Engineering [for] Effectiveness in Mathematics Education: Intervention at the Instructional Core in an Era of Common Core Standards

Jere Confrey and Alan Maloney

The Process of "Engineering [for] Effectiveness"

Improving schools has often been cast as a challenge of identifying effective programs, as captured by the general call for "What Works?" (www.whatworks.ed.gov). Many researchers, skeptical of this call, argue that the real question should not be "whether an intervention works," but instead, "what works, when, for whom, and under what circumstances" (Bryk, Gomez, & Grunow, 2011, p. 151). A shift to focus on specific outcomes that accrue under precise conditions and with specified resources rests on the assumption that educational outcomes result from (and often require) adaptations to circumstances; therefore to seek simple broad scientific principles or rules that apply across the board to a curriculum is of limited value. For example, Bryk, Gomez, and Grunow noted, "Treatises on modern causal inference place primacy on the word 'cause' while largely ignoring concerns about the applicability of findings to varied people, places and circumstances. In contrast, *improvement research* must take this on as a central concern if its goal is useable knowledge to inform broad scale change" (Bryk et al., 2011, p. 150; italics added).

Shifting the question to "what works, when, for whom, and under what conditions?" has profound implications for the meaning of effectiveness as a dependent variable. In establishing causal models, one determines, within the restrictions of a particular study's conditions, if an effect, controlling for other factors, can be

Based on a paper originally presented to the National Academies Board on Science Education and Board on Testing and Assessment for the conference, "Highly Successful STEM Schools or Programs for K-12 STEM Education: A Workshop"

J. Confrey (✉) • A. Maloney
North Carolina State University, Raleigh, NC, USA
e-mail: jconfre@ncsu.edu

J.A. Middleton et al. (eds.), *Large-Scale Studies in Mathematics Education*,
Research in Mathematics Education, DOI 10.1007/978-3-319-07716-1_17

rigorously linked to an antecedent condition. Instead of the causal structure of the phenomenon of interest, this focuses the study on its internal validity—hence "cause" and "effect." While studies typically can and do produce small but statistically significant effects, they often have nested within them more interesting conjectures about interactions and relationships among causes, effects, and co-relational phenomena. Those who demand causal design are often silent on the necessity of replication, which, strictly speaking, is required to realize the benefits of randomization; one study alone does not ensure generalizability.[1] Furthermore, in pursuit of causal models, researchers often rely on average effects, but doing so strips away more robust and potentially relevant differences that may apply to subsets of the whole.

Attempt to identify simple causal chains, and focus on strict control of study conditions, can lead those who attempt to implement research results astray. Too many policy makers and practitioners assume that an established treatment, as "cause," can be simply or directly applied to a practice and guarantee an effect. Perhaps some lack awareness that a study's internal validity does not assure its external validity. Consequently, most studies leave the practitioners themselves responsible to evaluate whether that study generalizes to their own settings. How they are supposed to do this responsibly is seldom addressed.

Regarding randomized field trials as the sole source—or the trump card—of assertions of a program's "effectiveness" poses a major dilemma. They are typically very costly, difficult, and time-consuming to conduct, leaving the public continually awaiting a sufficient set of scientifically "proven" empirical results. Randomized field trials seldom provide timely information in a quickly evolving context (especially for technology-enhanced programs)—by the time the results are available, the program typically is either outdated or has been significantly revised.

In contrast, in this chapter we argue that by developing and deploying explicit means of *engineering* [*for*] *effectiveness*, communities of practitioners and researchers can conduct ongoing local experiments at scale, which incorporate adequate design, as well as technologically enabled tools for real-time data collection and continuous analysis of patterns and trends.

Approaches similar to engineering [for] effectiveness have emerged under a variety of names. The study of complex and dynamic systems (Maroulis et al., 2010) has been addressed variously through continuous improvement models (Deming, 2000; Juran, 1962), implementation research (Confrey, Castro-Filho, & Wilhelm, 2000; Confrey & Makar, 2005), improvement research (Bryk et al., 2011), a science of improvement (Berwick, 2008), and Design-Educational Engineering and Development (DEED) (Bryk, 2009; Bryk & Gomez, 2008). When examined through the lenses of these various models, it becomes evident that the improvement of

[1] One can, of course, throw five heads in a row in a toss of five coins; only by replicating this experiment multiple times can one be certain that a generalized result of 50–50 emerges. Hence one experiment can never establish any form of cause and effect, a fact too frequently overlooked in discussions of the benefits of randomized field trials.

educational outcomes requires reexamination of approaches to just what is meant by "effectiveness." The following four ideas can be used to frame that reexamination:

1. *Education must be viewed as a complex system, with interlocking parts*. Study of a complex system requires one to locate a focus of attention without losing sight of the broader context. One must also attend to a variety of scales of events and time (Lemke, 2000). For instance, while summative and periodic results (large scale, longer time frames) may be useful as broad but crude policy levers that help in identifying trends and sources of inequities, formative results (smaller grain size, shorter time frames) are crucial to drive classroom processes forward. Measurement issues will vary according to these varying levels and orders of magnitude of phenomena (Lemke, 2000; Maroulis et al., 2010).
2. *Bands and pockets of variability should be expected, examined for causes and correlates, and used as sources of insight, rather than adjusted for, suppressed, or controlled*. Discerning how to characterize variability and its significance is key to knowing how to characterize a particular case or instance. "Most field trials formally assume that there is some fixed treatment effect (aka a standardized effect size) to be estimated. If pressed, investigators acknowledge that the estimate is actually an average effect over some typically nonrandomly selected sample of participants and contexts. Given the well-documented experiences that most educational interventions can be shown to work in some places but not in others, we would argue that *a more realistic starting assumption is that interventions will have variable effects and these variable effects may have predictable causes*" (Bryk et al., 2011, p. 24). Stephen J. Gould (1996) made a similar argument in *Full House*, discussing the diagnosis of his mesothelioma. He pointed out that, as a patient, broad survival rates were of less use to him than the smaller bands of variability that more specifically characterized his situation and provided more insight into his chances of survival. Similarly, analytic frames must therefore take into account patterns of antecedent and coincident conditions that mark potential variation in outcomes. "Effectiveness" is not unifaceted, but only understandable in the context of these causal networks.
3. *Causal or covarying cycles with feedback and interaction are critical elements of educational systems, in which learning is a fundamental process*. Furthermore, feedback loops mediate social cues and their behavioral outcomes, so one expects emergent phenomena (Maroulis et al., 2010). There is a contrast between constructions of simple cause-and-effect on the one hand, and causal cycles on the other. In the case of simple cause-and-effect, one assumes that a curriculum is implemented, and produces knowledge growth among students. In the case of causal cycles, the implementer is already aware of the types of outcomes measured, based on prior feedback, and implements and adapts the curriculum simultaneously, thereby raising the question "to what extent did the curriculum cause the effects, and to what extent did the outcome measures (through anticipation or feedback) cause the curriculum adaptation, and thence the effects (a causal cycle)?"
4. *Education should be treated as an organizational system that seeks, and is expected, to improve continuously*. As such, it is comprised of actors who must coordinate their expertise, set ambitious goals, formulate tractable problems

(Rittell & Webber, 1984), negotiate shared targets and measures of success (Bryk et al., 2011), make design decisions within constraints (Conklin, 2005; Penuel, Confrey, Maloney, & Rupp, 2014; Tatar, 2007), and develop and carry out protocols for inquiry. In such a "networked improvement community" (Bryk et al., 2011), one can position the causal cycles under investigation as "frames of action." Continuous improvement depends on iterations of collecting relevant, valid, and timely data, using them to make inferences and draw conclusions, and take deliberate actions, which, in turn, provide a refined set of data upon which to approximate some meaningful set of outcomes.

In analyzing the following examples of studies of curricular effectiveness, we will refer to these components as (1) complex systems with interlocking parts, (2) expected bands of variability, (3) focus on feedback, causal cycles, interactions, and emergence, and (4) continuous organizational improvement. We seek to show how these four components can inform us in designing and engineering [for] effectiveness and scale.

In this article, we focus our discussion of the redefinition of effectiveness research in the context of curriculum design, implementation, and improvement. We point out complementarities with the call for a change in "protocols for inquiry," in which Bryk et al. (2011) locate a "science of improvement" between models of traditional translational research and action research:

> In its idealized form, translational research envisions a university-based actor drawing on some set of disciplinary theory (e.g., learning theory) to design an intervention. This activity is sometimes described as "pushing research into practice" (see for example Coburn & Stein, 2010, p. 10). After an initial pilot, the intervention is then typically field-tested in a small number of sites in an efficacy trial. If this proves promising, the intervention is then subject to a rigorous randomized control trial to estimate an overall effect size. Along the way, the intervention becomes more specified and detailed. Practitioner advice may be sought during this process, but the ultimate goal is a standard product to be implemented by practitioners as designed. It is assumed that positive effects will accrue generally, regardless of local context, provided the intervention is implemented with fidelity.
>
> In contrast, action research places the individual practitioner (or some small group of practitioners) at the center. The specification of the research problem is highly contextualized and the aim is localized learning for improvement. While both theory and evidence play a role, the structures guiding inquiry are less formalized. Common constructs, measures, inquiry protocols and methods for accumulating evidence typically receive even less emphasis. The strength of such inquiry is the salience of its results to those directly engaged. How this practitioner knowledge might be further tested, refined and generalized into a professional knowledge, however remains largely unaddressed (Hiebert, Gallimore, & Stigler, 2002).
>
> A *science of improvement* offers a productive synthesis across this research-practice divide. It aims to meld the conceptual strength and methodological norms associated with translational research to the contextual specificity, deep clinical insight and practical orientation characteristic of action research. To the point, the ideas … are consistent with the basic principles of scientific inquiry as set out by the National Research Council (Shavelson & Towne, 2002, p. 22).
>
> Entire quote from Bryk et al. (2011), pp. 148–149 (italics added).

