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

Jere Confrey
Jere_confrey@ncsu.edu

Alan Maloney
Alan_maloney@ncsu.edu

The Friday Institute for Educational Innovation
College of Education
North Carolina State University
Raleigh, NC

A paper prepared for the National Academies
Board on Science Education and Board on Testing and Assessment for

"Highly Successful STEM Schools or
Programs for K-12 STEM Education: A Workshop"

Preface

The charge raised for this meeting was to identify "what are the best practices[1] that define effective STEM schools?" The juxtaposition of "best practice" and "effective" in this charge poses a particular challenge—best practices refer to the informal wisdom of the field delineating what "gets the job done." But effectiveness demands an evidentiary base complete with proof, justification and warrant. Thereby, as researchers, we claim that our charge can only be met definitively for "best practices" that have been researched. For this reason, our contribution to the meeting is to review, discuss, critique and refine what we can learn from a promising set of studies of curricular effectiveness, and propose a means to refocus our approach to effectiveness. We hope this will constitute a reasonable response to our charge.

The Process of "Engineering [for] Effectiveness"

Improving schools has been often cast as a challenge of identifying effective programs, as captured by the call for "What Works?" Many researchers, skeptical of this call, argue that the real question should be "what works, for whom, and under what conditions" (Means & Penuel, 2005; Bryk et al., 2011). A shift to focus on specific results that accrue under precise conditions and resources rests on the assumption that educational results require adaptations to circumstances, and therefore to seek broad scientific principles or rules that apply across the board is of limited value. For example, Bryk, Gomez, and Grunow noted, "Treatises on

[1] "Best practices are generally accepted, informally-standardized techniques, methods, or processes that have proven themselves over time to accomplish given tasks. Often based on common sense, these practices are commonly used where no specific formal methodology is in place or the existing methodology does not sufficiently address the issue." (Wikipedia)

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." [italics added] (Bryk et al., 2011)

Shifting the question to "what works, for whom, and under what conditions?" has profound implications for the meaning of effectiveness. In establishing causal models, one determines, within the restrictions of a particular study's conditions, if an effect, controlling for other factors, can be rigorously linked to a cause, and focuses on the internal validity of the study--hence "cause" and "effect." While studies typically can and do produce small, but "significant" effects, they often have nested within them more interesting conjectures about interactions (often correlational) and relationships (if they survive the editorial chopping block's insistence on narrow views of rigor). Ironically, those who demand causal design are often silent on the necessity of replication, which, strictly speaking, is required in order to realize the benefits of randomization; one study alone does not ensure generalizability.[2] 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.

Moreover, too many policy-makers and practitioners assume that an established treatment, as cause, can be directly applied to a practice and guarantee

[2] 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.

an effect. Consequently, most studies leave the practitioners responsible to evaluate whether that study generalizes to their own settings. How they are supposed to do this responsibly is seldom addressed. Bureaucrats will put a stamp of approval on the product, based on a study's internal validity, but fail to address the implications of the constraints on a study's external validity.

Due to the cost, time, and difficulty of conducting and analyzing randomized field trials, reliance on them as the only source of "effectiveness" leaves the public continually awaiting a sufficient set of scientifically proven empirical results.

In contrast, in this paper we argue that by developing and deploying explicit means of "engineering [for] effectiveness," communities of practitioners and researchers can conduct ongoing local experiments in context, that include adequate design, technologically-enabled tools for real-time data collection and continuous analysis of patterns and trends. As new findings emerge, they can be shared across common communities of practice.

Approaches similar to engineering [for] effectiveness have emerged under a variety of names: continuous improvement models (Juran, 1962; Deming, 2000), implementation research (Confrey et al., 2000; Confrey & Makar, 2005), improvement research (Bryk et al., 2011), a science of improvement (Berwick, 2008), Design-Educational Engineering and Development (DEED) (Bryk & Gomez, 2008; Bryk, 2009), and the study of complex and dynamic systems (Maroulis et al., 2010). When examined through the lenses of these various models, it becomes evident that the improvement of educational outcomes requires reexamination of approaches to "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 are 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 non-randomly-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.

3. *Causal or covarying cycles with feedback and interaction are critical elements of educational systems, in which learning is a fundamental process.* Furthermore, one expects emergent phenomena (Maroulis et al., 2010). There is a contrast between construction 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 "did the curriculum cause the effects, or 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; Tatar, 2007; Penuel et al., submitted), and develop and carry out protocols for inquiry. In such a "networked improvement community" (Bryk et al.), one

positions 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.

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, a 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 on redefining the approach to effectiveness in the context of curricular study. In this approach, there are complementarities with Bryk et al.'s (2011)call for a change in "protocols for inquiry" as they discuss how to carry out a "science of improvement," locating it between the 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, 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 et al., 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).

