**Interaction of Research, Practice, and Policy in Mathematics Education**

**Symposium Justification**

This symposium will address different perspectives, along with empirical studies, on the relationships between research, practice, and policy in mathematics education. Various governmental agencies, school personnel, and educational researchers have expressed a need for research-based pedagogical approaches and instructional materials for mathematics education. However, the phrase “research-based” has not been adequately defined, and perspectives on the research-and-development process (even if educators are well advised to use “research-to-practice” strategies) vary widely.

Presentation 1 will present a framework for a fundamental activity, the development of research-based curricula. This framework (a) posits that a valid scientific curriculum development program should address two basic issues—effect and conditions—in three domains, practice, policy, and theory and (b) consists of 10 phases of the development classified into four categories, A Priori Foundations, Learning Model, Formative Evaluation, and Summative Evaluation. One implication is that traditional strategies such as research-to-practice models are insufficient; more adequate is the use of multiple phases of the proffered framework.

Presentation 2 will report several studies conducted regarding one implementation of that framework. There may be no more challenging educational and theoretical issue than scaling up educational programs across a large number of diverse populations and contexts in the early childhood system in the U.S., avoiding the dilution and pollution that usually plagues such efforts to achieve broad success. A series of studies evaluated a research-based model to meet this challenge in the area of mathematics, with the intent the model generalize to other subject matter areas and other age groups.

Presentation 3 will report a large-scale randomized trial of an algebra curriculum that has been developed and field tested over a period of ten years. This study compares two primary theories about the sequence of the curriculum, (a) “Stretch Algebra,” in which students begin to study algebra from the beginning of the course, stopping throughout the course to backfill on intermediate skills and (b) “Catch up/Algebra 1” (the basis for the new curriculum), in which students take a structured “catch up class” during the first semester, followed by Algebra 1 during the second semester.

Presentation 4 will report on an initial evaluation study of Math Recovery (MR), a pullout, one-to-one tutoring program that has been designed to increase mathematics achievement among low-performing first graders, thereby closing the school-entry achievement gap and enabling participants to achieve at the level of their higher-performing peers in the regular mathematics classroom, with goals not merely to assess whether MR works, for whom, and under what circumstances.

The discussant, Larry Hedges, will critique these perspectives and presentations, offering commentary on how the individual presentations and the common themes advance our knowledge of causal relations important for educational effectiveness.
**Presentations**

**Introduction and Presentation 1:**
Douglas H. Clements (clements@buffalo.edu), University at Buffalo, The State University of New York

"Curriculum Research Framework—Beyond the Research-to-Practice Model"

**Presentation 2:**
Julie Sarama (jsarama@buffalo.edu), Douglas H. Clements, Mary Elaine Spitler, Alissa Lange, Christopher Wolfe, University at Buffalo, The State University of New York

"Evaluation of an Intervention based on the *Curriculum Research Framework*: Scale Up"

**Presentation 3:**
Ruth Curran Neild (rneild@csos.jhu.edu), Robert Balfanz (rbalfanz@CSOS.jhu.edu), Vaughan Byrnes, Johns Hopkins University

"Early Evidence from a Randomized Trial of Two Algebra Sequences for Underprepared Freshmen."

**Presentation 4:**
Thomas Smith (thomas.smith@vanderbilt.edu), Paul Cobb (paul.cobb@vanderbilt.edu), Dale Farran, David Cordray, Charles Munter, Sarah Green, Annie Garrison, and Alfred Dunn, Vanderbilt University

"Evaluating Math Recovery: Implications for Policy and Practice"

**Discussant:**
Larry Hedges (l-hedges@northwestern.edu), Northwestern
Abstract Title Page

Title:

Curriculum Research Framework: Beyond the Research-to-Practice Model

Author:

Douglas H. Clements (with Julie Sarama), University of Buffalo, State University of New York
Abstract Body

Background/context:

This presents a research framework that forms a theoretical and research design context for the empirically-based studies that constitute the remainder of the symposium. Also presented are one of a series of studies that evaluated an implementation of this framework.

Purpose:

Government agencies and members of the educational research community have petitioned for research-based curricula. The ambiguity of the phrase “research-based,” however, undermines attempts to create a shared research foundation for the development of, and informed choices about, classroom curricula. I will present a framework for the construct of research-based curricula. One implication is that traditional strategies such as research-to-practice models are insufficient; more adequate is the use of multiple phases of the proffered Curriculum Research Framework.

The Framework:

Our "Curriculum Research Framework" (CRF, Clements, 2007) rejects the sole use of commercially-oriented "market research" and "research-to-practice" strategies. Although included in the CRF, such strategies are inadequate. For example, because they employ one-way translations of research results, research-to-practice strategies are flawed in their presumptions, insensitive to changing goals in the content area, and unable to contribute to a revision of the theory and knowledge.