By defining a perspective of "engineering [for] effectiveness" we suggest that communities of practice, at a school, district, or state level, can build on what has

been learned from studies of curricular effectiveness. We review several studies associated with effectiveness research from mathematics education, and reinterpret their results and implications. Our focus will be the challenge of improving the *instructional core* (derived from Elmore, 2002; Cohen, Raudenbush, & Ball, 2003), by which we refer to *the daily classroom activities of implementing a curriculum, carrying out instruction, and applying formative assessment practices*.

Intervening at the Instructional Core

A model of the instructional core is shown below, in which the instructional core is situated between the Common Core State Standards (CCSS) and High Stakes tests. In general, standards (at the state level) and high stakes summative assessments are the "bookends" that constitute the accountability system. Policy levers of *No Child Left Behind* are designed to drive accountability through external pressure (sanctions and incentives) and to shed light on discrepant subgroup performances or lack of annual yearly progress. However, the bookends neglected and/or avoided the instructional core in relation to professional development, pedagogy, and classroom assessment. The absence of common standards fragmented the attention to curriculum (Reys, Reys, Lapan, Holliday, & Wasman, 2003). By squeezing the educational system by way of the bookends, the accountability system during the past decade and more produced some performance gains from the system. However, it failed to strengthen the instructional core with respect to capacity, unintentionally promoted a narrowing of the content taught, and, while calling for the use of "best practices" it failed to identify means to establish the credibility of practices identified as "best."

We chose the instructional core as a focus for this chapter because it can be readily recognized as a complex system, and should be analyzed as such. Its identifiable interlocking parts act at different levels of the educational system, from the standards and the summative tests to classroom practices and formative feedback. While one could view the instructional core as a temporal sequence of (a) curricular selection, (b) some level of professional development, (c) followed by implementation and assessments (both formative and summative), each of these components also interacts with and can generate (organized and explicit, or de facto and inadvertent) feedback to the other components. For instance, frequent formative results provide regular feedback to classroom practices, while data from high-stakes tests provide intermittent or periodic feedback and a much cruder level of nonspecific but severe institutional pressure. Resulting practices can be customized for groups according to curricular requirements and feedback from measures of learning. We have left the structure of improvement communities intentionally vague. The generality of the model allows for diverse institutional structure, as well as informal relationships among actors. Networked improvement communities are not explicitly identified in Fig. 1, but could be configured such that communities of practice could include practitioners, researchers, and administrators, who can plan together, share experiences, analyze data patterns, and discuss how to revise and adapt instructional approaches, curriculum, and schedules.

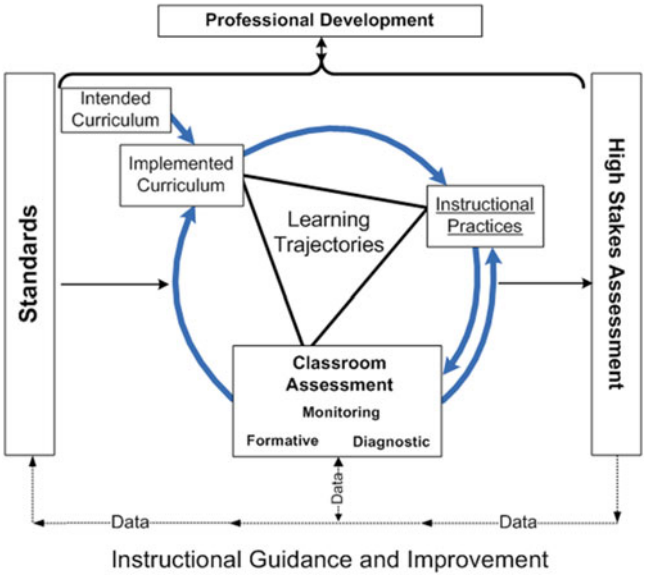

Fig. 1 Model of Classroom Educational System, illustrating position of the instructional core between the accountability "bookends" (Confrey & Maloney, 2012)

The adoption of the CCSS, by most of the states, positions educational communities, writ large, to create policy approaches and to reconsider the importance of focusing on improving the instructional core without overly constraining innovation, over-regulating curricular choice, or de-skilling teaching. By examining exemplars of research on the effectiveness of curricular programs, classroom instructional pedagogies, and formative assessment practices, and defining how these results can inform efforts to engineer [for] effectiveness, researchers could potentially jumpstart a movement towards school improvement in STEM disciplines.

Curricular Effectiveness Studies

"Curriculum matters" (Schmidt et al., 2001). It is the means by which students gain access to the knowledge and skills in a field and also the primary way they are attracted to pursue and persist. Since the publication of the NRC report that one of us (Confrey) chaired, *On Evaluating Curricular Effectiveness* (NRC, 2004), many mathematics educators have worked diligently to strengthen and improve research on and evaluation of curricular effects. That NRC report's framework called for designing evaluations to examine three components of curriculum: the program theory (through content analyses and comparison to standards), the program implementation (through a study of the program's implementation including

professional development and on-site staging, resources, and support), and the program outcomes (for alignment to standards and achievement of intended results). The report argued for the use of multiple methods in judging effectiveness, including content analyses, comparative studies, and case studies. It also called for the use of multiple and more sensitive outcome measures, and made a case for increased independence of evaluators, precise identification of comparison programs, and better measures of implementation. We have selected three studies that have taken these recommendations seriously and have moved research to a higher and more nuanced level. We report on their approaches, their principal findings, and identified limitations, and discuss how these can be interpreted to provide a solid foundation to next generation efforts to "engineer [for] effectiveness," that is, to iteratively design, monitor, analyze, and adjust components of the instructional core for more effective teaching and learning.

Case One: Single-Subject vs. Integrated Mathematics (COSMIC Study)

New studies of curricular implementation have advanced our understanding of curricular effectiveness. One such study is "Comparing Options in Secondary Mathematics: Investigating Curriculum," (COSMIC) (Grouws et al., 2010, 2013; Tarr, Grouws, Chávez, & Soria, 2013). The COSMIC project compared the effects of two curricula, one subject-specific and one integrated, on student learning in high school mathematics. Among the important contributions of this large quasi-experimental study was the development of multiple measures of curricular implementation and new types of curricular-appropriate tests to study the effects of curricular content organization on student learning in the first 2 years of high school mathematics.

A goal of the COSMIC study was to improve understanding about the relationships among curricular organization, curricular implementation factors, and gains in student learning. The study's research questions were the following for year 1 (Algebra 1 compared with Integrated Course 2) (Grouws et al., 2013). The research questions for year 2 student learning (Geometry compared with Integrated Course 2) were similar (Tarr et al., 2013):

1. Is there a differential mathematics learning effect when secondary school students study from an integrated textbook (Course 1) and when students study from a subject-specific textbook (Algebra 1)?
2. What are the relationships among curriculum type, curriculum implementation, and student learning? In particular,
   (a) What curriculum implementation factors are associated with high school students' learning in first-year mathematics courses?
   (b) What teacher characteristics and practices are associated with high school students' learning in first-year mathematics courses?

Participating schools all offered both a traditional high school curriculum (algebra 1, geometry, algebra 2) and an integrated curriculum (CORE-Plus), between which students chose freely (i.e., were not tracked by ability level).[2] In all, 11 schools in six districts across five regions of the country participated. The schools' student population demographics varied widely (e.g., the proportion of students eligible for free and reduced lunch (FRL) ranged from 17 to 53 % across the schools in the 2 year-levels of the study). Three distinct measures of student achievement (dependent variables) were used. Student results on those measures were compared to an index of prior achievement based on state-mandated eighth grade tests, normed against NAEP to provide comparability of student preparation across classes and states. Dependent measure data were analyzed using hierarchical linear modeling (HLM).

"*Fair test*" *as essential measure for comparing curricula*. The study generated a number of significant advances in research on curricular effectiveness. Researchers incorporated expertise in mathematics content and in learning effectively to design and select the study's outcome measures. They used multiple outcome measures: for each year level, two tests were designed specifically for the project (one of common content and another of reasoning and problem solving). The third test was a nationally normed standardized multiple-choice test, the Iowa Test of Educational Development [ITED]: Mathematics: Concepts and Problem Solving Form B, levels 15 (year 1) and 16 (year 2).

Drawing heavily on the NRC report's recommendations, the project began with content analyses of the printed curricula used in the schools. The project team then designed a "fair test" (NRC, 2004), "developed with the deliberate goal of not being biased towards either of the two curriculum programs studied" (Chávez, Papick, Ross, & Grouws, 2010, p. 4). To create the fair test, items were developed collaboratively by a research mathematician and mathematics educator to include content common to both curricula (i.e., that all students could be expected to have had the opportunity to learn (OTL) in both curriculum types) (Chávez et al., 2010). Items were constructed response instead of multiple choice, and often used realistic situations. Iteratively developed, the items were designed to permit adequate space and time for students to reveal potentially subtle differences in their understanding of underlying constructs, were piloted to ensure high face validity of the items, and were scored using a rubric construction method that assured careful internal and external review, and inter-rater reliability. An overall intent of the fair test was to allow inferences to be made about "student knowledge on constructs underlying the

[2] In the COSMIC year 1 study, the textbooks used were Core-Plus Mathematics Course 1 (Coxford et al., 2003) [20 classes], the integrated curriculum, and 5 different single subject curricula, Glencoe Algebra 1 (Holliday et al., 2005), [10]; McDougal Littell Algebra 1 (Larson, Boswell, Kanold, & Stiff, 2001) [6]; Holt Rinehart & Winston Algebra 1 Interactions (Kennedy McGowan, Schultz, Hollowell, & Jovell, 2001) [4]; and Prentice Hall Algebra 1 (Bellman, Bragg, Charles, Handlin, & Kennedy, 2004) [2 classes]. In the year 2 study, the textbooks were Core-Plus Course 2 (Coxford et al., 2003), and one of the following SS curricula Glencoe-McGraw Hill (Boyd, Cummins, Malloy, Carter, & Flores, 2005), Prentice Hall (Bass, Charles, Jonson, & Kennedy, 2004), Holt (Burger et al., 2007), and McDougal Littell (Larson, Boswell, & Stiff, 2001).

content of the tasks on the test, rather than merely…about student ability only on the tasks themselves" (Chávez et al., 2010, p. 8).