Likewise by defining a means of "engineering [for] effectiveness" we describe how communities of practice, at district or state level, can build on what has been learned from studies of curricular effectiveness. To do so, we review studies associated with effectiveness research from mathematics education and reinterpret their results and implications. Our focus will be on the challenge of improving the instructional core (Elmore, 2002; Cohen et al., 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 and the High Stakes tests. Together the two latter components of the educational system are the bookends that constitute the accountability system. Policy levers were 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, accountability measures that were driven by *No Child Left Behind* neglected and/or avoided the instructional core in relation to professional

development, pedagogy, and classroom assessment, and the absence of common standards fragmented the attention to curriculum (Reys et al., 2003). By squeezing the educational system by way of the bookends, the accountability system during the past 10 years produced some performance gains from the system. However, it failed to strengthen the instructional core with respect to capacity, led to a narrowing of enacted curriculum, and, while it called for the use of "best practices" it failed to identify a means to establish their credibility.

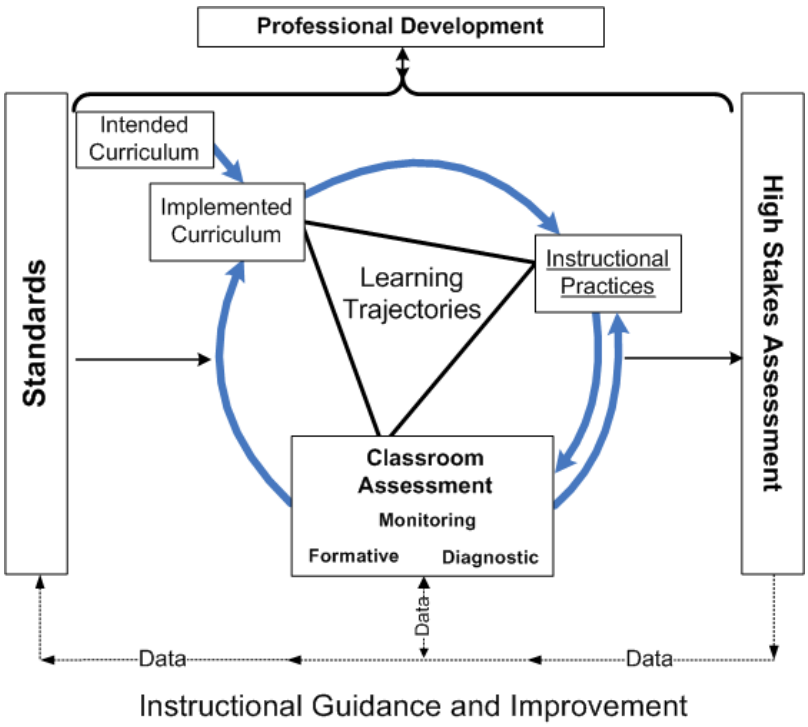

Figure 1. Model of Classroom Educational System, illustrating position of the instructional core between the accountability "bookends" (Confrey & Maloney, in press).

The instructional core was intentionally chosen as a focus for this paper 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 system, from the standards and the summative tests to classroom practices and formative feedback. While temporally one can view a sequence of curricular selection,

involving some degree of professional development, followed by implementation and assessments (both formative and summative), each of these components, also, interacts with and acts as feedback to the other components. For instance, frequent formative results provide regular feedback to classroom practices, while data from high-stakes tests provide intermittent feedback and a much cruder level of non-specific pressure. Resulting practices can be customized to groups according to documented needs. Networked improvement communities are not explicitly identified in the figure, but could be configured so that communities of practice can 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.

The adoption of the CCSS by numerous states positions us to create policy approaches and to reconsider the value of how to focus on improving the instructional core without overly constraining innovation, over-regulating curricular choice, or deskilling teaching. At the same time, we must adopt larger goals of becoming smarter about curricular effects, improving instruction, engaging students as individuals and as members of communities of scholars, increasing teaching capacity, and reducing.

By examining 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, one can jump-start a movement towards school improvement in STEM disciplines.

Curricular Effectiveness Studies.

As reported by Schmidt, "curriculum matters." 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,* the community of mathematics educators has worked diligently to strengthen and improve research on and evaluation of curricular effects. The report's framework called for evaluations designed to examine three components: 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 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 finer outcome measures, argued for increased independence of evaluators, precise identification of comparison programs, and better measures of implementation. We have selected three cases that have taken these recommendations seriously and moved research to the next level. We report on their approaches, their findings, and their limitations, and discuss how they can be interpreted so as to provide a solid foundation to next generation efforts to "engineer [for] effectiveness."

Case One: Single Subject vs. Integrated Math.

New studies of curricular effectiveness have advanced our understanding of curricular effectiveness. One such study is called "Comparing Options in Secondary Mathematics: Investigating Curriculum," (COSMIC;) (Grouws et al., 2010), in which Grouws and Tarr and colleagues compared the effects of two curricula, one subject-specific and one integrated, on student learning in high school mathematics. The study involved 11 schools in six districts across five regions of the country, in which these two curricula were used in parallel without tracking by ability level. The schools displayed a range of demographics; the proportion of students eligible for free and reduced lunch ranged from 19- 53%. The study design was quasi-experimental using prior achievement, normed against NAEP to create comparability across states, with data analyzed using hierarchical linear modeling (HLM). A goal of the study was the improved understanding regarding relationships among curricular organization, curricular implementation factors, and gains in student learning.

The study provided a number of significant advances in research on curricular effectiveness. First, researchers utilized expertise in mathematics content and learning effectively in the design and selection of their outcome measures. They used multiple outcome measures: two tests designed specifically for the project (one of content and one of reasoning and problem solving), and a standardized test, Test of Educational Development [ITED]: Mathematics: Concepts and Problem Solving. Drawing from the NRC report, the project team designed a "fair test" (which had been called a conservative test in UCSMP (NRC, 2004), defined as "developed with the deliberate goal of not being biased towards either of the two

curriculum programs studied." To build the fair test, the project began with content analyses of the two texts (Chavez et al., 2010, p. 4). In addition, their outcome measures were richer than many multiple choice-only tests, because they also relied on constructed responses scored using a rubric construction method that included careful internal and external review.

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 including Table of Contents Records, Textbook-Use Diaries, Initial Teacher Survey, and observations using a Classroom Visit Protocol (Mcnaught et al., 2010, p. 5)(p. 10). They were able to examine critical factors such as professional development, familiarity with standards, distribution of classroom time among lesson development, non-instruction, practice, and closure. In a sub-study, they defined, studied, and compared three related indices of curricular implementation: opportunity to learn (OTL), representing the percentage of textbook lessons taught; Extent of Textbook Implementation (ETI), representing the extent to which teachers followed their textbook using weighted averages; and Textbook Content Taught (TCT), representing the extent to which teachers, *when teaching textbook content*, followed their textbook, supplemented their textbook lessons, or used altogether alternative curricular materials. For this study, they reported that for OTL only 60.81% (19.98 SD) of the content of an integrated textbook was taught while 76.63% of the subject-specific textbook was taught. Based on the ETI, they showed that, across teachers, slightly more than one-third (35%) of the content was taught

primarily from the textbook, approximately one-fifth (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. The TCT showed that integrated content was taught more frequently directly from textbook (59%) as compared to subject-specific content (46%). One can see that interpreting findings on student learning outcomes related to a curricular treatment without considering information on textbook use could easily lead to unfounded conclusions.

In addition, because of the extent and richness of the data gathered at the teacher level with this curriculum evaluation model, the COSMIC researchers used the statistical technique of principle component analysis to simplify their model, and identified seven key components explaining 71.4 % of the variance. Four of these clustered around curricular implementation: standards-based instruction (extent of sense-making, student reasoning, presentation fidelity and closure), implementation fidelity (ETI, TCT, textbook satisfaction), technology and collaborative learning, and opportunity to learn. The other three factors were related to teacher characteristics: standards knowledge, experience, and professional development. These indicate the necessity of gathering data on these factors as potential mediators of curricular effects.

Their results report on students' mean residualized gain scores for each teacher, in recognition that the unit of analysis should not be the individual student(NRC, 2004). Those gain scores (adjusted for prior achievement) were positively correlated in favor of teachers of the integrated curricula, on the reasoning test

only, with no significant differences reported on the content and standardized tests. However, an examination of partial correlations found that when controlling for %FRL, the magnitude of the correlation between Curriculum Type and student outcomes became significantly different than 0 in favor of the integrated curricula; this was the case for all three tests. Moreover, the importance of OTL is substantially reduced with the partialing out of %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. Since teachers of integrated curricula covered significantly less textbook content than teachers of subject-specific curricula, a difference in coverage 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.

Overall, the COSMIC study illustrates that it is unwise to expect curricular studies to yield simple answers about curricular effectiveness. Technically, the study generalizes only to schools that offer both curricular options, if student choice rather than tracking determines which students enroll in the two curricula. Practitioners, however, ask whether an integrated program generates better, worse or the same outcomes as a single-subject approach, and unless the school offers both, the study offers no secure answer.

The COSMIC study however yields far more insight than its "curricular effects." These insights can be linked directly with the components 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 appropriate 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 is not interested in how two curricula perform on tests of common content but want to know how they affect performance on a measure that assesses common standards, as will be more likely with the implementation of the Common Core standards, the study describes how to recognize and pick 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 need closer inspection. The study demonstrates that the higher the percentage of students eligible for FRL, the lower the opportunity to learn. Further, the study suggests that the effects of the curriculum in favor of integrated math become more evident when FRL is partialed out.

Arguably, these findings suggest that using integrated math with students of poverty is a promising option, but would require teachers to receive substantial

assistance to increase students' "opportunity to learn." Furthermore, the COSMIC study informs readers about the make-up of curricular implementation comprised of standards-based instruction (extent of sense-making, student reasoning, presentation fidelity and closure), implementation fidelity (ETI, TCT, textbook satisfaction), technology and collaborative learning, and opportunity to learn. These results suggest that in addition to focusing on opportunity to learn, school leaders need to help teachers to understand the standards, focus on student reasoning and sense-making, and learn to reach closure. In fact, in a study in North Carolina, we found that, based on an analysis of reports from content specialists' monthly observations of teachers' practice, teachers using an integrated mathematics curriculum with students in poverty 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 studying multiple cases of teachers in these schools, Thomas (2010) showed that providing adequate support to teachers can transform practice but this is not easily accomplished due to weakness in teacher knowledge and to those teachers' views of instruction(Thomas, 2010). 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 Math-Science Partnership through the 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, and a semi-autonomous school organization committed to improving rural education. Our efforts could have

benefitted from richer and continuous data sources informed by research tools such as those developed for COSMIC.

Case Two: Comparing Effects of Four Curricula on 1st and 2nd grade Math Learning. A second major study on curricular effectiveness provides another example of the potential contributions of nuanced study beyond claims of cause and effect. The study "Achievement Effects of Four Early Elementary School Math Curricula: Findings for First and Second Graders," 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) (Agodini et al., 2009; Agodini et al., 2010). A total of 473 districts invited, but, only 12 agreed to participate in the study—a recruitment rate of 2.5 percent (Agodini et al., 2010, p. 10).[3] In all, 109 first grade elementary classes and 70 second grade classes were randomly assigned to a curriculum within districts. The authors (R. Agodini, B. Harris, M. Thomas, R. Murphy, L. Gallagher, and A. Pendleton) used four contrasting curricula including *Investigations in Number, Data, and Space* (*Investigations*; student-centered approach encouraging metacognitive reasoning and drawing on constructivist learning theory), *Math Expressions* (blending student-centered and teacher-directed approaches to mathematics), *Saxon Math* (Saxon) (scripted curriculum using direct instruction in procedures and strategies with guided and

[3] The authors acknowledge 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).

distributed practice) *and Scott Foresman-Addison Wesley Mathematics* (*SFAW*) (a basal curriculum that combines teacher-directed instruction with a variety of differentiated materials and instructional strategies).

The study addressed three broad questions: 1) What are the relative effects of the study's four math curricula on math achievement of first- and second-graders in disadvantaged schools? 2) Are the relative curriculum effects influenced by school and classroom characteristics, including teacher knowledge of math content and pedagogy? 3) What accounts for curriculum differentials that are statistically significant? Data were collected from fall and spring administrations of an adaptive test with items from the Early Childhood Longitudinal Study, student demographic and school data, teacher surveys, study-administered assessments of math content and pedagogical content, and an observation scale.

The study results were reported as pairwise comparisons between curricula (six possible pairwise comparisons). After one year of study participation, average spring 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 both *Investigations* and *SFAW* students. For a first grader at the 50th percentile in math achievement, these results mean that the student's percentile rank would be 4 points higher if the school 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 and 7 points higher if the school used *Math Expressions* or *Saxon Math*, respectively,[4]

This study, like COSMIC, examined curricular implementation, and reported on such factors as 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. The authors found that the teachers reported different coverage of math content areas across the curricula. They reported when differences in pairwise comparisons were significant, and determined, "There was no clear pattern to which curriculum differences are significant." (p.57).

For the table below, we selected some implementation differences that could have affected student-learning outcomes. For instance, teachers received twice as much initial professional development for *Expressions* than for other curricula, teachers of *Saxon Math* taught math an additional 20% each week, teachers of *Expressions* used more supplementation materials while *Investigations* teachers used less, and 16.2% of *Saxon Math* teachers and 21.1% of *SFAW* had taught with those curricula previously. Not surprisingly therefore, implementation reports show that higher percentages of first- and second- grade *Investigations* and *Expressions* teachers (22.3/23.2, 33.7/56.1) report feeling only "somewhat" or "not

[4] 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 1st 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. Compared 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 7-14% of the content as measured by the ECLS assessment.

at all" prepared to teach their curriculum, compared to teachers of *Saxon Math* (16.0/16.4) or *SFAW* 10.0/9.1).

	Respondents	*Investigations*	*Expressions*	*Saxon Math*	*SFAW*
Total amount of PD	All teachers	1 day	2 days	1 day	1day
Responded "Somewhat or not at all" adequately prepared after training"	1st grade teachers	23.3	33.7	16.0	10.0
	2nd grade teachers	23.2	56.1	16.4	9.1
Additional training	Reported by publishers	3-4 hours every 4-6 weeks	Twice a year (individually)	Once a year (individually)	3-4 hours every 4-6 weeks
Supplemented curriculum	1st grade teachers	14.8	32.1	24.8	27.5
	2nd grade teachers	11.7	55.6	30.5	24.6
Hours taught per week	1st grade teachers	5.1	5.0	6.1	5.3
	2nd grade teachers	5.4	5.5	6.9	5.5
Used the assigned curriculum the previous year	1st grade teachers	5.5	3.6	16.2	21.1

Table 1: Selected differences in implementation variables.

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 the designers ("program theory,") was distinguished from its application during implementation ("implementation fidelity"). The study measured adherence via a teacher survey and a classroom observation instrument. The data presented below suggests that teachers were more likely to adhere to designers' intentions in the *Saxon Math* program than in the *Expressions* program.

		Investigations	*Expressions*	*Saxon Math*	*SFAW*
Survey report	1st grade teachers	66	60	76	70
	2nd grade teachers	67	54	76	68
Observation	1st grade teachers	56	48	63	54
	2nd grade teachers	53	47	65	53
	Average	60.5	52.25	70	60.75

Table 2: Percentage of adherence to a curricular program's essential features (p. 65)

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 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 study by the authors indicated that some of these implementation factors act as mediators of achievement outcomes. Because of the design of the study, however, researchers could only examine the effects of one mediator at a time. The implications of this restriction means that while differences in professional development for *Expressions* mediated the curricular effect, the authors could not relate this to the meditational effects of less prior experience with and teachers' reports of less preparedness to teach the curriculum. Likewise, *Saxon Math* teachers are reported to have had 20% more instructional time, which mediated the *Saxon*

Math-SFAW difference in curricular effect. One cannot assess the combined effects of more instructional time and a higher likelihood of having taught a curriculum before. The authors interjected that a more rigorously designed study of mediation could disentangle these relationships among mediators (p. 102). In any case, these reported 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 study's many accomplishments of the Agodini et al. (2010) study was to identify and find a 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, it clearly 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, analogous to the distinction between a general outcome measure and a fair test.

The Agodini study also exhibits limitations and threats to its validity. It relied on a single outcome measure, and did not report on a method to check the "fairness" of that outcome measure across the curricula. This is 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, seven to fourteen gain points on the scaled score, could be the result of a few key assessment items.

While the study benefits from randomized assignment of curricula within the district, this 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 implementation, and providing only one to two days of professional development for primary teachers, weakens the confidence in the results. For instance, the report of high levels of supplementation with *Expressions* could be due to teachers' use of prior, more familiar materials. If this were the case, should one draw the conclusion that *Expressions* was "effective" under these conditions?

Furthermore, the authors also described teachers' reports, for each curriculum, on the frequency of teaching particular content topics (whole numbers, place value etc.). If an analysis of the test had been done, one might have been able to discern patterns in the relation between students' opportunity to learn the material and the outcome measure scores.

The Agodini et al. study offers far more insight into curricular effectiveness than is captured by its conclusions on "cause and effect". Like 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 measuring OTL as teachers implemented, and the other by relying on teacher reports of number of lessons by content area. 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 it relates to student learning (Stein & Kaufman, 2010). The study involved two districts using reform curricula, one using *Everyday Math* (*EM*) and the other using *Investigations.*

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 et al., 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, due to the very open-ended discourse they support, which is difficult to manage (Henningsen & Stein, 1997). In contrast, PWC tasks are more bounded and predictable, but are susceptible to "losing the connection to meaning"(Stein & Kaufman, 2010). Another consideration is to consider the level of support for teacher learning embedded in the two curricula. Stein and Kaufman documented that there is less professional development support in the *EM* materials than in the *Investigations* materials.

From these two analyses, the study authors characterized *EM* as low-demand, low-support, and *Investigations* as a high-demand, high-support curriculum. They 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 the implementation of the two reform curricula in two districts that contained comparable numbers of students eligible for free and reduced lunch. They studied implementation of the curricula by six teachers (one per grade level) in each of four schools over a period of two years. They coded their observations (with examples), of three consecutive lessons in each of fall and spring, 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) vested the "intellectual authority in mathematical reasoning," rather than in the text or the

teacher. Together, high values on these three dimensions defined high quality implementation.

Using surveys, observations, and interviews, they examined two teacher characteristics: teacher *capacity* (defined as comprising years of experience, mathematical knowledge for teaching (MKT), participation in professional development and educational levels), and teacher's *use of curriculum* (teachers' view of the curriculum's usefulness, percentage-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*)?" (p. 667), they found that the teachers from the district using *Investigations* were more likely to teach high quality lessons than teachers from the district using *Everyday Math* (note that the relationship of instructional performances to student outcome performance was not investigated). 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 to 3), and to establish norms for the authority of mathematical reasoning (1.2> .4 on a scale of 0 to 2) .

In examining their second question across the two districts and curricula, "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?", they found that *none of the teacher capacity* variables was consistently and significantly related to the quality of implementation. In the district using *EM*, higher performance on MKT surveys was *negatively* correlated to the use of student thinking and to establishing the authority of mathematical reasoning in the classroom. In the district using *Investigations,* the correlations with teacher capacity were positive but not significant. In examining the relationship between either hours or type of professional development to implementation quality, they found no relationship in the district using *EM*, but in the district using *Investigations*, the amount of professional development was (positively) significantly correlated with all three components of implementation quality.

Across both districts, the discussion of big ideas in lesson planning was the \ *only teacher's use of curriculum variable* that was significantly and positively correlated to two out of three of the implementation quality components (attention to student thinking and the authority of mathematical reasoning). Further the authors reported that this tendency was more in evidence 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.

The study shows that implementation quality cannot be inferred from content topic analysis alone but depends also on how the tasks are structured. In addition, it appears to relate more to the extent of professional development support,

facilitated by the district and afforded by the materials, than to teachers' education, experience and the mathematical knowledge of teaching. Finally, the study revealed that the extent to which teachers use the materials to look for "big ideas" correlated with implementation quality across both curricula.

The authors suggest that their study points to a way to re-conceptualize curricular effectiveness as a process of improvement rather than as a product's warranted claim: "We 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). They further suggest that curricula could be viewed not only as programs to be implemented, but as tools to change practice.

Overall Conclusions from the Three Cases

Juxtaposing the three cases reviewed here has provided an opportunity to synthesize advice for future conduct of effectiveness studies. There is a temptation in the calls for, and interpretation of, effectiveness studies to try to identify something that works--that is, to identify one or more curricula that can be implemented with the expectation of subsequent, direct major improvements in student outcomes.

Initially, we 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. The review of these cases demonstrated how tentative causal conclusions are and reminded us that all

studies have flaws and limitations. The quest for the perfect curricular effectiveness study is highly unlikely to yield results that are robust or extensive enough to guide practice. Each study, at best, provides insight into some specific conditions under which certain outcomes occurred, depending on how constructs were defined and measured surrounding the implementation of the curricula.

The COSMIC study, for instance, provides evidence of relative effectiveness of an integrated curriculum compared to subject-specific curriculum when students are provided both options. However, were only one option to be provided, or if tracking had been implemented with the two curricular options, we do not know what the results would be. Or, based on the COSMIC results, it could be that if teachers of integrated curricula were able to cover the same percentage of their text during a year as teachers of a subject-specific curriculum, performance of students in integrated math would be relatively even stronger than of those in the subject-specific curriculum. A practitioner choosing to apply this study to make a curriculum selection decision has to weigh these considerations as he or she contextualizes the results to apply and adapt to his or her setting.

Similarly, the Agodini 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, in the case of *Saxon Math,* due to increased instructional time. Alternatively all outcomes of this study could be attributable to

some curricula being a better fit with the ECLS outcome measure; on another end-of-year assessment the results could have been quite different. Another possible interpretation of this study is that by studying the effectiveness of curricula in their first year of implementation, researchers skewed their results in favor of *Saxon Math* and *SFAW* which had higher levels of prior use, and that therefore the results would differ if monitored over a longer period of implementation or in another setting, and shift the ordering of the effectiveness outcomes.

All studies are open to multiple interpretations and unknown limits to their generalizability to new settings. In the Stein and Kaufman study, the stronger implementation quality of *Investigations* could have been attributed to its design of the curricular tasks, affordances for focus on big ideas, and support for professional development. But perhaps the district that offered *Investigations* simply supported its implementation with better quality and more extensive professional development.

Do the conflicting interpretations of these studies mean that they are pointless and a waste of time and money? Does the simple fact that we cannot know whether the results of a study will accrue in a setting that differs from the original and may require a level of adaptation from the conditions for the study mean that such studies are of limited importance?

If the goal of curricular effectiveness studies is to decide unequivocally whether a single product, a curricular program and related materials, can be placed into classrooms across the country and produce predictable gains in learning, then these studies fail to establish curricular effectiveness.

Rather, we argue that these studies, especially taken together, demonstrate why simple causality is an insufficient model for judging effectiveness of a curriculum. The message to be taken from them, then, is that many things matter to the implementation of a curriculum and to the learning that students can accomplish with different curricula. Context matters (the extent to which one is serving poorer students and needs more resources or stronger teachers). Resources matter. Teaching quality matters.

Most significantly, these studies contribute substantially to an understanding of the instructional core. By the very fact that these experts have gained purchase on modeling these systems, 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 getting smarter about *how curricular use in particular settings can improve instructional quality and student outcomes*.

In particular, these studies emphasize, beyond a doubt, the following lessons

1. Outcome measures matter—and with the availability of Common Core State Standards, we can create a variety of measures in a cost effective way across districts and states. These studies demonstrate that studies need multiple outcome measures which should: a) include measures that act as "fair" tests (Chavez et al) to ensure non-biased comparison of student performance on topics common to all curricula being examined, b) be normed against relevant populations (college-intending, ELS students) and used to make systemic decisions;(such as statewide end-of-course exams or new

assessments of Common Core State Standards) c) assess the development of big ideas over time, such as learning progressions, and d) assess other dimensions of mathematics learning such as the mathematical practices in CCSS, student attitudes, or intentions to pursue further study or certain STEM careers. The studies showed that the categories by which results were disaggregated were critical, and were sensitive to interactions, such as by race and FRL. At the very least, therefore, relevant data should be gathered in relation to performance measures, ethnic and racial diversity, gender, ELL, and FRL, to support investigation of relevant bands of variability in effects.

2. Monitoring what was actually taught, and <u>why it was taught,</u> is crucial to making appropriate attributions in examining effectiveness. Monitoring should include measures of curricular coverage (OTL and adherence), the degree and type of supplementation and reasons behind these choices. Different ways 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 implementation of curricula will add a wealth of insight to explanatory frameworks of effectiveness. Some factors should directly measure the extent to which implementation captures a designer's specific intent, while other should measure qualities that apply across all curricula. Many innovative methods of data collection were undertaken in the studies: surveys, intermittent and

extended classroom observations with various coding schemes, reports of instructional time usage, and interviews. In one case, these were coded in the predetermined categories of maintaining cognitive demand of tasks, eliciting student thinking, and vesting authority in mathematical reasoning. In the COSMIC and Agodini et al studies, statistical techniques (factor and principle component analyses) were used to identify and name implementation variables in clusters (standards-based instruction, implementation fidelity, peer collaboration, technology use, student-centered instruction, teacher-directed instruction, and classroom environment). These areas of research on identifying, defining and studying implementation factors, promise to continue to grow and add to our understanding of curricular effects.

4. Issues of professional development and teacher capacity are critical in judging curricular effectiveness, and their influence varies depending in part on whether they are viewed as a resource within a curriculum and its implementation or a factor that interacts with implementation. Teachers' capacity, as a pre-determined resource defined as teacher knowledge, experience, and education, did emerge as influential in two studies (COSMIC, Agodini et al.), but in the third study (Stein & Kaufman), in one district it did not correlate in a significant positive way with implementation quality one district, and in the other, some of its factors correlated negatively and significantly. In that study, it was instead the *amount* of professional development time, access to assistance and support, and the way in which

teachers used the materials in planning (a focus on big ideas) and communicated with others about curricular use that emerged as most closely associated with implementation quality. Yet again, in the COSMIC study, professional development was associated negatively with curricular impact. Thus, additional work is needed to clarify how professional development and capacity relate to curricular implementation. These studies do suggest important value in designs which incorporate and distinguish three perspectives on professional development and capacity: one in which it is viewed as a resource to implementation, one in which curricular implementation is seen as a tool for changing capacity and as a source of professional development, and one in which it could be viewed as a factor that interacts with implementation.

5. How a study is situated in relation to educational structures and organizations may eventually be important at a meta-level in understanding the curricular effects and the conclusions drawn. Each study's location was driven by experimental design issues—for instance, the offering of two curricular options without tracking (COSMIC), the dependence of a study on the willingness of districts to randomly assign teachers to treatments (Agodini et al.), and the choice by districts to support extended observations over two years and guarantee researchers access to extensive teacher data. While these factors are given as part of the study's design, over time such factors may emerge as major influences on understanding key organizational factors concerning curricular implementation, such as

governance, decision-making, funding, support and data use.

Engineering [for] Effectiveness: Summary and Recommendations

These studies remind us how remarkably complicated is the interplay of curricula, instruction, classroom assessment practices, and professional development. They demonstrate that the instructional core is in fact 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 would be wise therefore, to treat and study the entire instructional core, therefore as a complex system. We would suggest that rather than seek a grand causal effect from these studies, one should use them to learn more about possible ways to model complex systems.

A proposal that derives from that conclusion is to focus more on how to engineer [for] effectiveness, that is, to design our way into a greater understanding of the operation and improvement of the instructional core. These studies have provided some critical elements of such an endeavor, by identifying a number of critical constructs and creating measures to gauge and monitor them.

Many of the instruments outlined in these studies can be applied using networked technological systems to be gathered in real time. For instance, measures of curricular monitoring and adherence can be easily recorded by teachers on an on-going basis. Rather than impose pacing guides, based on external and untested models of sequencing and timing, let teachers report what

they do, and learn from it. Asking them to record when and why they supplement, become delayed, or experience difficulty with an area of study would be a way to use ongoing practice to inform future implementation. There is little doubt that in the near future, with curriculum delivered electronically much of this monitoring could even be done automatically.

Likewise, the studies ask teachers to complete a number of surveys to learn about their knowledge of standards, their beliefs about instructions, their approaches to certain kinds of practices, in addition to core information about teacher capacity and participation in professional development. By requesting these periodically within technologically networked practitioner communities, the data from these surveys can be factored into the models.

Perhaps the most difficult data gathering tasks will be those that involve the kind of real-time observational data required for the analysis of many of the implementation factors. While surveys and teacher reports can shed light on these issues, the collection of observational data, and analysis with established and reliable rubrics, will continue to be an essential, if costly, element. And while 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), the use of video from such observations to guide professional development may be a major driver in our efforts to engineer for effectiveness going forward.

Using technological means of data gathering can enhance the kinds of outcomes recorded, measured, and reported. In this paper, we concentrated on measures to

permit comparison of curricular effectiveness, and stressed the importance of ensuring curricular sensitivity and an alignment of outcome measures to systemic factors. In a subsequent paper, we will review the research on the effectiveness of formative assessment practices, and discuss in more detail approaches to designing and supporting formative assessment practices, learning progressions, and diagnostic assessment.

