbaum.

U.S. Department of Education (2012). Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. Retrieved from http://ies.ed.gov/funding/12rfas.asp.

Zeichner, K. M. (2005). A research agenda for teacher education. In M. Cochran-Smith & K. Zeichner

(Eds.), *Studying teacher education: The report of the AERA Panel on Research and Teacher Education* (pp. 737-759). Mahwah, NJ: Lawrence Erlbaum.