Treatment integrity (*multiple measures of implementation fidelity*). COSMIC researchers also intensified the degree to which they addressed *treatment integrity* (NRC, 2004) using multiple data sources to gauge teachers' implementation of curricular materials. These included Table of Contents Records, Textbook-Use Diaries, an Initial Teacher Survey, a Mid-course Teacher Survey and observations using a Classroom Visit Protocol (McNaught, Tarr, & Sears, 2010, p. 5). The research team was able to examine critical factors such as professional development, familiarity with standards, and teachers' distribution of classroom time among lesson development, non-instruction, practice, and closure. In a sub-study across two consecutive school years, the authors defined, studied, and compared three related indices of curricular implementation: OTL Index, "the percentage of textbook lessons taught without considering teachers' use of supplemental or alternative curricular materials" (the topics or lessons that students thus had an OTL); Extent of Textbook Implementation (ETI) Index, to provide a sense of how closely the textbook was related to the implemented curriculum (a weighted index to indicate the extent to which lessons were taught directly from textbook or with varying degrees of supplementation, including lessons that were not taught at all); and Textbook Content Taught (TCT), representing the extent to which teachers, *when teaching textbook content*, followed their textbook, supplemented their textbook lessons with additional materials, or used altogether alternative curricular materials (McNaught et al., 2010; Tarr et al., 2013). Differences in all these indices could then be folded into the analysis of factors contributing to student learning outcomes.

For example, for the entire study (3 years), for OTL 60.81 % (19.98 SD) of the content of the integrated textbooks was taught while 76.63 % (17.02 SD) of the content of the subject-specific textbooks was taught. The ETI index showed that across all teachers, "(35 %) of the textbook content was taught primarily from the textbook, …(21 %) of the content was taught with some supplementation, a small portion (12 %) was taught from alternative resources, and 32 % of the content was not taught at all." (Overall ETI values were 50.37 (20.20) for integrated and 57.15 (18.94) for single subject (SS)). The TCT index showed that when integrated content was taught, it was more frequently directly from textbook (59 %) as compared to when subject-specific content was taught (46 %). Furthermore, 28 % of integrated lessons were taught with some supplementation, while 33 % of subject-specific lessons were so taught (overall, 81.96 (14.50) for integrated, 74.93 (18.29) for (SS)) (McNaught et al., 2010, pp. 12–13). However, there was considerable variation in curriculum implementation between year-levels 1 and 2. Year 1 teachers' implementation index values were much closer, and higher than the summary values for all teachers in the study, whereas year 2 teachers had wide variation in OTL and ETI, with values for teachers of year 2 integrated much lower that those for teachers of SS. This study provided a major opportunity to interpret student learning outcomes in relation to variation in implementation fidelity, and led to the conclusion that unless information on textbook use is considered, interpreting findings on student learning outcomes related to a curricular treatment can easily lead to unfounded conclusions.

Teacher, *classroom*, *and student data*: *explaining variation in student outcomes*. The COSMIC project design required the accumulation of a wide variety of student- and teacher/classroom-level factors as potential moderators of curricular effects (eventually analyzed using HLM). The project gathered extensive teacher-level data (nearly 30 variables) from an initial and mid-year teacher survey, teachers' self-reports on curriculum implementation (the three indices developed from Table of Contents records), and classroom observations. The teacher data were subjected to principal components analysis and eventually were reduced to seven key teacher-level factors that explained approximately 70 % of the variance in the original data set. The factors clustered around two themes: curriculum implementation (the classroom learning environment, implementation fidelity, use of technology, and OTL) and teacher characteristics (their adherence to and practice of NCTM Standards-based instruction, their teaching and curriculum experience, and professional development) (Grouws et al., 2013; Tarr et al., 2013). Student achievement on the dependent measures was subsequently examined for their relationship to the student- and teacher (classroom)-level factors.

Overall, the extent and richness of student, teacher implementation, and classroom observation data gathered through the curriculum evaluation model, COSMIC was able to develop a more textured understanding of curricular effectiveness than had been accomplished to date.

COSMIC reported on student outcomes by adjusting the scores for students' prior achievement and then aggregating them by teacher (Tarr et al., 2010). The outcomes were reported as residualized gain scores by *teacher*, in recognition that the unit of analysis should not be the individual student (NRC, 2004).

For year 1 course comparisons, the following represent some of the noteworthy results: over all three measures of learning, (1) while several student-level variables were statistically significant predictors of students performance, consistent with previous studies (prior achievement, gender, ethnicities, and special needs); (2) the organization of the curriculum was the single most important factor in the modeling of performance on the tests, with large effect sizes for the test of common objectives and the problem solving test, and somewhat smaller for the Iowa Test of Educational Development. Numerous other factors were statistically significant predictors of performance on some, but not all the measures, and there were statistically significant interactions of factors for performance on one or another of the tests.

For the study of the year 2 courses (Geometry and Core-Plus 2), similar results were seen. However, while many of the individual student level variables were statistically significant predictors of performance on one or more of the measures, for year 2 course students, the CPA index was by far the strongest predictor, with effect sizes greater than 0.5 on all three measures. And perhaps most notably, the curriculum type had little effect on the outcomes on either the test of common objectives or the problem solving test for this year level, but the integrated curriculum had a significant favorable effect on performance on the Iowa Test.

An examination of partial correlations found that when controlling for %FRL, the magnitude of the correlation between Curriculum Type and student outcomes became significantly significant in favor of the integrated curricula, for all three tests.

OTL independent of curriculum was also significantly and positively correlated with higher performance on all three outcome measures.

The importance of OTL is substantially reduced with the partialing out of Class-level %FRL, suggesting that %FRL and OTL may be closely related. While it is possible that the relationship between OTL and %FRL may be attributable to a differential (slower) pace of content coverage in classes with higher percentages of FRL students, the result—less opportunity to have learned the material—suggests there is a need for active intervention to address this resulting inequity of opportunity (note: the study did not address school effects). Since teachers of integrated curricula covered significantly less textbook content than teachers of subject-specific curricula, a difference in coverage (as a percent of the curriculum topics that were taught) may have moderated the effect of Curriculum Type. Further, this study indicates that by controlling for OTL and %FRL, one can more carefully measure the impact of curriculum on student learning.

The year 1 study showed students studying from the integrated curriculum outperforming students studying from single subject curricula on all three measures, but the year 2 results were less clear-cut—while there was a significant effect of the integrated curriculum on the standardized test, there was no significant effect of curriculum on the two project-developed tests. However, prior achievement was a very strong predictor of student learning on all three tests, for both year-level studies. The COSMIC study produced many other results, showing more subtle correlations of student- and teacher-level factors with the student outcomes, as well as more interesting pairwise interactions, than can be discussed here.

No simple answers. Policy makers, administrators, and even practitioners ask whether an integrated program generates (causes) better, worse, or the same learning (outcomes) as a single-subject curriculum. Overall, the COSMIC study illustrates that it is unwise to expect curricular studies to yield such simple answers about curricular effectiveness. The authors note further that the study generalizes only to schools that offer both curricular options, and only if student choice (rather than tracking decisions) determines which students enroll in the two curricula. Unless these conditions are met, the study offers no definitive answer.

However, the COSMIC study yields far more contributions and insight than its statistical "curricular effects." These insights reflect the nature of complex systems. Consider what one could learn from this study that pertains to "engineering [for] effectiveness." COSMIC researchers have provided a protocol for creating and using appropriate multiple outcome measures to compare two curricula, first determining the extent to which they cover the same material, and, second, by selecting common topics by which to create a "fair test." If a district instead wants to know how curricula affect performance on a measure that assesses common standards, such as the Common Core State Standards, the study describes how to recognize and select such a reliable and valid test. It also illustrates how the choice of outcome measure interacts with the curriculum's effects. In systems with causal cycles, measures can also drive the system towards improvement, so such insights into analyzing outcome measures can facilitate important discussions of high-priority goals.

The COSMIC study also illustrates the value of disaggregated data for revealing and identifying relevant bands of variability that may warrant closer inspection. The study reinforces many other findings that the higher the percentage of students eligible for FRL, the lower the OTL. However, OTL was typically a significant moderating effect on student performance on one or more of the tests, while FRL did not have a statistically significant effect. The study further suggests that the effects of the curriculum in favor of integrated math become more evident when FRL measured at the classroom level is partialed out. Arguably, these findings suggest that using integrated mathematics curriculum could be a considerable educational benefit to students with low SES, but may nonetheless require teachers to receive substantial assistance to increase students' "opportunity to learn." At the class-level, experience (in teaching, and in teaching the specific curriculum) was a significant moderating factor, with students taught by experienced teachers (3 or more years of experience) achieving more than students of inexperienced teachers.

Practitioners and policy makers ask whether an integrated program generates better, worse, or the same outcomes as a single-subject approach. The COSMIC study design reflected the complex nature of curriculum organization and implementation, illustrating that it is unwise to expect curricular studies to yield simple general answers. It provides further insight into the inherent weakness of any simple statement that a curriculum is more or less "effective" than another.

The COSMIC study informs readers about the complexity of curricular implementation, as comprising the classroom learning environment (focus on sense-making, reasoning about mathematics, students' thinking in instruction, and presentation fidelity), implementation fidelity (ETI, TCT, textbook satisfaction), technology and collaborative learning, and OTL. These results suggest that in addition to focusing on OTL, school leaders need to help teachers to understand the standards, focus on student reasoning and sense-making, and learn to achieve closure during instruction. In relation to Fig. 1, this suggests that the factors involved in implementation rest within the circle and that their connections to the two bookends in the drawing provide guidance and feedback.

Overall, the COSMIC study results so far suggest that the use of integrated mathematics in year 1, at least, and possibly year 2, may offer considerable learning opportunities for students across the spectrum. Implementation of the integrated curriculum is not a simple matter. In a North Carolina study, based on an analysis of reports from content specialists' monthly observations of teachers' practice, we found that teachers using an integrated mathematics curriculum with low SES students often lost a great deal of time in transitioning to problems in integrated math, tended to be reluctant to turn over authority to students, and missed opportunities to establish closure (Krupa & Confrey, 2010). In a case study of one teacher, instructional coaches engaged in specific and targeted activities with the classroom teacher, and the teacher was able to transform her instructional practices and in fact became a role model for new teachers at the school (Krupa & Confrey, 2012). In studying multiple cases of teachers in these schools, Thomas (2010) showed that providing adequate support to teachers *can* transform practice, but that this is very difficult to accomplish, due to weakness in teacher knowledge and to those teachers' views of instruction. Disentangling these complex relationships may be easier to accomplish

in studies seeking improvement over time in the context of smaller studies. Our studies, funded as a Mathematics-Science Partnership through a state department of education, permitted us to form a networked community for improvement, among University researchers, faculty from the state School of Science and Mathematics, a semi-autonomous school organization committed to improving rural education, and—critically—in-service teachers and principals. Our efforts could have greatly benefitted from richer and more continuous data sources informed by research tools such as those developed for COSMIC.

Case Two: Comparing Effects of Four Curricula on First- and Second-Grade Math Learning

A second major study on curricular effectiveness provides another example of the potential contributions of nuanced study that goes beyond simple claims of cause and effect. The study "Achievement Effects of Four Early Elementary School Math Curricula: Findings for First and Second Graders" (Agodini et al., 2009, 2010), examined whether some early elementary school math curricula are more effective than others at improving student math achievement in disadvantaged schools (57 % of schools included in the study were school-wide title 1 eligible, compared to 44 % nationwide). The authors (R. Agodini, B. Harris, M. Thomas, R. Murphy, L. Gallagher, and A. Pendleton) studied the implementation of four contrasting curricula: *Investigations in Number*, *Data*, *and Space* ("*Investigations*"), featuring a student-centered approach encouraging metacognitive reasoning and drawing on constructivist learning theory (Wittenberg et al., 2008), *Math Expressions*, blending student-centered and teacher-directed approaches to mathematics (Fuson, 2009a, 2009b), *Saxon Math* (*Saxon*), a scripted curriculum relying heavily on direct instruction in procedures and strategies with guided and distributed practice (Larson, 2008), and *Scott Foresman-Addison Wesley Mathematics* (*SFAW*), a basal curriculum that combines teacher-directed instruction with a variety of differentiated materials and instructional strategies (Charles et al., 2005a, 2005b). *Math Expressions* and *Investigations* are both "reform" curricula whose development had been either initially funded by the National Science Foundation or based on research with considerable NSF funding. A total of 473 districts were invited, but only 12 agreed to participate in the study—a recruitment rate of 2.5 % (Agodini et al., 2010, p. 10).[3] In all, 109 first-grade classes and 70 second-grade classes were randomly assigned to a curriculum within districts.

[3] The authors acknowledge that this low rate leaves an "open issue, which cannot be examined with the study's data, is whether the potential differences between participating and nonparticipating sites are related to the study's findings" (p. 14). The conditions of the study, in particular the need for a district to assign different curricula to schools at random, could be viewed by many districts as unacceptably burdensome or arbitrary, and conflict with their own judgment about the most useful curriculum, or simply be at odds with district policy and/or fiscal constraints.

The study addressed three broad questions (Agodini et al., 2010, pp. 4–5):

1. What are the relative effects of the study's four mathematics curricula on first- and second-graders' mathematics achievement in disadvantaged schools?
2. Are the relative curriculum effects influenced by school and classroom characteristics, including teacher knowledge of math content and pedagogy?
3. [Based on subsequent statistical analysis—] What accounts for curriculum differentials that are statistically significant?

Student mathematics achievement outcomes were based on fall and spring administrations (pre- and post-administrations) of the ECLS-K assessment (developed for the National Center for Education Statistics' Early Childhood Longitudinal Study-Kindergarten Class of 1998–1999), a nationally normed adaptive test.[4] Other data were drawn from student demographic and school data, teacher surveys, study-administered assessments of math content and pedagogical content, and scales of instructional practices and approaches derived from classroom observations.

The study results were reported as pairwise comparisons of the curricula, for student outcomes (six pairwise comparisons) for each grade. After 1 year of schools' participation, average first-grade math achievement scores of *Math Expressions* and *Saxon Math* students were similar and higher than those of both *Investigations* and *SFAW* students. In first-grade classrooms, average math achievement scores of *Math Expressions* students were 0.11 standard deviations higher than those of *Investigations* and *SFAW* students. These results were interpreted to mean that, for a first grader at the 50th percentile in math achievement, the student's percentile rank would be 4 points higher if the school had used *Math Expressions* instead of *Investigations* or *SFAW*. In second-grade classrooms, average math achievement scores of *Math Expressions* and *Saxon Math* students were 0.12 and 0.17 standard deviations higher than those of *SFAW* students, respectively. For a second grader at the 50th percentile in math achievement, these results mean that the student's percentile rank would be 5 or 7 points higher if the school used *Math Expressions* or *Saxon Math*, respectively.[5]

[4] The test is adaptive in that students are initially administered a short, first-stage routing test that broadly measures each student's achievement level. Based on the first-stage scores, students are then assigned one of three second-stage tests: (1) easy, (2) middle-difficulty, or (3) difficult. Scale calibration among the second-stage is accomplished through overlap of items on the second stage tests and item response theory (IRT) techniques, by which scores from different tests are placed on a single scale.

[5] Another way the authors interpreted these differences was to consider the average score gain by grade in the lowest quintile of SES on ECLS (16 points in first grade) and to convert the .1 effect size into points using the reported standard deviation of 10.9, getting a difference of 1.09 scale points. Comparing 1.09 to an average gain of 16 scale points, they describe an effect size of .10 as having an effect of 7 % of the gain over first grade. Thus the differences in student results reported between curricula account for between 7 and 14 % of the content as measured by the ECLS assessment.

This study, in some ways similar to the COSMIC study, examined curricular implementation, and reported on such factors as the use of the curriculum, the amount, frequency, and stated reasons for supplementation, the availability of support, amount of professional development, distribution of uses of instructional time, and focus on particular content areas. Teachers reported varying coverage of math content areas across the curricula. They determined that variation in coverage (number of lessons on a topic) of 19 out of 20 content areas was significantly different across all four curricula. However, in pairwise comparisons of the curricula, "there was no clear pattern [regarding] which curriculum [coverage] differences are significant." (p. 57): some pairwise differences in coverage were statistically significant and others were not.

For Table 1 below, we selected some implementation differences that could have affected student-learning outcomes. For instance, teachers received twice as much initial (voluntary attendance) professional development for *Expressions* than for other curricula (with >90 % of first-grade teachers reporting attendance at initial training sessions for all the curricula, but 80–97 % of second-grade teachers attending, with Math Expressions having the highest attendance rate). Teachers of *Saxon Math* taught math an additional 20 % of the time each week, teachers of *Math Expressions* used more supplementation materials while *Investigations* teachers used less, and 16.2 % of *Saxon Math* teachers and 21.1 % of *SFAW* teachers had taught with those curricula previously, compared to less than 6 % for each of the other two curricula. It should be noted that *Math Expressions* and *Investigations* are based more intensively on student-centered instructional approaches and represent pedagogical approaches that require extensive teacher preparation. Not surprisingly therefore, implementation reports show that higher percentages of first- and second-grade *Investigations* and *Expressions* teachers report feeling only "somewhat" or "not at all" prepared to teach their curriculum, compared to teachers of *Saxon Math* or *SFAW*.

The study's authors also conducted an analysis of the extent to which teachers adhered to their assigned curriculum. "Adherence" referred to the extent to which a teacher taught the curriculum using practices consistent with the curriculum developers' model. (In the NRC report, the philosophy of a curriculum's designers ("program theory") was distinguished from the application of the curriculum during implementation ("implementation fidelity").) The study measured adherence via a teacher survey and a classroom observation instrument, as the extent to which essential features of the assigned curriculum were implemented. The results shown in Table 2 suggest that teachers were more likely to adhere to designers' intentions in the *Saxon Math* program than in the *Expressions* program.

In an exploratory look at what might account for the relative curricular effects, the researchers examined the instructional practices that occurred across different curricular types (in contrast to adherence) based on the observational data. They conducted a factor analysis, yielding four factors: (1) student-centered instruction, (2) teacher-directed instruction, (3) peer collaboration, and (4) classroom environment. The analysis across the curricular pairs indicated that student-centered instruction and peer collaboration were significantly higher in *Investigations* classrooms

Table 1 Selected differences in implementation variables for different curricular implementations

	Respondents	Investigations	Expressions	Saxon Math	SFAW
Initial PD offered (prior to first day of school)	All teachers	1 day	2 days	1 day	1 day
Responded "Somewhat or not at all" adequately prepared after training (%)	First-grade teachers	23.3	33.7	16.0	10.0
	Second-grade teachers	23.2	56.1	16.4	9.1
Additional training offered	Reported by publishers	3–4 h every 4–6 weeks (group)	Twice a year (one-on-one meetings)	Once in fall (one-on-one meetings)	3–4 h every 4–6 weeks (group)
Follow-up training % participated/days	First-grade teachers	95.5/2.6	90.5/0.6	74.3/0.4	99/2.1
	Second-grade teachers	97.4/2.3	82.4/0.5	66.7/0.4	91.9/2.0
Supplemented curriculum (%)	First-grade teachers	14.8	32.1	24.8	27.5
	Second-grade teachers	11.7	55.6	30.5	24.6
Hours taught per week	First-grade teachers	5.1	5.0	6.1	5.3
	Second-grade teachers	5.4	5.5	6.9	5.5
Used assigned curriculum the previous year (%)	First-grade teachers	5.5	3.6	16.2	21.1

Table 2 Adherence to a curricular program's essential features, as percentage of features implemented

		Investigations	Expressions	Saxon Math	SFAW
Survey (self-report)	First-grade teachers	66 (3)	60 (4)	76 (1)	70 (2)
	Second-grade teachers	67 (3)	54 (4)	76 (1)	68 (2)
Observation of daily essential features	First-grade teachers	56 (2)	48 (4)	63 (1)	54 (3)
	Second-grade teachers	53 (2–3)	47 (4)	65 (1)	53 (2–3)
	Average	60.5 (3)	52.25 (4)	70 (1)	60.75 (2)

Numbers in parentheses indicate the relative ranks of the curricula for each row (pp. 65–67)

than in classrooms using the other three curricula. Teacher-directed instruction was significantly higher in *Saxon Math* classrooms than in classrooms using the other three curricula. The classroom environment did not differ across curricula.

Additional analysis indicated that some of these implementation factors act as mediators of achievement outcomes. The study's design, however, permitted examination of only one mediator at a time. This constraint meant that while differences in professional development for *Expressions* mediated the curricular effect, the authors could not relate this to the mediational effects of less prior experience with, and teachers' reports of less preparedness to teach, the curriculum. Likewise, *Saxon Math* teachers were reported to have had 20 % more instructional time, which mediated the *Saxon Math-SFAW* difference in curricular effect. The study design, however, does not permit assessment of *combined* effects of interactions between instructional time and likelihood of having taught a curriculum before. The authors interjected that a more rigorously designed study of mediation could disentangle the relationships among the mediators (p. 102). In any case, the examinations of implementation variables as mediators of curricular effects make it clear that one must always interrogate the results to understand the nuances in a causal study's assumptions and claims.

Among the many accomplishments of the Agodini et al. (2010) study was the identification of means to measure a considerable number of factors that comprise classroom practice. The study reports on a variety of factors that are worth examining, even if they were not demonstrated to be statistically significant contributors to differentiated curricular effects. For instance, the study reports low levels of mathematical knowledge on the part of elementary teachers, and while this was not differentially related to curricular effectiveness in the study, this is a persistent issue in elementary teaching that needs to be addressed. The study also makes a useful distinction between implementation factors that apply to *any* curriculum, and *adherence*, which pertains to the specific intentions of each curriculum's design; the latter is a curriculum-specific measure of teachers' fidelity of implementation of specific features/activities.

The Agodini study also exhibits limitations and threats to its validity: reliance on only a single student outcome measure (the ECLS-1 and -2), and the absence of a method to check the "fairness" of that outcome measure across the curricula. These are in contrast to the call in *On Curricular Effectiveness* for multiple measures and

for outcome measures that demonstrate "curricular validity of measures" (also called "curricular sensitivity") and "curricular alignment with systemic factors" (NRC, 2004, p. 165). Such a notable weakness with regard to the outcome measures unfortunately leads to major problems with the interpretation of the study's conclusions. The size of the curricular effect, 7–14 gain points on the scaled score, could be the result of a few key assessment items.

The study benefits—as an experimental study—from randomized assignment of curricula to teachers (classrooms) within the district, but this feature of the study came at a high cost to its external validity. Few districts were willing to randomly assign curriculum to teachers, calling into question the generalizability of the study's results. Secondly, conducting a study of curricular effectiveness during the first year of a curriculum's implementation, and providing only 1–2 days of professional development for primary teachers, must weaken confidence in the validity of comparisons of curricular effectiveness. For instance, reports of high levels of supplementation by *Expressions* teachers could be due to the teachers' use of prior, more familiar materials. If this were the case, should one draw the conclusion that *Expressions* itself was "effective" under these conditions?

Furthermore, the authors also described teachers' reports, for each curriculum, of the frequency of teaching particular content topics (whole numbers, place value, etc.). If an analysis of the test had been performed, and included in the study, one might have been able to discern patterns in the relationship between students' OTL the different topics and the outcome measure scores.

The Agodini et al. study offers far more insight into curricular effectiveness than is captured by its conclusions of "cause and effect." As with the COSMIC study, it makes progress on establishing implementation factors. Both studies identify similar factors, such as adherence vs. implementation fidelity, the use of student collaboration, and the use of general instructional approaches (student-centered and teacher-directed vs. standards-based instruction). Both examine content variations, one by conducting content analyses and then measuring OTL as teachers implemented, and the other by relying on teacher reports of number of lessons by content area and adherence to essential features of each curriculum. By designing different means of capturing the variations in these factors, these studies help us to progress in our understanding of the complexity of curricular use.

Case Three: The Relationship Among Teacher's Capacity, Quality of Implementation, and the Ways of Using Curricula

A third study, "Selecting and Supporting the Use of Mathematics Curricula at Scale," is a study of curricular impacts on implementation *quality* with respect to teachers' capacity and ways of using the materials, rather than a study of *effectiveness* (as based on student outcomes) (Stein & Kaufman, 2010). The study involved two districts using reform curricula, one using *Everyday Math* (*EM*) and the other using *Investigations*, in order to begin to answer the question of "What curricular materials work best under which kinds of conditions?" (p. 665)

The authors initially analyzed the two curricula with respect to the frequency of two kinds of high cognitive-demand tasks: "procedures with connections to concepts, meaning and understanding" (PWC) tasks and "doing mathematics" (DM) tasks (Stein, Grover, & Henningsen, 1996). They characterized PWC tasks as "… tend[ing] to be more constrained and to point toward a preferred—and conceptual—pathway to follow toward a solution," and identified 79 % of the tasks in *Everyday Math* as PWC tasks. They characterized DM tasks, in contrast, as "…less structured and [not containing] an immediately obvious pathway toward a solution" (Stein & Kaufman, 2010, p. 665), and identified 84 % of the tasks in *Investigations* as DM tasks. Based on these differences, they conjectured that it would be less difficult for teachers to learn to teach with *EM* than with *Investigations*. DM tasks are more difficult to implement faithfully, because they support open-ended discourse, which is often difficult to manage and require more of the teacher's own learning (Henningsen & Stein, 1997). In contrast, PWC tasks are more bounded and predictable, but are susceptible to "losing the connection to meaning" (Stein & Kaufman, 2010). Stein and Kaufman also documented that there is less professional development support embedded in the *EM* materials than in the *Investigations* materials, mirroring the conventional wisdom that teaching with the *EM* is less challenging than with *Investigations* curricula.

From these analyses, the study characterized *EM* as a low-demand, low-support curriculum, and *Investigations* as a high-demand, high-support curriculum. They then investigated how the implementation of these two contrasting reform curricula might differ, particularly with respect to the quality of implementation and its relationship to teacher characteristics.

Using classroom observations, interviews, and surveys, the researchers compared implementation of the two reform curricula in two districts that were similar in terms of the (high) percentage of students eligible for FRL (86 and 88 %). They studied implementation of the curricula by six teachers (one per grade level) in each of four elementary schools in each district over a period of 2 years. Observations (with examples) were conducted on three consecutive lessons in each of fall and spring, and coded for the extent to which teachers were able to (1) sustain high cognitive demand through the enactment of a lesson, (2) elicit and use student thinking, and (3) vest the "intellectual authority in mathematical reasoning," rather than in the text or the teacher. Together, high values on these three dimensions characterized high quality implementation.

Using surveys, observations, and interviews, they examined two teacher characteristics: teachers' *capacity* (defined as comprising years of experience, mathematical knowledge for teaching (MKT), participation in professional development, and educational levels) and their *use of curriculum* (teachers' views of the curriculum's usefulness, percentage of time teachers actually used the curriculum in lessons, and what teachers talked about with others in preparing for lessons—including non-mathematical details, materials needed for the lesson and articulation, and discussion of big ideas.)

In answering their first question, "How does teachers' quality of implementation differ in comparisons between the two mathematics curricula (*Everyday Mathematics* and *Investigations*)?" (Stein & Kaufman, 2010, p. 667), they found that teachers

from the district using *Investigations* were more likely to teach high quality lessons than teachers from the district using *Everyday Math* (it must be noted again, however, that this study did not investigate the relationship of instructional performances to student outcome performance, but rather the "less-studied link between curricula and instruction," p. 668). Teachers implementing *Investigations* were more likely to maintain the cognitive demand (6.7 > 4.9, on a scale of 2–8), to utilize student thinking more (1.1 > .5, on a scale of 0–3), and to establish norms for the authority of mathematical reasoning (1.2 > .4, on a scale of 0–2).

Their second question across the two districts and curricula was, "To what extent are teachers' capacity and their use of curricula correlated with the quality of their implementation, and do these correlations vary in comparisons between the two mathematics curricula?" (p. 667). The study found that most of the teacher capacity variables were not consistently and significantly related to the quality of implementation. In the district using *EM*, higher performance on MKT surveys was *negatively* correlated with the use of student thinking and with establishing the authority of mathematical reasoning in the classroom. In the district using *Investigations*, correlations of implementation quality with teacher capacity were positive but not significant. And while no clear relationship was found between either hours or type of professional development to implementation quality in the district using *EM*, in the district using *Investigations*, the amount of professional development was (positively) significantly correlated with all three components of implementation quality.

The study shows that implementation quality cannot be inferred from content *topic* analysis alone but depends also on the kinds of tasks (how the tasks are structured) that are used to promote student learning of those topics. It also suggests that implementation quality appears to relate more strongly to the extent of professional development support both facilitated by the district and afforded within the materials, than to other traditional capacity variables such as teachers' education, experience, and their MKT.

Across the two districts and curricula, the discussion of big ideas during lesson planning was the only teacher's-use-of-curriculum variable that was significantly and positively correlated to implementation quality components (and then to only two of those: attention to student thinking and authority of mathematical reasoning). Further, the authors reported that this tendency was more evident in the district using *Investigations*. In explaining this difference, they reported that teachers using *Everyday Math* indicated that frequent shifts in topics in the spiral curriculum tended to make identification of big ideas more difficult, while in *Investigations*, the "doing math" tasks led teachers to focus more on big ideas. These findings were somewhat counterintuitive because it had been thought that *Investigations* was more difficult to implement because it has a much higher percentage of DM tasks than does *EM*. The consideration of big ideas during instructional planning was strongly linked to high quality implementation of both curricula, and was also more engaged in by teachers implementing the curriculum that focused more extensively on DM tasks (*Investigations*).

Stein and Kaufman (2010) note that this work "provides evidence that one cannot draw a direct relationship between curriculum and student learning" (p. 688).

They asked, "…what elements of teacher capacity interact with particular curriculum features to influence what teachers do with curriculum. Thus, our focus is on *which program leads to better instruction under what conditions*" (p. 668, italics added). They suggest reorienting the concept of teacher capacity to incorporate the interaction of curriculum as tool with how teachers use the curriculum, and that study of how curriculum use over time interacts and promotes improved instruction would be a very fruitful path of research. In essence, by suggesting that "…curricula could be viewed not only as programs to be implemented, but as tools to change practice" (p. 688), they are suggesting that curricular effectiveness might eventually be considered not a static value or a product's claim, but instead a *process of improvement* of instruction through interaction between curriculum and how it is used.

Overall Conclusions from the Three Cases

Juxtaposing the three cases reviewed here provides an opportunity to synthesize advice for the conduct of future effectiveness studies. There has been a strong temptation in the calls for, and the interpretation of, effectiveness studies, to try to identify *some*thing that works—that is, to identify one or more curricula (or in fact, a single most effective curriculum for grade level or range) that can be adopted with the expectation of subsequent, direct major improvements in student learning outcomes. Calls for randomized field trials of curricular effectiveness have carried with them the assumption that such experimental designs will provide the best evidence that a curriculum is "effective." We have asserted, and taken together, the studies discussed in this chapter have shown this approach to be poorly conceptualized, underestimating the collective and cumulative impacts of coverage/OTL, implementation fidelity, and quality of instruction, not to mention differences in curricular structure, pedagogy, and content rigor.

We initially examined the three studies from a perspective of causality, to understand whether and how they might inform us about the results of implementing and comparing two or more curricula. Reviewing these cases, however, demonstrated how tentative causal conclusions are, and reminded us that all studies have flaws and limitations. The quest for the perfect curricular effectiveness study—and a quest for a single most-effective curriculum—is highly unlikely to yield results that are robust or extensive enough to guide practice. Each study provides insight into some *specific conditions* under which certain factors played roles and certain outcomes occurred, and that these depend on how constructs surrounding the implementation of the curricula were defined and measured.

The COSMIC study provides evidence of relative effectiveness of an integrated curriculum compared to subject-specific curriculum when students are provided a choice between those options. However, had multiple curricular alternatives been available, or had ability tracking been used to assign students to the two curricular options, the authors note, we do not know what the results of the study would have been. It could also be the case that if teachers of integrated curricula were able to

cover the same percentage of their text during a year as did teachers of a subject-specific curriculum, student performance in integrated math would be even stronger relative to that in the subject-specific curriculum. Practitioners choosing to apply this study to their own curriculum selection decisions must weigh these considerations, and must contextualize the results to develop expectations relevant to their own settings.

Similarly, the Agodini et al. study reported that students taught using *Expressions* outperformed students taught using the other curricula in both first and second grades, with the exception of students using *Saxon Math* in second grade. It is possible however, that this effect may have resulted from the extra day of professional development time or additional supplementation reported to be used by teachers for *Expressions*, or from increased instructional time, in the case of *Saxon Math*. Alternatively, it is feasible that all outcomes of this study could be attributable in large part to the degree of fit of the curricula with the single ECLS outcome measure used; if the study had used a different end-of-year assessment (or multiple measures as in the COSMIC study) the results might have been quite different. Another possible interpretation is that by examining the effectiveness of curricula for only the first year of implementation, the study's results were necessarily skewed in favor of *Saxon Math* and *SFAW*, which had higher levels of prior use and scripting, and that the student outcomes would evolve considerably over a longer study period (allowing more teacher experience with the assigned curricula), potentially re-ordering the student learning results.

All studies are open to multiple interpretations; most are subject to various predictable (or emergent) limits to generalizability. In the Stein and Kaufman study, for example, the stronger implementation quality of *Investigations* could have been attributed to its design of curricular tasks, affordances for focus on big ideas, and/or support for professional development. But perhaps the district that offered *Investigations* simply supported its implementation with higher quality, more extensive professional development.

These studies demonstrate further the complexity of curriculum's relationship to student learning. But some may ask whether the fact that these studies have some conflicting interpretations or that they do not provide generalizable recommendations, means that such investigations are not useful, or even a waste of time and money. Are such studies of limited importance because we cannot know whether a study's results will accrue in a setting that differs from the original—and may require a level of adaptation from the conditions for the study?

If the goal of curricular effectiveness studies were to decide unequivocally whether a single product—a curricular program and its related materials—can be simply dropped into classrooms and be expected to yield predictable learning gains, then these studies fail to establish curricular effectiveness. More to the point, however, these studies instead bolster the recognition that this assumption about curricular effectiveness and its generalizability is mistaken, a false apprehension. Their design is to provide more insight into the factors affecting effectiveness (and possibly leading to redefining the use of the term); their design and their execution make them highly valuable to that end.

We argue that these studies, especially when taken together, demonstrate why simple causality is an insufficient model for judging effectiveness of a curriculum. The message to be taken from them is that the instructional core is a complex system, that many things matter to the implementation of a curriculum and to the learning that students can accomplish with different curricula, and that what matters appears to depend in large degree on multiple factors, and different factors in different situations. Context matters—the extent to which one serves disadvantaged students, requires more resources, or requires teachers with stronger capacity or settings in which professional development is supportive and sustained. Resources matter. The quality of instruction, and the quality of curricular implementation, matter.

Most importantly, these studies contribute substantially to an understanding of the instructional core. By the very fact that the experts who conducted the studies have gained purchase on modeling the instructional core, they provide us insights into the complexity of instructional systems. They identify interlocking factors, loci of possible interventions, and a set of measures and tools that can help in the process of becoming smarter and wiser about *how curricular use in particular settings can improve instructional quality and student outcomes*.

These studies, we believe, provide the following lessons:

1. Outcome measures matter—and with the availability of Common Core State Standards, we have the opportunity and the responsibility to create a variety of measures in a cost effective way across districts and states (this is one of the premises of the Common Core assessment initiatives). The COSMIC study in particular reinforces the notion that implementation or effectiveness studies require multiple outcome measures which should (a) include measures that act as "fair" tests (Chávez et al., 2010) to ensure non-biased comparison of student performance on topics common to all curricula being examined, (b) include project-designed measures of reasoning and problem-solving, (c) be normed against relevant populations (e.g., college-intending students, ELS students) and used to make systemic decisions (such as statewide end-of-course exams or new assessments of Common Core State Standards), (d) assess the development of big ideas over time; learning progressions are one way to conceptualize coherent curricular experiences and their development over time, and (e) assess other dimensions of mathematics learning, such as the mathematical practices in the CCSS, student attitudes, or student intentions to pursue further study or certain STEM careers. The studies showed that the categories by which outcomes were disaggregated were critical, and were sensitive to interactions, such as by ethnicity and socioeconomic factors. At the least, therefore, relevant data gathered in relation to performance measures should include ethnic and racial diversity, gender, ELL, and FRL status, to support the investigation of relevant bands of variability in effects and outcomes.
2. Monitoring what was actually taught, and *why* it was taught, is crucial to making appropriate attributions in examining effectiveness. Monitoring should include measures of curricular coverage (such as OTL and adherence), and of the type and degree of supplementation (and the reasons for choices regarding these variables). Different methods of monitoring curricular coverage and supplementation included table-of-contents reports, surveys of relative emphasis, and textbook use diaries.

3. A better understanding of the factors involved in the implementation of curricula will add a wealth of insight to explanatory frameworks of curricular effectiveness. Some factors should directly reflect the extent to which implementation captures a designer's specific intent, while others should address qualities that apply across all curricula. These studies undertook many innovative methods of data collection: surveys, intermittent and extended classroom observations with various coding schemes, reports of instructional time usage, and interviews. In one case, these were coded in predetermined, theoretically relevant categories—maintaining cognitive demand of tasks, eliciting student thinking, and vesting authority in mathematical reasoning. In the COSMIC and Agodini et al. studies, high numbers of variables were identified a priori, and embedded in other instruments (teacher surveys, for instance). Modeling the factors that can explain the majority of observed variation for different levels of analysis (student, class/teacher, school level, for instance) requires statistical techniques (factor analysis, principal component analysis) to reduce the dimensionality of the vast amounts of resulting data, and to identify and sort critical variables into appropriate clusters (classroom learning environment, implementation fidelity, peer collaboration, technology use, student-centered instruction, and teacher-directed instruction). Selection of appropriate units of analysis, and hierarchical (multilevel) linear modeling were essential (COSMIC, Agodini et al.) for modeling the relationship and interactions of student- and teacher-level factors and their contributions to the dependent measures of student learning. Research on identifying, defining, and studying implementation factors (perhaps as latent variables) promises to continue to grow and add to our understanding of curricular effects.
4. Issues of teacher capacity and professional development are critical in judging curricular effectiveness, but not necessarily in a predictably simple or straightforward way; their influence varies depending in part on whether they are viewed as a resource within a curriculum and its implementation, or as a factor that interacts with implementation. Teacher capacity, a term that subsumes teacher MKT, experience, education, and professional development, did emerge as influential in two studies (COSMIC,[6] Agodini et al.). In the third study (Stein & Kaufman) however, its influence was mixed: while most teacher capacity factors did not correlate in a significant positive way with implementation quality in one district/curriculum, but some of the component factors correlated *negatively* and significantly in the other. In that study, the amount of professional development time, teachers' access to assistance and support, and the ways in which teachers used materials in planning (i.e., the degree of their focus on big ideas) and communicated with each other about curricular use emerged as the factors most closely associated with implementation quality. On the other hand, professional development was not significantly associated with student outcomes in the

[6] Though early results suggested that teacher experience was not significantly correlated with student outcomes (Tarr et al., 2010), completed HLM analyses of year 1 data revealed that teaching experience was a significant predictor of student outcomes on all three measures (Grouws et al., 2013; Tarr et al., 2013).

COSMIC study; teachers reported that they perceived little to no impact of their professional development activities on their teaching practices, in part, because they perceived the activities merely confirmed what they were already doing (Tarr et al., 2013), and the study measured professional development in terms of quantity not quality. The studies incorporate three perspectives on professional development and teacher capacity—one in which these factors are viewed as a resource for curricular implementation, one in which they could be viewed as a factor that interacts with implementation, and one in which curricular implementation is seen as a tool for changing capacity and as a source of professional development. To clarify how professional development and teacher capacity can relate to curricular implementation and effectiveness will require additional investigation.

5. How a study is situated in relation to educational structures and organizations may eventually be important at a meta-level of understanding curricular effects and the conclusions drawn. The location of each of the studies described here was driven by issues of experimental design—for instance, the availability of two curricular options without tracking (COSMIC), the dependence of a study on districts' willingness to randomly assign teachers to treatments (Agodini et al.), to support extended observations over 2 years, and to provide researchers with access to extensive teacher data. These issues were reported as features of the studies' designs, but over time such they may themselves emerge as organizational factors that are as important to curricular implementation as traditional organizational characteristics as governance, decision-making, funding, and data use.

Engineering [for] Effectiveness: Summary and Recommendations

These studies remind us how remarkably complicated are the interplay of curricula, instruction, classroom assessment practices, and professional development. They demonstrate that the instructional core is a complex system, exhibiting the first-order traits of complex systems including interlocking parts, bands of variability, feedback, causal cycles, interactions and emergent phenomena, and the need for focus on continuous improvement. It is incumbent on policy makers, system leaders, teachers, professional development and curriculum designers, and researchers, to treat the entire instructional core accordingly: as a complex system. We suggest therefore that rather than seek any grand causal effect from these or similar studies, one should use them to learn more about possible ways to model and improve the instructional core at the classroom, school, and district level throughout the USA.

We have come to believe that while curricular effectiveness has seemed an important focus for study, we suggest that with the instructional core as the complex system of which curriculum is one part, the focus for improvement should be the functioning of the system itself. The proposal that follows from this is to focus on how to engineer the *instructional core* for improved teaching and learning effectiveness—that is,

to iteratively design and improve our way to a greater understanding of the operation and strengthening of the instructional core. The studies recounted here have provided some critical elements of such an endeavor, including identification of a number of critical constructs, and creating measures to gauge and monitor them. Other researchers have argued for the importance of multiple methodologies (NRC, 2004) including such approaches as design studies, which are useful in identifying mechanisms to describe and explain interactions at the classroom level.

Many of the instruments outlined in the studies can be applied using networked technological systems to gather data in real time. For instance, teachers could easily record measures of curricular monitoring and adherence on an on-going basis. Rather than impose lockstep pacing guides, based on external and untested models of sequencing and timing (and instead of focusing on punitive responses if a teacher or class falls off the pace), districts could require teachers to report and interpret how they implement a curriculum, and learn from it. Records of when and why teachers supplement curricular materials, become delayed, or experience difficulty with one or more topics would generate more informative district-wide data about curricular use, and become a means to use ongoing practice to inform future implementation, especially from combining monitoring and supplementation data with disaggregated student and school data. In the near future, along with electronically delivered curriculum, the bulk of such monitoring could even be done automatically.

The studies asked teachers to complete a number of surveys regarding their knowledge of standards, their beliefs about instructions, and their approaches to certain kinds of practices, as well as core information about teacher capacity and about their participation in professional development. Data from such surveys, gathered periodically within technologically networked practitioner communities, could be factored into models of curricular implementation, professional development planning, and overall teacher community organization, with the goal of instructional improvement at the individual teacher or classroom level, and at higher levels of organization such as departments, schools, and districts.

Perhaps the most difficult data gathering tasks will be the collection of the kind of real-time observational data required for analysis of many of the implementation factors. While surveys and teachers' own monitoring reports can shed light on these issues, the collection of observational data, and its analysis via established, reliable rubrics, will continue to be an essential, and costly, element. It will be challenging to gather and use observational data to help define curricular, or, more broadly, *instructional* effectiveness (even with some of new technologies for classroom video recording becoming available). The use of video from such observations to guide professional development may turn out to be a major driver in our efforts to engineer for effectiveness going forward.

In this chapter, we concentrated on measures to permit comparison of curricular implementation and effectiveness, and emphasized the importance of ensuring curricular sensitivity and the alignment of outcome measures to systemic factors. But one can imagine that technological means of data gathering can enhance or transform the kinds of outcomes recorded, measured, and reported.

Treating the instructional core as a complex system will support efficient design and implementation of such new innovations in curricular implementation and

prototype systems for gathering and analyzing relevant data. As these are created, with the aim of engineering the instructional core for improved effectiveness, it will be essential to consider the use scenarios of innovations—to ensure that the data gathering fits into the work flow of engaging classroom activities (i.e., does not become onerous for teachers' workloads, *and* in fact reinforces their instructional efforts), that the data neither artificially reduce nor diminish the complexity of the instructional core, and that the statistical analytic approaches are robust and appropriate.

The ongoing improvement of the complex instructional core requires a "capacity to inform improvement" (Bryk, 2009) that establishes regular flow of information, feedback, and consultation within and among different levels of the educational organization. This argues for the establishment, in schools and districts, of networked improvement communities that include practitioners, researchers, technologists, and leaders who all participate throughout the work of achieving common goals, the design, testing, and implementation of the innovations, recognizing patterns and identifying sources of variability (Bryk et al., 2011).

All major complex systems (websites, health systems, communications, consumer marketing, climate analysis, disaster relief) are moving to the use of data-intensive systems with related analytics. What is most compelling in the studies described here is that it is possible to infer from them how we should be developing and deploying technologically enabled systems of data collection that will permit us to (a) gather more complete types and quantities of data about what is happening in classrooms, (b) become aware when a system exhibits patterns or trends toward improvement, stagnation, or deterioration over time, and (c) learn how to drive those systems towards improvement. Learning to undertake this level of analysis would constitute second-order traits of these complex systems.

Several principles continually surface in considering the goal of improving the instructional core: (a) curriculum matters; (b) instructional materials matter, because these best express the enacted curriculum, and their importance grows as the scale of implementation increases to the district level; (c) coherence matters, because it is critical in any complex system that all the moving parts align and mutually support each other; (d) multiple processes combine to result in observed outcomes (Bryk et al., 2011); (e) focusing solely on outcome data is not sufficient to support instructional improvement; and (f) managing and monitoring the implementation of tools/programs/curricula is a key function of school and district leadership.

This review leads us to the conclusion that it should be a high priority to design and implement technologically enabled systems that extend the capability of district and state data systems to gather data that can inform *improvement of the instructional core*, focused on curricular selection, use and implementation.[7]

[7] The components outlined here would not be a complete set to drive improvement in the instructional core. In an earlier version of this paper, we sought to discuss formative assessment and tie it to the construct of learning trajectories, diagnostic assessments and instructional practices, but it was too ambitious for a single paper. This second analysis will lead to an additional set of factors and data elements to this system, and we hope to complete that paper as a companion to this one in the near future.

Based on this review, we believe that districts could make significant progress on such an agenda in the areas of outcome measures, curricular monitoring, curricular implementation factors, and professional development and capacity issues. To this end, we outline a set of proposed actions.

Steps in a Strategic Plan to Strengthen the Instructional Core in Relation to Curricular Use, Implementation, and Outcomes

1. Form "networked improvement communities," (Bryk et al., 2011) to define tractable problems on which to focus, establish common targets and develop precise, measurable goals for the instructional core, across multiple levels of the system (teachers and classrooms, researchers, schools, districts).
2. Construct databases of assessment items linked directly to Common Core State Standards (using a set of relevant tags that distinguish among the features and measures), a variety of outcome measures to yield fair tests, and tests aligned to the CCSS. Focus on creating automated means of scoring that support the use of varieties of item types (multiple choice, as well as constructed and extended response) and concentrate on how to get meaningful data to teachers and students.
3. Develop and implement a means of analyzing, documenting, and notating the alignment of a curriculum to the CCSS, and of creating a standardized means of analyzing and representing content analysis of a curricular program.
4. Build a data system to gather and monitor data on curricular use, supplementation, and reasons for supplementation, gathered in real time.
5. Collect data on implementation factors such as those identified in the above studies.
6. Link the data system and various data categories and outcome measures to student, classroom, school, and district demographic data.
7. Link the data system to teacher demographic and survey data.
8. Find/develop/implement ways to conduct valid classroom observations (by teachers, supervisors, principals, specialists) for professional development purposes, and to triangulate these observations with teacher self-reports.

Finally, we argued that the value of the work rested in building models of the complex system known as the instructional core, and in engineering that instructional core for effectiveness by designing and implementing data systems using the constructs and measures developed by the studies. We suggest that treating the instructional core as a complex system, and taking a stance of engineering the instructional core for greater effectiveness of mathematical teaching, learning, and reasoning—studying what is happening in the classrooms in terms of patterns, trends, emergent behaviors, with deliberate sensitivity to variations in contexts—is a means to accelerate improvement in instruction and student learning. Ironically, by doing so, one could create a next generation of "best practices," this time with a focus on a continuously improving community in which research and practice draw more directly and iteratively from each other.

References

Agodini, R., Harris, B., Atkins-Burnett, S., Heaviside, S., Novak, T., Murphy, R., et al. (2009). *Achievement effects of four early elementary school math curricula: Findings from first graders in 39 Schools*. Washington, DC: IES National Center for Education Evaluation and Regional Assistance.

Agodini, R., Harris, B., Thomas, M., Murphy, R., Gallagher, L., & Pendleton, A. (2010). *Achievement effects of four early elementary school math curricula*. Washington, DC: IES National Center for Education Evaluation and Regional Assistance.

Bass, L., Charles, R. I., Jonson, A., & Kennedy, D. (2004). *Geometry*. Upper Saddle River, NJ: Prentice Hall.

Bellman, A. E., Bragg, S. C., Charles, R. I., Handlin, W. G., & Kennedy, D. (2004). *Algebra 1*. Upper Saddle River, NJ: Prentice Hall.

Berwick, D. M. (2008). The science of improvement. *The Journal of the American Medical Association, 299*(10), 1182–1184.

Boyd, C. J., Cummins, J., Malloy, C., Carter, J., & Flores, A. (2005). *Geometry*. Columbus, OH: Glencoe-McGraw Hill.

Bryk, A. S. (2009). Support a science of performance improvement. *Phi Delta Kappan, 90*(8), 597–600.

Bryk, A. S., & Gomez, L. M. (2008). Ruminations on reinventing an R&D capacity for educational improvement. In F. M. Hess (Ed.), *The future of educational entrepreneurship: Possibilities of school reform* (pp. 127–162). Cambridge, MA: Harvard University Press.

Bryk, A. S., Gomez, L. M., & Grunow, A. (2011). Getting ideas into action: Building networked improvement communities in education. In M. Hallinan (Ed.), *Frontiers in sociology of education*. New York: Springer.

Burger, E. B., Chard, D. J., Hall, E. J., Kennedy, P. A., Leinwand, S. J., Renfro, F. L., et al. (2007). *Geometry*. Austin, TX: Holt.

Charles, R., Crown, W., Fennell, F., Caldwell, J. H., Cavanagh, M., Chancellor, D., et al. (2005a). *Scott Foresman-Addison Wesley mathematics: Grade 1*. Glenview, IL: Pearson Scott Foresman.

Charles, R., Crown, W., Fennell, F., Caldwell, J. H., Cavanagh, M., Chancellor, D., et al. (2005b). *Scott Foresman-Addison Wesley mathematics: Grade 2*. Glenview, IL: Pearson Scott Foresman.

Chávez, O., Papick, I., Ross, D. J., & Grouws, D. A. (2010). *The essential role of curricular analyses in comparative studies of mathematics achievement: Developing "fair" tests.* Paper presented at the Annual Meeting of the American Educational Researcher Association, Denver, CO.

Coburn, C. E., & Stein, M. K. (Eds.). (2010). *Research and practice in education: Building alliances, bridging the divide*. Lanham, MD: Rowman & Littlefield.

Cohen, D. K., Raudenbush, S. W., & Ball, D. L. (2003). Resources, instruction, and research. *Educational Evaluation and Policy Analysis, 25*(2), 119–142.

Confrey, J., Castro-Filho, J., & Wilhelm, J. (2000). Implementation research as a means to link systemic reform and applied psychology in mathematics education. *Educational Psychologist, 35*(3), 179–191.

Confrey, J., & Makar, K. (2005). Critiquing and improving the use of data from high stakes tests with the aid of dynamic statistics software. In C. Dede, J. P. Honan, & L. C. Peteres (Eds.), *Scaling up success: Lessons learned from technology-based educational improvement* (pp. 198–226). San Francisco: Jossey-Bass.

Confrey, J., & Maloney, A. P. (2012). Next generation digital classroom assessment based on learning trajectories in mathematics. In C. Dede & J. Richards (Eds.), *Digital teaching platforms: Customizing classroom learning for each student* (pp. 134–152). New York: Teachers College Press.

Conklin, E. J. (2005). *Dialogue mapping: Building shared understanding of wicked problems*. New York: Wiley.

Coxford, A. F., Fey, J. T., Hirsch, C. R., Schoen, H. L., Burrill, G., Hart, E. W., et al. (2003a). *Contemporary mathematics in context: A unified approach (Course 1)*. Columbus, OH: Glencoe.

Coxford, A. F., Fey, J. T., Hirsch, C. R., Schoen, H. L., Burrill, G., Hart, E. W., et al. (2003b). *Contemporary mathematics in context: A unified approach (course 2)*. Chicago: Everyday Learning.

Deming, W. E. (2000). *Out of the crisis*. Cambridge, MA: MIT Press.

Elmore, R. F. (2002). *Bridging the gap between standards and achievement (report)*. Washington, DC: Albert Shanker Institute.

Fuson, K. C. (2009a). *Math expressions: Grade 1*. Orlando, FL: Houghton Mifflin Harcourt Publishing.

Fuson, K. C. (2009b). *Math expressions: Grade 2*. Orlando, FL: Houghton Mifflin Harcourt Publishing.

Gould, S. J. (1996). *Full house: The spread of excellence from Plato to Darwin*. New York: Three Rivers Press.

Grouws, D. H., Reys, R., Papick, I., Tarr, J., Chavez, O., Sears, R., et al. (2010). *COSMIC: Comparing options in secondary mathematics: Investigating curriculum*. Retrieved 2010, from http://cosmic.missouri.edu/

Grouws, D. H., Tarr, J. E., Chávez, Ó., Sears, R., Soria, V. M., & Taylan, R. D. (2013). Curriculum and implementation effects on high school students' mathematics learning from curricula representing subject-specific and integrated content organizations. *Journal for Research in Mathematics Education, 44*(2), 416–463.

Henningsen, M., & Stein, M. K. (1997). Mathematical tasks and student cognition: Classroom-based factors that support and inhibit high-level mathematical thinking and reasoning. *American Education Research Journal, 28*(5), 524–549.

Hiebert, J., Gallimore, R., & Stigler, J. W. (2002). A knowledge base for the teaching profession: What would it look like and how can we get one? *Educational Researcher, 31*(5), 3–15.

Holliday, B., Cuevas, G. J., Moore-Harris, B., Carter, J. A., Marks, D., Casey, R. M., et al. (2005). *Algebra 1*. New York: Glencoe.

Juran, J. M. (1962). *Quality control handbook*. New York: McGraw-Hill.

Kennedy, P. A., McGowan, D., Schultz, J. E., Hollowell, K., & Jovell, I. (2001). *Algebra one interactions: Course 1*. Austin, TX: Holt.

Krupa, E. E., & Confrey, J. (2010). *Teacher change facilitated by instructional coaches: A customized approach to professional development*. Paper presented at the Annual Conference of North American Chapter of the International Group for the Psychology of Mathematics Education.

Krupa, E. E., & Confrey, J. (2012). Using instructional coaching to customize professional development in an integrated high school mathematics program. In J. Bay-Williams & W. R. Speer (Eds.), *Professional collaborations in mathematics teaching and learning 2012: Seeking success for all, the 74th NCTM Yearbook*. Reston, VA: National Council of Teachers of Mathematics.

Larson, N. (2008). *Saxon math*. Orlando, FL: Harcourt Achieve, Inc.

Larson, R., Boswell, L., Kanold, T. D., & Stiff, L. (2001). *Algebra 1*. Evanston, IL: McDougal Littell.

Larson, R., Boswell, L., & Stiff, L. (2001). *Geometry*. Evanston, IL: McDougal Littell.

Lemke, J. L. (2000). *Multiple timescales and semiotics in complex ecosocial systems.* Paper presented at the 3rd International Conference on Complex Systems, Nashua, NH.

Maroulis, S., Guimera, R., Petry, H., Stringer, M. J., Gomez, L. M., Amaral, L. A. N., et al. (2010). Complex systems view of educational policy research. *Science, 330*, 38–39.

McNaught, M., Tarr, J. E., & Sears, R. (2010). *Conceptualizing and measuring fidelity of implementation of secondary mathematics textbooks: Results of a three-year study.* Paper presented at the Annual Meeting of the American Educational Research Association, Denver, CO.

NRC. (2004). *On evaluating curricular effectiveness: Judging the quality of K-12 mathematics evaluations*. Washington, DC: The National Academies Press.

Penuel, W. R., Confrey, J., Maloney, A. P., & Rupp, A. A. (2014). Design decisions in developing assessments of learning trajectories: A case study. *Journal of the Learning Sciences, 23*(1), 47–95.

Reys, R. E., Reys, B. J., Lapan, R., Holliday, G., & Wasman, D. (2003). Assessing the impact of standards-based mathematics curriculum materials on student achievement. *Journal for Research in Mathematics Education, 34*(1), 74–95.

Rittell, H. W. J., & Webber, M. M. (1984). Planning problems are wicked problems. In N. Cross (Ed.), *Developments in design methodology* (pp. 135–144). New York: Wiley.

Schmidt, W. H., McKnight, C. C., Houang, R. T., Wang, H., Wiley, D., Cogan, L. S., et al. (2001). *Why schools matter: A cross-national comparison of curriculum and learning*. San Francisco: Jossey-Bass.

Shavelson, R. J., & Towne, L. (Eds.). (2002). *Scientific research in education*. Washington, DC: National Academy Press.

Stein, M. K., Grover, B. W., & Henningsen, M. (1996). Building student capacity for mathematical thinking and reasoning: An analysis of mathematical tasks used in reform classrooms. *American Education Research Journal, 33*(2), 455–488.

Stein, M. K., & Kaufman, J. H. (2010). Selecting and supporting the use of mathematics curricula at scale. *American Education Research Journal, 47*(3), 663–693.

Tarr, J. E., Grouws, D. H., Chávez, Ó., & Soria, V. M. (2013). The effects of content organization and curriculum implementation on students' mathematics learning in second-year high school courses. *Journal for Research in Mathematics Education, 44*(4), 683–729.

Tarr, J. E., Ross, D. J., McNaught, M. D., Chávez, O., Grouws, D. A., Reys, R. E. et al. (2010). *Identification of student- and teacher-level variables in modeling variation of mathematics achievement data*. Paper presented at the Annual Meeting of the American Educational Research Association, Denver, CO.

Tatar, D. (2007). The design tensions framework. *Human-Computer Interaction, 22*(4), 413–451.

Thomas, S. M. (2010). *A study of the impact of professional development on integrated mathematics on teachers' knowledge and instructional practices in high poverty schools.* Unpublished doctoral dissertation, North Carolina State University, Raleigh, NC.

Wittenberg, L., Economopoulos, K., Bastable, V., Bloomfield, K. H., Cochran, K., Earnest, D., et al. (2008). *Investigations in number, data, and space* (2nd ed.). Glenview, IL: Pearson Scott Foresman.