As we begin to build prototypes of systems to gather data so that we might engineer [for] effectiveness, it will be essential to consider the use scenarios—to ensure that the data gathering does not become too onerous for teachers, fits into the work flow of engaging classroom activities, and that the data neither reduce nor diminish the complexity of the instructional core. Treating the instructional core as a complex system will support efficient design and implementation of such new prototypes, for this goal calls for establishment of a networked improvement community that includes practitioners, researchers, and technologists, who all participate throughout the work of design, testing, and implementation of these innovations,

All major complex systems (websites, health systems, communications, consumer marketing, climate analysis) are moving to the use of data-intensive systems with related analytics. What is most compelling from the studies described here is that they all suggest how we should be engineering 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) seek to become aware when a system exhibits patterns or trends toward improvement, stagnation, or

deterioration over time, and c) seek to learn how to drive those systems towards improvement. Learning to undertake this level of analysis of complex systems would constitute second-order traits of these complex systems.

What can be learned from this review is that the priority should be to design and implement technologically-enabled systems that extend current district and state data systems to be able to gather data that can inform *improvement at the instructional core,* focused on curricular selection, use and implementation.[5] Based on this review, we could make immediate progress on such 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. Construct databases of assessment items linked directly to Common Core State Standards using a strong set of tags that distinguish among the features and measures, a variety of outcome measures to yield fair tests, and tests aligned to CCSS. Focus on creating automatized means of scoring that support the use of varieties of item types (multiple choice, constructed and extended response) and concentrate on how to get meaningful data to teachers and students.

[5] The components outlined here would not be a complete set to drive improvement in the instructional core. In our original version of this paper, we sought also to discuss formative assessment and tied 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.

2. 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.
3. Build a data system to gather and monitor data on curricular use, supplementation, and reasons for supplementation, gathered in real time;
4. Collect data on implementation factors such as those identified in the above studies.
5. Link the data system and various data categories and outcome measures to student, classroom, school, and district demographic data.
6. Link the data system to teacher demographic and survey data.
7. Find ways to conduct valid classroom observations (by teachers, supervisors, principals, specialists) for professional development purposes, and to triangulate these observations with teacher self-reports.
8. Form "networked improvement communities,"
9. Through the improvement communities, define tractable problems on which to focus, and identify and implement appropriate continuous improvement models.

By beginning with a review of studies of curricular effectiveness, we have illustrated what the studies have shown us about the operation of the instructional core, especially with respect to the implementation of curricula and its relationship to student outcomes. These studies shed light on the complexity of that system and

the degree to which interactions among the key variables constrain the validity of even simple conclusions on cause and effect. They also permit one to see why an approach to simple causality that lacks sufficient attention to causal cycles, interactions, and meditational effects, can limit the usefulness of the findings. Each study's results were bounded by the limits of generalizability of the results, with major implications for any practitioner who makes decisions based solely on those results. It was also demonstrated that the studies have a great deal to offer about the core areas of outcome measures, curricular monitoring and adherence, implementation factors, professional development and capacity, and the beginnings of an identification of organizational factors.

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 [for] effectiveness--studying what is happening in the classrooms in terms of patterns, trends, emergent behavior and deliberate sensitivity to variations in contexts--is a means to boost our speed of improvement. Ironically by doing so, one could create a next generation of "best practices", and this time in 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., & Pendleton, A. (2009). *Achievement Efects of Four Early Elementary School*

Math Curricula: Findings from First Graders in 39 Schools. Washington, DC: IES National Center fo 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 fo Education Evaluation and Regional Assistance.

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

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*. Cambridge: 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, NY: Springer.

Chavez, O., Papick, I., Ross, D. J., & A., G. D. (2010). *The Essential Roel 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 Publishers.

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. (in press). Next generation digital classroom assessment based on learning trajectories in mathematics. In C. Dede & J. Richards (Eds.), *Steps toward a digital teaching platform*. New York: Teachers College Press.

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

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

Elmore, R. F. (2002). *Bridging the Gap Between Standards and Achievement*. Washington, DC: Albert Shanker Institute.

Gould, S. J. (1996). *Full House: The Spread of Excellence for Plato to Darwin*. New York: Three Rivers Press.

Grouws, D. H., Reys, R., Papick, I., Tarr, J., Chavez, O., Sears, R., Soria, V. M., & Taylan, R. D. (2010). COSMIC: Comparing Options in Secondary Mathematics: Investigating Curriculum, 2010, from http://cosmic.missouri.edu/

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.

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

Krupa, E. E., & Confrey, J. (2010). Using Instructional Coaching to Customize Professional Development in an Integrated High School Mathematics Program *NCTM Yearbook* (Vol. 74): NCTM.

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., & Wilensky, U. (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.

Means, B., & Penuel, W. R. (2005). Scaling up technology-based educational innovations. In C. Dede, J. P. Honan & L. C. Peters (Eds.), *Scaling up technology-based educational innovations*. San Francisco: Jossey-Bass.

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

Penuel, W. R., Confrey, J., Maloney, A. P., & Rupp, A. A. (submitted). Design Decisions in Developing Assessments of Learning Trajectories: A Case Study. *International Journal of the Learning Sciences*.

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, 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.

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.

Tatar, D. (2007). The Design Tensions Framework. *Human-Computer Ineractions 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.* North Carolina State University, Raleigh, NC.