Goals of the framework. Such knowledge building is—alongside the development of a scientifically-based, effective curriculum—a critical goal of a scientific curriculum research program. Indeed, a valid scientific curriculum development program should address two basic issues—effect and conditions—in three domains, practice, policy, and theory, as described in Table 1 (please insert Table 1 about here).

Structure of the framework. Our Curriculum Research Framework (CRF) constitutes a theoretical structure for such a complete program. The CRF consists of 10 phases of the development research process that warrant claiming that a curriculum is based on research, as outlined in Table 2 (please insert Table 2 about here). These 10 phases are classified into four categories, A Priori Foundations, Learning Model, Formative Evaluation, and Summative Evaluation.

The first category, A Priori Foundations, includes three variants of the research-to-practice model, in which extant research is reviewed and implications for the nascent curriculum development effort drawn. (1.) In General A Priori Foundation, developers review broad philosophies, theories, and empirical results on learning and teaching. Based on theory and research on early childhood learning and teaching (Bowman, Donovan, & Burns, 2001; Clements, 2001), we determined that Building Blocks’ basic approach would be finding the mathematics in, and developing mathematics from, children's activity, such as "mathematizing" everyday tasks. (2.) In Subject Matter A Priori Foundation, developers review research and consult with experts to identify mathematics that makes a substantive contribution to students'
mathematical development, is generative in students’ development of future mathematical understanding, and is interesting to students. We determined subject matter content by considering what mathematics is culturally valued (e.g., NCTM, 2000) and empirical research on what constituted the core ideas and skill areas of mathematics for young children (Baroody, 2004; Clements & Battista, 1992; Clements & Conference Working Group, 2004; Fuson, 1997), including hypothesized syncretism among domains, especially number and geometry. We revised the subject matter specifications following a content analysis by four mathematicians and mathematics educators, resulting in learning trajectories in the domain of number (counting, subitizing, sequencing, arithmetic), geometry (matching, naming, building and combining shapes), patterning, and measurement. (3.) In Pedagogical A Priori Foundation, developers review empirical findings on making activities educationally effective—motivating and efficacious—to create general guidelines for the generation of activities. As an example, research using computer software with young children (Clements, Nastasi, & Swaminathan, 1993; Clements & Swaminathan, 1995; Steffe & Wiegel, 1994) showed that preschoolers can use computers effectively and that software can be made more effective by employing animation, children’s voices, and clear feedback.

In the second category, Learning Model, developers structure activities in accordance with empirically-based models of children’s thinking in the targeted subject-matter domain. This phase, (4) Structure According to Specific Learning Model, involves creation of research-based learning trajectories, which, of course, have been described in detail in this book.

In the third category, Evaluation, developers collect empirical evidence to evaluate the appeal, usability, and effectiveness of a version of the curriculum. Past phase (5) Market Research is (6) Formative Research: Small Group, in which developers conduct pilot tests with individuals or small groups on components (e.g., a particular activity, game, or software environment) or sections of the curriculum. Although teachers are involved in all phases of research and development, the process of curricular enactment is emphasized in the next two phases. Studies with a teacher who participated in the development of the materials in phase (7) Formative Research: Single Classroom, and then teachers newly introduced to the materials in phase (8) Formative Research: Multiple Classrooms, provide information about the usability of the curriculum and requirements for professional development and support materials. We conducted multiple case studies at each of these three phases (e.g., Clements & Sarama, 2004a; Sarama, 2004), revising the curriculum multiple times, including two distinct published versions (Clements & Sarama, 2003, 2007a). In the last two phases, (9) Summative Research: Small Scale and (10) Summative Research: Large Scale, developers evaluate what can actually be achieved with typical teachers under realistic circumstances. Phases 9 and 10 use cluster randomized trials, which provide the most efficient and least biased designs to assess causal relationships (Cook, 2002), in an increasing number of classrooms, with increasing diversity, and decreasingly “ideal” conditions.

An initial phase-9 summary research project (Clements & Sарамa, 2007c), yielded effect sizes between 1 and 2 (Cohen’s d1977). This study only involved 4 classrooms, however. We briefly describe a larger phase-9 study in the following sections.

Setting and Participants:

This study (Clements & Sarama, 2008) involved 36 diverse classrooms serving children
from low-income households, including Head Start and state-funded programs in New York.

Research Design:

Classrooms were randomly to one of three conditions. The experimental group used *Building Blocks* (Clements & Sarama, 2007b). The comparison group used a different preschool mathematics curriculum—the same as we previously used in the PCER research (mainly Klein, Starkey, & Ramirez, 2002). The control used their schools existing curriculum.

Findings:

Two observational measures indicated that the curricula were implemented with fidelity and that the experimental condition had significant positive effects on classrooms' mathematics environment and teaching. The experimental group score increased significantly more than the comparison group score (effect size, .47) and the control group score (effect size, 1.07).

Conclusions:

*Conclusions of the previous study:* Focused early mathematical interventions, especially those based on a comprehensive model of developing and evaluating research-based curricula, can increase the quality of the mathematics environment and teaching and can help preschoolers develop a foundation of informal mathematics knowledge (Clements & Sarama, 2008).

*Relationships of the CRF to design experiments.* Design experiments are an important component of curriculum research and development activity within the proposed Curriculum Research Framework (CRF). However, this framework shows that other research and development strategies are necessary to meet the goals of a complete curriculum research and development program. Design experiments cannot control the many variables in their complex settings; the large amount of data collected rarely can be analyzed fully before the next cycle of revision, enactment, and analysis takes place (Collins, Joseph, & Bielaczyc, 2004); and different participants may have different data and perspectives, so that the ultimate paths and products may be arbitrary to an extent and generalization may be difficult (Kelly, 2004). Experiments (randomized trials designs) provide some of what design experiments cannot. In addition, the use of phases in the A Priori Foundations and Learning Model categories of the CRF provide useful constraints and theoretical groundings for design experiments.

Conversely, design experiences, as well as other methods such as teaching experiments and classroom-based teaching experiments, can help accomplish what randomized trials cannot. In the CRF context, these methods include conceptual and relational, or semantic, analysis, and thus are theoretically grounded. As such, they allow researchers to build models of the child’s mathematics, of mental actions on objects, of learning, and of teaching interactions. Because it includes a coherent complement of methods, the CRF has built-in checks and balances that address the limitations of each method, with concentration on the learning model especially useful for maintaining theoretical and scientific foci.

In summary, embedding design experiments within the CRF encourages the conduct of a comprehensive research and development program. Such a program is theoretically sound and scientifically defensible. Further, it addresses the full range of curriculum research questions, about effects and conditions in the three domains of policy, practice, and theory (Clements,
CRF-based programs thus contribute to both the research field and produce a curriculum that is a contribution to practice.

Implications:

1. **Using the multiple phases in the proposed Curriculum Research Framework (CRF) will help developers improve curricula and contribute to the field of curriculum research.** Particular research designs and methods are suited for specific kinds of investigations and questions, but can rarely illuminate all the questions and issues in a line of inquiry. This is why different methods are used in various phases of the CRF (National Research Council, 2004). For example, although iterating through one or two of the phases here, such as phase 8, *might* lead to an effective curriculum, this would not meet all the goals outlined in Table 1. The curriculum might be effective in some settings, but not others, or it might be too difficult to scale up. Moreover, we would not know *why* the curriculum is effective.

Using the CRF not only documents if the design is successful in attaining achievement goals, but also traces whether that success can be attributed to the posited theory-design connections. This necessitates developers accepting new responsibilities, such as expanding their knowledge of the subject matter, psychology, and cognitive science, instruction, implementation, and scaling up, as well as of the variety of scientific research methods in the CRF’s phases. Even if multiple phases are used, if they are all *a priori foundations*, for example, they are inadequate. As noted, subtle differences in activities can enhance or sabotage effectiveness (Sarama, 2000). Achieving the goals of the CRF (Table 1) requires refining and especially elaborating principles by ongoing research and development work that tracks the effectiveness of every specific implementation, consistently maintaining links to the hypothesized theories and models, through progressively expanding social contexts. Ensuring that the research trajectory described by the CRF is coherent and connected throughout the development process maintains unbroken threads of argumentation.

2. **Achieving the goals of CRF requires both qualitative and quantitative methodologies** (National Research Council, 2004, makes similar recommendations, albeit for summative research only). In response to theorists who celebrate the “defeat of quantitative research in the curriculum field and the victory of qualitative research” (Pinar, Reynolds, Slattery, & Taubman, 1995, p. 52), we paraphrase Mark Twain to say that the report of its death is an exaggeration. Both approaches can make valid, rigorous contributions to scientific research (Darling-Hammond & Snyder, 1992; Johnson & Onwuegbuzie, 2004; National Research Council, 2004). Quantitative methodologies provide experimental results, garnered under conditions distant from the developers, that are useful in and of themselves and in that they can generate political and public support. Randomized experiments are more powerful and less biased than alternative
designs and also can uncover unexpected and subtle interactions not revealed by qualitative investigations (Clements & Nastasi, 1988; Nastasi, Clements, & Battista, 1990; Russek & Weinberg, 1993).

Qualitative methodologies are important for three reasons. First, curriculum research seeks to understand individual students’ interpretations and learning and how these change in the context of, and as a result of, interactions among teachers and students around a specific curriculum. Qualitative research describes the nature of the “it” when researchers ask, “Did it work?” (Erickson & Gutierrez, 2002); validity is suspect without this information (especially given the possibility of unintended and immeasurable outcomes, Taba, 1962; van Oers, 2003; Walker, 1992). Second, such research helps explain why it works and how and why it works differently in different contexts. Third, qualitative research in a triangulation context may serve to validate or invalidate quantitative results, more so than the inverse (Russek & Weinberg, 1993), and such methodologies complement experiments in ruling out alternate explanations. Experiments control a necessarily small fraction of an indefinite number of contextual variables, and one will rarely identify limiting or catalytic conditions and curricular features (including the aforementioned “subtle differences”) optimally by considering only focal experimental variables (Greenwald, Pratkanis, Leippe, & Baumgardner, 1986). In summary, given its inherently complex and creative nature, its interpretive goals, the small number of students involved in many of its techniques, and the progressive breadth of concerns combined with the consistent need for sensitivity to new findings and insights, curriculum research requires qualitative methodologies and openness to emergent findings throughout the phases (Smith, 1983).

Finally, quantitative and qualitative method are integrated throughout the CRF’s phases. Every experiment benefits from collecting ethnographic data. Conversely, the validity of qualitative methodologies, such as case studies, is increased if they are conducted within the context of an experiment (Cook, 2002). Finally, the use of summative evaluation without other phases is usually premature, wasteful, and misleading. (The medical research model, oft-cited as the gold standard, uses randomized trials, especially large-scale experiments, only after non-random, discovery strategies, exploratory clinical research, dose-response trials, etc., Giorgianni & Granna, 1999; Zaritsky, Kelly, Flowers, Rogers, & O’Neil, 2003.) Thus, although randomized experiments remain the best design for evaluation of causal interpretations, placing them in the context of a complete CRF mitigates the limitations and misuses of randomized experiments (The Design-Based Research Collective, 2003).

3. Increasing academe’s support for curriculum research will improve curricula, research, and the public’s opinion of educational research. There is a long history of bias against design sciences in academe (Simon, 1969; Wittmann, 1995).

4. Curriculum research could be more successful if funding agencies reconsidered time frames and funding requirements for this enterprise. Curriculum research needs increased funding (Feuer, Towne, & Shavelson, 2002). The proportion of funds presently allocated to research in education is inconsistent with virtually any other enterprise (Dow, 1991; President’s Committee of Advisors on Science and Technology—Panel on Educational Technology, 1997; Schoenfeld, 1999).

5. To benefit from curriculum research, the entire education community needs to support and expect research-based curriculum development—and expect that the specific methods used and results obtained are fully explicated. Lack of a connection between research and curriculum
development and adoption is a major reason that curriculum, and ultimately student achievement, in the U.S. do not improve (Battista & Clements, 2000; Clements, 2002; Clements & Battista, 2000) and that curriculum reforms usually fail, with “genuine achievements…thrown out along with excesses and failures” (Walker, 2003, p. 116).

6. Curriculum developers must accept new responsibilities, including reviewing all relevant research and remaining receptive to successes of varied approaches.

7. Policy makers should support and insist on research-based curricula.
Appendices

Appendix A: References


### Appendix B. Tables

**Table 1**

**Goals of Curriculum Research**

<table>
<thead>
<tr>
<th>Practice</th>
<th>Policy</th>
<th>Theory</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Effects</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a. Is the curriculum effective in helping children achieve specific learning goals? Are the intended and unintended consequences positive for children? (What is the quality of the evidence?—Construct and internal validity.) (6-10) *</td>
<td>c. Are the curriculum goals important? (1, 5, 10)</td>
<td>f. Why is the curriculum effective? (all)</td>
</tr>
<tr>
<td>b. Is there credible documentation of both a priori research and research performed on the curriculum indicating the efficacy of the approach as compared to alternative approaches? (all)</td>
<td>d. What is the effect size for students? (9, 10)</td>
<td>g. What were the theoretical bases? (1, 2, 3)</td>
</tr>
<tr>
<td><strong>Conditions</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i. When and where?—Under what conditions is the curriculum effective? (Do findings generalize?—External validity.) (8, 10)</td>
<td>j. What are the support requirements (7) for various contexts? (8-10)</td>
<td>k. Why do certain sets of conditions decrease or increase the curriculum’s effectiveness? (6-10)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>l. How do specific strategies produce previously unattained results and why? (6-10)</td>
</tr>
</tbody>
</table>
Table 2

Categories and Phases of the Curriculum Research Framework (CRF)

<table>
<thead>
<tr>
<th>Categories</th>
<th>Questions Asked</th>
<th>Phases</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A Priori Foundations.</strong></td>
<td>What is already known that can be applied to the anticipated curriculum?</td>
<td>Established review procedures (e.g., Light &amp; Pillemer, 1984) and content analyses (NRC, 2004) are employed to garner knowledge concerning the specific subject matter content, including the role it would play in students’ development (phase 1); general issues concerning psychology, education, and systemic change (phase 2); and pedagogy, including the effectiveness of certain types of activities (phase 3).</td>
</tr>
<tr>
<td>In variants of the research-to-practice model, extant research is reviewed and implications for the nascent curriculum development effort drawn.</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Goals</strong></td>
<td>Phase b c f g 1</td>
<td></td>
</tr>
<tr>
<td><strong>Goals</strong></td>
<td>Phase b f g 2</td>
<td></td>
</tr>
<tr>
<td><strong>Goals</strong></td>
<td>Phase b f g 3</td>
<td></td>
</tr>
<tr>
<td><strong>Learning Model.</strong></td>
<td>How might the curriculum be constructed to be consistent with models of students’ thinking and learning (which are posited to have characteristics and developmental courses that are not arbitrary and therefore not equally amenable to various instructional approaches or curricular routes)?</td>
<td>In phase 4, the nature and content of activities is based on models of children’s mathematical thinking and learning (cf. James, 1892/1958; Tyler, 1949). In addition, a set of activities (the hypothetical mechanism of the research) may be sequenced according to specific learning trajectories (Clements &amp; Sarama, 2004b). What distinguishes phase 4 from phase 3, which concerns pedagogical a prior foundations, is not only the focus on the child’s learning, rather than teaching strategies alone, but also the iterative nature of its application. That is, in practice, such models are usually applied and revised (or, not infrequently, created anew) dynamically, simultaneously with the development of instructional tasks, using grounded theory methods, clinical interviews, teaching experiments, and design experiments.</td>
</tr>
</tbody>
</table>
**Evaluation.** In these phases, empirical evidence is collected to evaluate the curriculum, realized in some form. The goal is to evaluate the appeal, usability, and effectiveness of an instantiation of the curriculum.

<table>
<thead>
<tr>
<th>Goals</th>
<th>Phase</th>
</tr>
</thead>
<tbody>
<tr>
<td>b c f</td>
<td>5</td>
</tr>
</tbody>
</table>

How can market share for the curriculum be maximized?

**Phase 5** focuses on marketability, using strategies such as gathering information about mandated educational objectives and surveys of consumers.

<table>
<thead>
<tr>
<th>Goals</th>
<th>Phase</th>
</tr>
</thead>
<tbody>
<tr>
<td>a b f h k l</td>
<td>6</td>
</tr>
</tbody>
</table>

Is the curriculum usable by, and effective with, various student groups and teachers? How can it be improved in these areas or adapted to serve diverse situations and needs?

Formative phases 6 to 8 seek to understand the meaning that students and teachers give to the curriculum objects and activities in progressively expanding social contexts; for example, the usability and effectiveness of specific components and characteristics of the curriculum as implemented by a teacher who is familiar with the materials with individuals or small groups (*phase 6*) and whole classes (*phase 7*) and, later, by a diverse group of teachers (*phase 8*). Methods include interpretive work using a mix of model testing and model generation strategies, including design experiments, microgenetic, microethnographic, and phenomenological approaches (*phase 6*), classroom-based teaching experiments and ethnographic participant observation (*phase 7*), and these plus content analyses (*phase 8*). The curriculum is altered based on empirical results, with the focus expanding to include aspects of support for teachers.

<table>
<thead>
<tr>
<th>Goals</th>
<th>Phase</th>
</tr>
</thead>
<tbody>
<tr>
<td>a b f h j k l</td>
<td>7</td>
</tr>
<tr>
<td>a b f i j k l</td>
<td>8</td>
</tr>
</tbody>
</table>

(con’t)
What is the effectiveness (e.g., in affecting teaching practices and ultimately student learning) of the curriculum, now in its complete form, as it is implemented in realistic contexts?

Summative phases 9 and 10 both use randomized field trials and differ from each other most markedly on the characteristic of scale. That is, phase 10 examines the fidelity or enactment, and sustainability, of the curriculum when implemented on a large scale, and the critical contextual and implementation variables that influence its effectiveness.

Experimental or carefully planned quasi-experimental designs, incorporating observational measures and surveys, are useful for generating political and public support, as well as for their research advantages. In addition, qualitative approaches continue to be useful for dealing with the complexity and indeterminateness of educational activity (Lester & Wiliam, 2002).

*Goals refer to the specific questions in Table 1, answers to which are the goals of the CRF.
Title:

Evaluation of an Intervention Based on the *Curriculum Research Framework*: Scale Up

Author:

*Julie Sarama* (with Douglas H. Clements, Mary Elaine Spitler, Alissa Lange, Christopher Wolfe), University at Buffalo, The State University of New York
Abstract Body

Background/context:

Although the successes of some research-based educational practices have been documented, equally recognized is the “deep, systemic incapacity of U.S. schools, and the practitioners who work in them, to develop, incorporate, and extend new ideas about teaching and learning in anything but a small fraction of schools and classrooms” (see also Berends, Kirby, Naftel, & McKelvey, 2001; Cuban, 2001; Elmore, 1996, p. 1; Tyack & Tobin, 1992). There may be no more challenging educational and theoretical issue than scaling up educational programs across a large number of diverse populations and contexts in the early childhood system in the U.S., avoiding the dilution and pollution that usually plagues such efforts to achieve broad success. We created a research-based model to meet this challenge in the area of mathematics, with the intent the model generalize to other subject matter areas and other age groups. The field needs transferable, practical examples of scale up (McDonald, Keesler, Kauffman, & Schneider, 2006); empirical evidence of the effectiveness of these examples; and focused research on critical variables—all leading to refined, generalizable theories and models of scale up. Our research plan describes a project designed to meet those needs.

The specific goal of our implementation of the TRIAD (Technology-enhanced, Research-based, Instruction, Assessment, and professional Development) model is to increase math achievement in young children, especially those at risk, by means of a high-quality implementation of the Building Blocks math curriculum, with all aspects of the curriculum—mathematical content, pedagogy, teacher’s guide, technology, and assessments—based on a common core of learning trajectories. The TRIAD intervention provides (a) these curriculum materials; (b) ongoing professional development, including scalable distance education, a web-based application with extensive support for teaching based on learning trajectories, and classroom-based coaching during the school year; and (c) supportive roles and materials for parents and administrators.

TRIAD’s theoretical framework (Sarama, Clements, Starkey, Klein, & Wakeley, 2008) is an elaboration of the Network of Influences model (Sarama, Clements, & Henry, 1998), illustrated in Figure 1 (please insert figure 1 here). It is consistent with, but extends in levels of detail, such theories as diffusion theory and the overlapping spheres of influence (Rogers, 2003; Showers, Joyce, & Bennett, 1987).

Purpose:

There are three primary research questions, each a cluster of several components.

1. **Can the intervention be implemented with high fidelity and have substantial positive effects on teachers’ beliefs and knowledge?** Previous research indicated that most teachers implemented the curriculum with high fidelity and made significant gains in knowledge (Clements & Sarama, 2008; Sarama, et al., 2008). We evaluated whether all aspects of the intervention could be replicated in greater numbers of classrooms, in distant, diverse sites. That is, did teachers participate in the complete professional development program, receive the prescribed support, implement the curriculum with high fidelity, and provide a richer classroom environment for mathematics? Do their knowledge and beliefs about early childhood mathematics develop as a result of the intervention?

2. **What are the immediate effects of the scale up intervention, as implemented under diverse conditions, on achievement and the achievement gap?** In our previous IERI-funded evaluation of TRIAD, children in the experimental group outperformed children in the control group in math achievement at the end of pre-K. We determined whether these findings are replicated with the same effect size under full scale up. Do experimental children from low-income homes and ethnic and linguistic
minority groups narrow the achievement gap that separates them from children from middle- and high-income and non-minority homes?

3. **What are the long-range (persistence of) effects of the intervention, with and without follow through, on achievement and the achievement gap?** These critical questions have not been addressed for TRIAD. The present research included two experimental groups (and one control group). In both, pre-K teachers participated in the intervention. In the Follow-Through experimental group, teachers in grades K and 1 were taught about the pre-K intervention and ways to build upon it. Do both experimental groups outperform those in the comparison group in math achievement on the average, at the end of kindergarten (first grade data are still being cleaned)? Do children in the experimental TRIAD Follow-Through group on the average outperform children in the TRIAD (non-follow through) experimental group? Do children in both experimental groups from low-income homes and ethnic and linguistic minority groups narrow the achievement gap? Does the gap narrow more for children in the follow-through experimental group?

**Setting:**

The study took place in pre-K classrooms in two urban school districts, the Buffalo Public School system in Buffalo, NY and the Boston Public School system in Boston, MA (a third site, in Nashville, TN/Vanderbilt University, started later and those data are being integrated at the time of this writing).

**Participants:**

In the Buffalo Public Schools, all schools whose pre-K teachers had not previously been involved in Building Blocks (e.g., Clements & Sarama, 2007; Clements & Sarama, 2008; Sarama, 2004; Sarama & Clements, 2002) or TRIAD (Sarama, et al., 2008) research or development projects were included. Extending the evaluation to distal sites—essential for generalizing to the target population of all U.S. pre-Ks—all Boston, MA schools that were not adopting a new pre-K curriculum that year and whose principals agreed to participate were included. For the pre-K study, there was only 5% attrition (from control, 12 from Buffalo and 6 from Boston from TRIAD, 37 in Buffalo and 15 from Boston schools, for a total of 70), most moved out of state, but some were ill for the entire posttest period, leaving a total of 1305 children with complete data on both pretest and posttest). Analyses revealed no significant difference between those who left and remained in mean pretest achievement ($F = 2.09(2), p = .148$); further, the small effect size, ES < .01, indicates that any effect on the findings was negligible. By the end of the kindergarten year, the population remaining in the original schools include 348 children in the TRIAD Follow Through group, 335 in the TRIAD group, and 286 in the control group.

**Intervention:**

The specific goal of our implementation of the TRIAD (Technology-enhanced, Research-based, Instruction, Assessment, and professional Development) model is to increase math achievement in young children, especially those at risk, by means of a high-quality implementation of the Building Blocks math curriculum. This includes all aspects of the curriculum—mathematical content, pedagogy, teacher’s guide, technology, and assessments—based on a common core of learning trajectories. Building Blocks, the $I$ in the TRIAD acronym, was based on a comprehensive Curriculum Research Framework (Clements, 2007) and its efficacy validated by two Cluster Randomized Trial (CRT) evaluations, yielding effect sizes ranging from .5 to over 2 (Clements & Sarama, 2007, 2008). The Assessment component of TRIAD includes both formative assessment performed by the teachers training to use learning trajectories for this purpose, supplemented by the Building Blocks Software management system. TRIAD’s professional Development includes multiple forms of training (15 full days over two years, the first year a “gentle introduction” with no data collection) and support (coaching and mentoring). Each of these uses the software application, Building Blocks Learning Trajectories (BBLT), which presents and connects all
components of the innovation. BBLT provides scalable access to the learning trajectories via descriptions, videos, and commentaries. The two main aspects of each learning trajectories—the developmental progressions of children’s thinking and connected instruction—are linked to the other (see Fig. 2—please insert Figure 2 about here).

**Research Design:**

In a CRT design, schools within each district were ordered on the basis of their average scores on state-based math achievement tests and then publicly assigned to one of three treatment groups using a randomized block design (using a table of random numbers, with blind pointing to establish the starting number). We attempted to implement components of TRIAD implemented in ways that would be available under normal condition. The first year was a pilot/training year, as our current work confirms that of others indicating that teachers take at least two years before they begin implementing the curriculum conceptually and completely. Assessors were trained and validated during this time.

**Data Collection and Analysis:**

All assessments were completed the second year, including two measures of teachers’ classroom practices (e.g., implementation fidelity), knowledge, and beliefs—Fidelity of Implementation, Classroom Observation of Early Mathematics Environment and Teaching (COEMET) and two groups of measures of child outcomes, math (pre-K and Knsg., Research-based Elementary Math Assessment, REMA), and language and literacy (pre-K: PALS-Pre-K; MCLASS: CIRCLE; beginning of Knsg., Renfrew Bus Story).

To answer question 1, Can the intervention be implemented with high fidelity and have substantial positive effects on teachers’ practice, beliefs and knowledge?, factorial repeated measures analyses were conducted on the Fidelity (intervention groups) and COEMET (both groups) T-scores. (A Teacher Questionnaire was also employed to collect data on demographics and teacher beliefs.)

The other two questions were answered with hierarchical linear modeling (HLM, Stephen W. Raudenbush, Bryk, Cheong, & Congdon, 2000; S. W. Raudenbush & Liu, 2003). All level-2 predictors were centered around their group means. All interactions were computed on mean-centered transformations of the variables involved. Effect sizes were computed for significant main effects by dividing the regression coefficient by the pooled posttest standard deviation.

**Findings/results:**

**Pre-K.** HLM analyses revealed that the two groups differed significantly in math achievement and that both improved significantly and substantially (more than 1 SD). The TRIAD group outperformed the control group \( p < .0001 \), with an effect size of .69. There were no significant main effects or interactions for SES or LEP (percentage of students with Limited English Proficiency). At the child level, there were no significant interactions for sex or disability status and only one interaction for ethnic group—African American children in the control group learned significantly less than other children in the control group, but African American children in the TRIAD group scored significantly learned significantly more than other children in the TRIAD group.

Fidelity scores were relatively high (on Likert scale items, averaging “agree”). COEMET scores were significantly higher for the TRIAD group at both time periods, Fall and Spring \( p < .0001, \text{ ES = 1.13} \). COEMET scores were a significant mediator, although use of the Building Blocks curriculum alone accounted for more variance. Of the items, those that accounted for the most variance were: the number of specific math activities (SMAs), the number of computers running the Building Blocks software, and the total classroom culture score. The role of the SMAs raised the question of whether this was merely an
indicator of the ubiquitous “time on task” variable. To check this possibility, we returned to the original COEMET and computed the total instructional time for all activities. This time-on-task variable was not significantly related to child gain.

There were no differences between the groups on letter recognition. There were no differences between the groups on three measures from the oral language (Bus Story; recall this was measured at the beginning of the kindergarten year) measure: sentence length, listening, and story duration. The TRIAD group outperformed the control group on three language measures: information (ES = .33), complexity (ES = .16), and independence (ES = .36).

Kindergarten. As in pre-K, the TRIAD group outperformed the control group in math achievement at the end of kindergarten (ES = .38, p <.01). In addition, the TRIAD Follow-Through group outperformed both the control group (ES = .55, p <.01) and the TRIAD group (ES = .33, p <.05).

Conclusions:

High levels of fidelity of implementation resulted in consistently higher scores in the intervention classrooms, compared to control classrooms, on the COEMET, and statistically significant and substantially greater gains in children’s math achievement in the intervention, compared to the control, children. The greatest treatment effects on the observation instrument were on several classroom culture items (responsiveness to children, use of teachable moments, and especially use of computer to teach math), and items regarding specific math activities (the number of activities, eliciting and extending children’s mathematical thinking, and especially observing and listening to children and using formative assessment).

The most significant contributors to children’s learning were the number of specific math activities (SMAs), the number of computers running the Building Blocks software, and the total classroom culture score. The role of the SMAs raised the question of whether this was merely an indicator of the ubiquitous “time on task” variable. This time-on-task variable was not significantly related to child gain, so that it was the number of SMAs, not the total instructional time, that affected children’s learning. There are several possibilities to account for this phenomenon. It may be that developmental limitations in attention constrain young children’s ability to learn past TRIAD had a positive effect on math achievement. Probably due to both districts’ new emphasis on pre-K math, the control group gained a surprising standard deviation (especially compared to national pre-K math gains). Thus, the effect size of .69 for TRIAD is noteworthy. Teachers teaching more and better math accounted for some of the children’s gains; just using the curriculum accounted for the rest.

Extensive time spent on mathematics in pre-Kindergarten does not appear to hinder early literacy and language performance. Children exposed to the Building Blocks early math curriculum did not differ significantly from children who participated in the regular district math curriculum on letter recognition or on a number of language measures. Children who received the Building Blocks curriculum outperformed the control group on two language measures, information and independence. The Building Blocks group had an advantage over the control group in recall of key content from the narrative, a skill that has been linked to academic success at age 15. The higher independence scores for children in the Building Blocks group indicates that may be more confident than the control group in verbalizing their thoughts. Aspects of the curriculum, such as the focus on verbal explanations for solutions to math problems, may account for these differences. Results on these two measures suggest that children who learn math through Building Blocks may also develop important competencies that can transfer to other academic areas, such as oral language.

The theory of scale up employed held up under a CRT experiment. Centering on a common core of learning trajectories and the combination of workshops, coaching, mentoring, and Internet tools
focused on a common curriculum, yielded a successful scale up with effects not dissimilar to ideal conditions.

There are five basic recommendations. (1) Curriculum and policy should ensure that children, especially those living in poverty, should be provided with research-based, focused early mathematical interventions which can increase their knowledge of multiple mathematical concepts and skills (including, but also going beyond number) without harming—and actually increasing—their language and literacy skills. (2) Substantial professional development may be necessary—15 full days were barely adequate to the task in this study, consistent with previous research (cf. an average of 53 hours of teaching training yielding an average effect size of .53 on student math achievement in Yoon, Duncan, Lee, Scarloss, & Shapley, 2007), suggesting a minimal duration for effective professional development. (3) The Curriculum Research Framework (Clements, 2007) upon which the curriculum was based has been repeatedly empirically supported and may serve as a useful guide to policy makers, curriculum and software developers, and administrators. (4) The learning trajectories at the core of the curriculum and TRIAD model may constitute a useful construct in future research, curriculum development, and professional development efforts.

(5) Finally, we need more research on the conditions that children from early interventions enter in the primary school years. