Measuring the Scale Outcomes of Curriculum Materials

Natalie Pareja Roblin, University of Pittsburgh, natalie.pareja@gmail.com
Christian Schunn, University of Pittsburgh, schunn@pitt.edu
Susan McKenney, University of Twente, susan.mckenney@utwente.nl

Abstract: Learning sciences research often focuses on learning environments that aim for substantial impact on learners and is increasingly concerned with doing so at scale. While learning scientists regularly measure teacher outcomes and learner outcomes, data are rarely collected on other dimensions of scale that are important for uncovering practices and designs that achieve wide-reaching, long-term impact on learning, notably: spread of underlying norms and beliefs, sustainability of use, and shift of ownership into the hands of the practitioners using the learning environments. This paper describes an instrument for measuring these scale outcomes with a focus on a particular component of learning environments: curriculum materials. By presenting both the instrument and results from its pilot use, this paper offers a conceptual as well as a methodological contribution toward developing much-needed instrumentation for evaluating the outcomes of learning environments at scale.

Introduction

A core commitment of the learning sciences is to envision and create learning environments with transformative outcomes. A key challenge to upholding this commitment has been the development of environments that both substantially impact the learners and do so in ways that are sustainable at scale. To date, the learning sciences have problematized what it means for learning environments to be productive across many contexts (Tatar, Roschelle & Hegedus, 2014). Further, the field has also conceptualized the learning of the system that occurs when attending to various stakeholders in a larger ecology, such as principals, districts, states (Marx et al., 2004). Yet in order to have a science of scaling, we need to develop measures of scale that are meaningful across specific project instances. This is no easy task. By their very nature, scale outcomes are broadly distributed across place and time (i.e., potentially across the world and across years), so the feasibility of measures is particularly problematic. This paper describes the Scale Outcomes Rubric, an instrument to measure the outcomes of curriculum materials as well as initial results from pilot testing the rubric. As such, this paper offers both a lens and a tool for collecting data on scale outcomes of curriculum innovation through materials. Here, the term curriculum materials, refers to resources designed for use by teachers in the classroom to guide their instruction, including textbooks, supplementary units, and instructional media (Remillard, Harris & Agodini, 2014).

Theoretical framework: A multidimensional view on scale

Impacting learning at scale of major interest for education policy, practice, and research. This interest stems from the desire of education policymakers and curriculum developers to create deep and lasting changes in teaching and learning (Sanders, 2012; Sabelli & Harris, 2015), and is reflected both in a growing literature concerned with scaling up educational innovations (Coburn, 2003; Lee & Krajcik, 2012; Levin, 2013; Looi et al., 2014; Lynch, Pyke & Grafton, 2012; McDonald, Kessler, Kauffman, & Schneider, 2006; Sanders, 2012), and in increasing funding for research and development of innovative interventions that are brought to scale (e.g., Investing in Innovation program at the US Department of Education, ITEST SPrEaD program at the US National Science Foundation). Most research concerned with the scaling up of curriculum innovations has focused either on the provision of convincing evidence of their effectiveness in various settings and with a variety of teacher and student populations (e.g., Geier et al., 2008; Lynch et al., 2012; Plass et al., 2012; Tartar et al., 2008), or more generally on the challenges of going to scale (e.g., Datnow, 2002; Lee & Krajcik, 2012; Levin, 2013; Lynch, 2012). In these, scale has been viewed primarily as increasing the number of implementing schools and districts.

Others have argued that going to scale cannot be undertaken after proof-of-concept, but that it is an essential concern that must be factored into the initial innovation design (Clarke & Dede, 2009; McKenney, 2018). Doing so requires a more comprehensive view of scale that acknowledges its complexity and attends to the qualitative changes in educational practice resulting from the reform (Coburn, 2003; Cohen & Ball, 2006; Elmore, 1996; Sabelli & Harris, 2015). To capture this comprehensive view, Coburn (2003) proposed a multidimensional conceptualization of scale that includes attention to the *depth* of changes in classroom instruction; to issues of *sustainability*; to *spread* of norms, principles and beliefs; and to *shift in ownership* from external designers to schools and districts so that reform can become self-generative. These dimensions are taken as points of departure for creating an initial framework for measuring the outcomes of (science) curriculum materials. Learner outcomes and teacher outcomes—critical to the depth of changes in classroom instruction—are frequently measured (though teacher outcomes are measured much less often). Dimensions of sustainability, spread, and shift in ownership are of critical importance for supporting learning at scale, but rarely measured. We initially operationalized these as follows:

- Sustainability: Curriculum functions without implementation scaffolds put into place by the designers.
- Spread: Curriculum or underlying ideas have been adopted for use by at least twice as many units (teachers, schools, districts) as were involved in the pilot testing
- Shift: Systems are in place for local maintenance of curriculum implementation

We then undertook the study described here to further conceptualize and measure these three dimensions.

Study purpose

A nuanced view of innovations at scale acknowledges the complexity of multiple dimensions involved. An important first step in deciding whether or not an innovation holds potential for implementation at scale, is already being taken through evaluation studies of curriculum materials which focus primarily on demonstrating effectiveness with respect to student learning, and (though less often) teacher learning. Yet, understanding scale also requires consideration of sustainability, spread of key pedagogical principles and beliefs, and shift in ownership (Coburn, 2003; Clarke & Dede, 2009). Tools to systematically describe and assess the *outcomes* of curriculum materials on these three dimensions of scale are currently lacking. At the same time, they are urgently needed. To design effectively for scale, we need feedback on the characteristics of our designs which enable or hinder innovations in reaching large numbers of learners across multiple contexts. While designers may able to envision and create learning environments with transformative outcomes in the short term, they need to be able to test sustainability, spread and shift to understand if and how their designs thrive in the long term. To address this need, the present study set out to develop an instrument to efficiently describe the outcomes of curriculum materials on sustainability, spread, and shift in ownership. The instrument was validated with a sample of K-12 research-based curriculum materials for K-12 science developed with federal funding in the United States.

Methods

Sample

To test the instrument, a broad sample of research-based curriculum materials was obtained, focusing on those funded by the National Science Foundation (NSF) and the Institute of Education Sciences (IES). We focused on these grant-funded materials because 1) NSF and IES are the largest funders of research-based curriculum materials in the United States (Feder, Ferrini-Mundy, & Heller-Zeisler, 2011), and 2) the extra accountability associated with public funds would likely motivate projects to document evidence of impact. The sample was limited to projects awarded between 2001 and 2010 to ensure access to documentation on scale outcomes.

We obtained and analyzed in detail documentation from 51 projects concerned with the design of K-12 science curriculum materials for classroom use. The identification and selection of these projects followed a five-step procedure. In the first step, all NSF and IES grant awards concerned with curriculum design for K-12 science education were sought in the official databases of each funding agency. This resulted in 1,301 hits. In the second step, project abstracts were screened to identify relevant awards, using the following inclusion criteria: the project targets mainstream K-12 science education, has curriculum design as an important goal, and focuses on curriculum materials for classroom use. This narrowed the selection to 162, after which different awards linked to the same project were then merged in step 3, yielding 146 projects. To make the process of in-depth analysis of project scale outcomes manageable, step 4 involved a random subsample of projects based on principal investigators. To ensure efficient use of principal investigators' time for member-check interviews, when a principal investigator received multiple awards within our 10-year time window, all the awards were included, resulting in 83 projects. We then (step 5) carefully examined available documentation (e.g., academic publications, evaluation reports, sales figures) that could be reporting on the outcomes of the curriculum materials, which reduced the sample to 51 projects with any documented curriculum outcomes (some projects focused on theory testing) and 26 projects with any evidence of sustainability, spread, and/or shift in ownership. These materials were highly diverse, ranging from elementary through high school, from short modules to multi-year sequences, from textbooks to simulations, and from primarily student focused materials to also including many forms of rich teacher support.

Instrument development

Building on our theoretical framework, we developed an instrument to describe the outcomes of science curriculum materials on three dimensions of scale typically understudied: sustainability, spread, and shift in ownership. Feasibility was an important consideration for instrument development, which followed an iterative process and was informed by 1) a literature review (only briefly summarized in the theoretical framework section), 2) a systematic analysis of documentation produced by a sample of research-based science curriculum development projects, and 3) feedback from experts in the field of (science) education research and curriculum design. First, key components were identified for each of the three dimensions of scale based on relevant literature about scaling up educational innovations. Next, inductive analyses of documentation produced by a small subset of science curriculum development projects (N=3) that were identified to have successfully achieved scale were used to refine our operationalization of each dimension and to identify the types of outcomes evidence typically reported by development projects. This resulted in a first draft instrument that was then discussed with experts in the fields of science education research and curriculum design (for more details, see validity and reliability section below). Based on the expert feedback, further refinements were made to the instrument. The final version of the instrument comprises three main sections that describe the outcomes of curriculum materials on sustainability, spread, and shift in ownership. Each dimension is operationalized into a set of two or more components that describe key project outcomes theoretically linked to the respective dimension (see Table 1).

Table 1: Operationalization of Sustainability, Spread and Shift in Ownership Outcomes in the rubric

Components	Indicators
Sustainability	
Intention to continue (partial) use of curriculum materials	Individual teachers express *intention* to continue use of the curriculum materials after direct project support ends.
Sustained (partial) use of curriculum materials	Individual teachers *continue use* of the curriculum materials or parts of them after direct project support ends.
Sustained use of key pedagogical ideas	Individual teachers express *intention* to or *continue use* of key pedagogical ideas underlying the curriculum materials after direct project support ends.
Spread	
Spread of *curriculum materials*	Curriculum materials are used by new teachers (not previously involved in the project), possibly in other grade levels than originally intended or in other subject areas than originally intended.
Spread of *key pedagogical ideas*	Key pedagogical ideas, not previously in place, are used by new teachers (not previously involved in the project) and/or in other grade levels or subject areas than originally intended.
Shift in ownership	
Formal decision to adopt the materials after direct project support ends	Decisions made by a district, school, or department to adopt the curriculum (or key practices/features of the curriculum) beyond the pilot or field trial.
Formal decisions to adapt the curriculum materials to local needs	Decisions made by a district, school, or department to adapt (components of) the curriculum to local needs.
Maintenance of professional development and physical support structures	Local systems are put into place for continued teacher professional development (e.g., school-based professional development, district level supervisors or coaches) and/or maintenance of the physical aspects of the curriculum (e.g., kit refurbishment, website access, etc.).
Presence of *local champions*	Local champions (e.g. teachers, parent groups, change agencies, district administrators) support the dissemination, adoption and implementation of the curriculum materials.

Content validity refers to the extent to which a measurement instrument captures all the facets of a construct (Wilson, Pan, & Schumsky, 2012). One way of achieving content validity involves a panel of subject matter experts that independently consider the importance of individual items within an instrument (Ayre & Scally, 2014). In the present study, feedback was sought from experts in the field of (science) education research and curriculum design in order to validate the rubric. The validation procedure involved two steps. In a first step, science education researchers were asked to 1) assess the importance of the components identified to describe projects' outcomes on sustainability, spread and shift in ownership, and 2) identify relevant components not yet included in the instrument and/or components requiring further clarification. Expert feedback was gathered through individual telephone interviews (N=5), and a focus group meeting (N=8). Experts started by individually rating each component as "essential", "useful but not essential", or "not necessary". Lawshe's content validity ratio (CVR) was then calculated to ascertain the degree of experts' agreement on how essential each component is (Ayre & Scally, 2014; Wilson et al., 2012). CVR values range between −1 (perfect disagreement) and +1 (perfect agreement), with CVR values above zero indicating that over half of panel members agree on a component being essential (Ayre & Scally, 2014). For a panel of 13 experts, a critical value of .54 or above indicates that agreement was not likely to occur by chance, using a one-tailed alpha level of .05 (Ayre & Scally, 2014). The CVR for all the components in the instrument met this criterion, except for one ("Formal endorsement of key principles underlying the curriculum materials in school/district/State policy", CVR= -.07). This component was therefore removed from the instrument. Next to rating the importance of the components identified for each dimension of scale, experts also made suggestions of other components that they believed should be added to the instrument (e.g., the emergence of local champions of the curriculum materials as another key indicator of shift in ownership), or that they believed required clarification.

In a second step, the revised instrument was presented to a group of curriculum designers to assess its overall comprehensiveness and clarity (N=5). This resulted in minor textual edits to improve the clarity of the components and their indicators. The final version of the instrument was then used to describe the scale outcomes of a highly diverse sample of K-12 research-based curriculum materials (N=51) based on an extensive review of project documentation. This further contributed to determine whether the instrument could be used to describe and assess the outcomes of a wide range of development projects concerned with the design of curriculum materials for classroom use. To assess reliability and further establish meaningfulness of rubric to the curriculum designers, the results of the coding were verified with the principal investigators in a telephone interview. Member-check interviews were completed for 45 projects (88%); a few principal investigators were not available or did not return e-mails.

During the member-check interviews, principal investigators were asked to verify if the description of their project's scale outcomes was accurate, and to share access to additional documentation that could be relevant for our analyses (e.g. evaluation reports, sales figures, analytics). To facilitate this process, a few days prior to the interview, principal investigators were provided with an excel file containing: 1) a list of all the documentation found for the project, and of those documents included as relevant to the outcomes analyses; 2) an overview of the project's outcomes on sustainability, spread and shift in ownership; and 3) an overview of the evidence used to code the project's scale outcomes. Interviews lasted between 60 to 90 minutes and focused on a discussion of the coded outcomes for each dimension, and the identification of other documentation that could provide additional insights about the project's scale outcomes.

Data collection and analysis

To identify and select relevant project documents, studies and reports produced by the 51 sampled projects were sought on project websites, principal investigator's personal websites, and scientific databases (Google Scholar and Web of Science). Keywords used to search for project documentation in scientific databases included: award number, project title, and/or (co-) principal investigator's name. In a member-check interview, principal investigators were asked to go over the search results and identify additional publications that could be relevant to our analyses. This resulted in the identification of over 500 references to publications and project documents. We were unable to obtain the full text for 43 of these references, and the majority of these missing references (53%) were conference presentations.

Available publications and documents were then screened to determine relevance to our study. To be deemed relevant, documents had to report on project outcomes related to sustainability, spread, and/or shift in ownership (see Table 2). In order to ensure a comprehensive review and prevent potential publication bias, all documents reporting on the scale outcomes of the project, regardless of the publication type, were included in the analyses (e.g., journal articles, book chapters, conference proceedings, evaluation reports, sales reports, web analytics). However, when multiple document types reported results from the same study, peer-reviewed and/or most recently published sources were prioritized, and duplicates were removed.

Next, documents were screened for analyzability. Available project documentation was reviewed to describe the project outcomes on three dimensions of scale: sustainability, spread, and shift in ownership. When a document reported multiple studies, results for each study were coded separately. Document screening resulted in the identification of 42 documents reporting on sustainability, spread, and/or shift. Only 16% of the documentation reporting on these dimensions was peer-reviewed (e.g., journal articles, edited book chapters, conference proceedings), while the largest proportion consisted of non-peer reviewed sources including project (evaluation) reports (62%), web or sales analytics (17%), or unpublished manuscripts (5%).

Finally, documents were analyzed. Because of the limited evidence available, project outcomes for each of the dimensions were scored as either absent or present, although more levels of each dimension were originally conceptualized (e.g., sustained use in only a few contexts vs. sustained use in most contexts). Table 2 contains examples of the types evidence collected for each of the dimensions studied.

Table 2: Examples of project evidence for outcomes related to sustainability, spread and shift

Component	Indicator	Example
Sustainability		
Intention to continue (partial) use of curriculum materials	Individual teachers express *intention* to continue use of the curriculum materials after direct project support ends.	"In the last year of the program, the second cohort of teachers unanimously stated that they plan to continue using the materials after the program officially ends. In fact, they were very concerned about retaining access once the funding was over."
Sustained (partial) use of curriculum materials	Individual teachers *continue use* of the curriculum materials or parts of them after direct project support ends.	"Even as the grant funding has decreased, teachers and students continue to use resources developed during this ITEST project. The curricular materials, as intact projects or as discrete activities, are currently in use by 8th, 9th, and 10th grade teachers and students."
Sustained use of key pedagogical ideas	Individual teachers express *intention* to or *continue use* of key pedagogical ideas underlying the curriculum materials after direct project support ends.	No evidence available.
Spread		
Spread of *curriculum materials*	Curriculum materials are used by new teachers (not previously involved in the project), possibly in other grade levels than originally intended or in other subject areas than originally intended.	"Due to the results of the impact of the program on student outcomes in reading on the state reading test as well as the positive findings from the teacher surveys, the district is expanding the program to all 72 buildings."
Spread of *key pedagogical ideas*	Key pedagogical ideas, not previously in place, are used in other grade levels or subject areas than originally intended and/or by new teachers (not previously involved in the project).	"In terms of the effects of participation on the respondents' teaching in other areas, the teachers answered with a significant (meaning, with a greater-than-expected consensus) that participation had a moderate effect on their teaching other content (…) When asked to elaborate on what the changes were, respondents tended to cite an increased use of inquiry, or uses of the engineering design process applied to areas not normally characterized as unique to engineering."
Shift		
Formal decision to adopt the materials after direct project support ends	Decisions made by a district, school, or department to adopt the curriculum (or key practices/features of the curriculum) beyond the pilot or field trial.	"The Director of Curriculum and Training and the [name of curriculum] specialist at [publisher] also reported that the materials are being distributed throughout the nation, with large school districts making adoptions of [curriculum X] (e.g., in Arizona, Oregon, Georgia, New York, Illinois, and Ohio)."
Decisions to adapt the curriculum materials to local needs	Decisions made by a district, school, or department to adapt (components of) the curriculum to local needs.	No evidence available

Maintenance of professional development and physical support structures	Local systems are put into place for continued teacher professional development and/or maintenance of the physical aspects of the curriculum (e.g., kit refurbishment, website access, etc.).	"…additional help with facilitating professional development workshops was provided by consultants from [Publisher], but the internal leadership team was the primary deliverer of the professional development. This leadership team met monthly to plan professional development for the other teachers. These planning sessions were focused on the identified needs of teachers and students in the district."
Presence of *local champions*	Local champions (e.g. teachers, parent groups, change agencies, district administrators) support the dissemination, adoption and implementation of the curriculum materials.	"Texas teacher leaders have presented the results of their work in the [project name] project at geoscience professional meetings […] The teacher leaders are not only serving as effective [project name] Climate ambassadors to achieve broad dissemination of the [project name] Climate modules but also are evolving into respected colleagues, as evidenced by their professional accomplishments."

Results

As shown in Figure 1, roughly a quarter to a third of the studied projects had any evidence on each of sustainability (33%), spread (29%), and shift (27%).

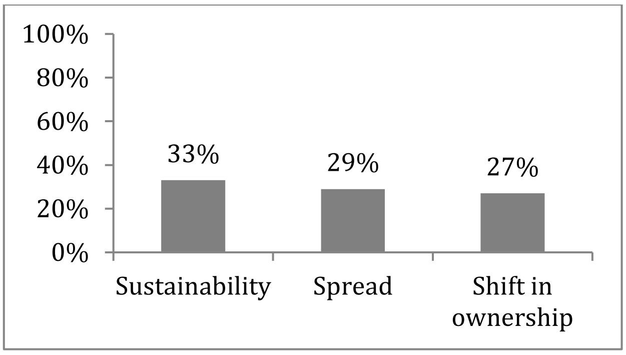

Figure 1. Proportion of projects with evidence of sustainability, spread and/or shift in ownership (N=51).

Table 3 shows the percentages of projects with any evidence on each component of sustainability, spread, or shift. Within each dimension, degree of availability varied substantially across components.

Table 3: Proportion of projects with available evidence for each component (N=51)

Dimension	Components	Proportion of projects with available evidence
Sustainability	*Intention to continue (partial) use* of curriculum materials	25%
	Sustained (partial) use of curriculum materials	10%
	Sustained use of key pedagogical ideas	0%
Spread	Spread of *curriculum materials*	27%
	Spread of *key pedagogical ideas*	4%
Shift in ownership	*Formal decision to adopt* the materials after direct project support ends	6%
	Formal decisions to adapt the curriculum materials to local needs	0%
	Maintenance of professional development and physical support structures	8%
	Presence of *local champions*	24%

Table 4 demonstrates how we might examine scale outcomes in light of other information. In this case, the scale outcomes are examined in relation to start year of the award and the total amount awarded. This particular 10-year funding window experienced substantial changes in funding priorities (e.g., a shift from allowing development-only to requiring research alongside development). Additionally, one might have predicted that high levels of project funding is required to aim for broad use and have resources to study scaling outcomes. Only evidence of sustainability showed any changes over time, with projects in the second half of the time window being 30% more likely to provide evidence of sustainability. Follow-up analyses showed this change was the result of shift towards materials that could be distributed online, since web-metrics provide easy access to sustainability information; interestingly, though, this did not lead to significantly greater information about spread. Finally, there was no evidence that larger grants produced more scaling outcomes evidence; the larger projects focused the project on other aspects of the work (e.g., developing more materials or conducting more research on student and teacher outcomes). Overall, these analyses show how the instrument can give insight not only into specific project outcomes, but also how it might be used to investigate the scale outcomes of funding programs and mechanisms.

Table 4: Exp(b) (i.e., odds ratio) from logistic regression analyses of the effects of award start year and total amount awarded on availability of evidence of sustainability, spread and/or shift in ownership

	Exp(B)		
Predictors	Evidence of Sustainability	Evidence of Spread	Evidence of Shift in Ownership
Award Start Year	1.30*	1.12	.92
Total Amount Awarded	1.00	1.00	1.00

**p<.05*

Discussion

This study was undertaken to develop an instrument to describe the outcomes of curriculum materials on sustainability, spread and shift in ownership. The content validity results seem promising, but further research is necessary to assess and refine its reliability between multiple coders. In so doing, it would be useful to seek lessons learned during the development of related instruments. Although we know of none that specifically measure sustainability, spread and shift, the Designing for Sustained Adoption Assessment Instrument (Stanford, et al., 2016) and i3 Sustainability Rubric (https://i3community.ed.gov/insights-discoveries/1835) do seem relevant.

The general lack of evidence is a notable, though not necessarily surprising finding. In light of recent criticism about the limited evidence available about the actual scale outcomes of development projects (Stanford et al., 2016), our findings suggest that this is not only due to lack of tools for measuring such outcomes, but also because few projects pay attention to them in any formal way. In the interviews, most PIs expressed frustrations with wanting to know what happened to their work but not having any means to investigate that, often noting that the funding requests for proposals paid little attention to scale and therefore limited how much of the budget could be devoted to these aspects. Therefore, this issue seems an important consideration for funders. We take the stance that all stakeholders would benefit if funders were to support impact studies of both federally-funded and commercial curriculum materials.

Further research on scale outcomes could help both curriculum designers and policymakers alike. For designers, it would be useful to know: What characteristics of curricula contribute to positive outcomes at scale? Equally? Or are particular curriculum features more regularly associated with sustainability, spread or shift? For policymakers, broader analysis of funded curriculum design projects and their outcomes could give important feedback on past funding support and help identify key concerns to be taken up in future programs that support the design of learning environments in general and of curricula in particular.

This study takes a first step toward measuring the scale outcomes of curriculum materials. Although, the Scale Outcomes Rubric was tested with grant-funded K-12 science curriculum materials, it could be used in other areas and possibly also for higher education. The instrument may help to frame new data collection on the impact of curriculum materials, or to develop models of factors that may contribute to yielding the desired outcomes in the long term, after innovations are left to thrive on their own. While more work is needed, this study brings us one step closer to being able to measure the extent to which innovations are able to reach large numbers of learners across multiple contexts.

References

Ayre, C., & Scally, A. J. (2014). Critical values for Lawshe's content validity ratio: revisiting the original methods of calculation. *Measurement and Evaluation in Counseling and Development*, *47*(1), 79-86.

Clarke, J. & Dede, C. (2009). Design for scalability: A case study of the River City curriculum. *Journal of Science Education and Technology*, *18*(*4*), 353–365.

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

Cohen, D. K., & Ball, D. L. (2007). Educational innovation and the problem of scale. In B. Schneider, B. & McDonald, S. K. (Eds.). *Scale-up in education: Ideas in principle*, 1 (19-36). Rowman & Littlefield Publishers.

Datnow, A. (2002). Can we transplant educational reform, and does it last?. *Journal of Educational Change*, *3*(3), 215-239.

Elmore, R. (1996). Getting to scale with good educational practice. *Harvard Educational Review*, *66*(1), 1-27.

Feder, M., Ferrini-Mundy, J. & Heller-Zeisler, S. (2011). *The federal science, technology, engineering, and mathematics (STEM) education portfolio*. A Report from the Federal Inventory of STEM Education Fast-Track Action Committee. Committee on STEM Education, National Science and Technology Council.

Geier, R., Blumenfeld, P. C., Marx, R. W., Krajcik, J. S., Fishman, B., Soloway, E., & Clay-Chambers, J. (2008). Standardized test outcomes for students engaged in inquiry-based science curricula in the context of urban reform. *Journal of Research in Science Teaching*, *45*(8), 922-939.

Lee, O., & Krajcik, J. (2012). Large-scale interventions in science education for diverse student groups in varied educational settings. *Journal of Research in Science Teaching*, *49*(3), 271-280.

Levin, B (2013). *What Does It Take to Scale Up Innovations? An examination of Teach for America, the Harlem Children's Zone, and the Knowledge Is Power Program.* Boulder, CO: National Education Policy Center. Retrieved from http://nepc.colorado.edu/publication/scaling-up-innovations

Looi, C. K., Sun, D., Wu, L., Seow, P., Chia, G., Wong, L. H., ... & Norris, C. (2014). Implementing mobile learning curricula in a grade level: Empirical study of learning effectiveness at scale. *Computers & Education*, *77*, 101-115.

Lynch, S.K. (2012). Metaphor and theory for scale-up research: eagles in the Anacostia and activity systems. In B.J. Fraser et al. (eds.), *Second International Handbook of Science Education* (pp. 913-929), Springer International Handbooks of Education.

Lynch, S. J., Pyke, C., & Grafton, B. H. (2012). A retrospective view of a study of middle school science curriculum materials: Implementation, scale-up, and sustainability in a changing policy environment. *Journal of Research in Science Teaching*, *49*(3), 305-332.

Marx, R. W., Blumenfeld, P. C., Krajcik, J. S., Fishman, B., Soloway, E., Geier, R., & Tal, R. T. (2004). Inquiry-based science in the middle grades: Assessment of learning in urban systemic reform. *Journal of research in Science Teaching*, *41*(10), 1063-1080.

McDonald, S. K., Keesler, V. A., Kauffman, N. J., & Schneider, B. (2006). Scaling-up exemplary interventions. *Educational Researcher*, *35*(3), 15-24.

McKenney, S. (2018). How can the learning sciences (better) impact policy and practice? *Journal of the Learning Sciences*, 27, 1-8.

Plass, J. L., Milne, C., Homer, B. D., Schwartz, R. N., Hayward, E. O., Jordan, T., Verkuilen, J., Ng, F., Wang, Y., & Barrientos, J. (2012). Investigating the effectiveness of computer simulations for chemistry learning. *Journal of Research in Science Teaching, 49*(3), 394-419.

Remillard, J., Harris, B., & Agodini, R. (2014). The influence of curriculum material design on opportunities for student learning. *ZDM Mathematics Education, 46*(5), 735–749.

Sabelli, N. H., & Harris, C. J. (2015). The Role of Innovation in Scaling Up Educational Innovations. In C.-K. Looi, L.W. Teh (eds.), *Scaling Educational Innovations* (pp. 13-30). Springer: Singapore.

Sanders, M. G. (2012). Achieving Scale at the District Level A Longitudinal Multiple Case Study of a Partnership Reform. *Educational Administration Quarterly*, *48*(1), 154-186.

Stanford, C., Cole, R., Froyd, J., Friedrichsen, D., Khatri, R., & Henderson, C. (2016). Supporting sustained adoption of education innovations: The Designing for Sustained Adoption Assessment Instrument. *International Journal of STEM Education*, *3*(1), 1-13.

Tatar, D., Roschelle, J. & Hegedus, S. (2014). SIMCALC: Democratizing access to advanced mathematics. *International Journal of Designs for Learning*, 5, 83-100.

Wilson, F. R., Pan, W., & Schumsky, D. A. (2012). Recalculation of the critical values for Lawshe's content validity ratio. *Measurement and Evaluation in Counseling and Development*, *45*(3), 197-210.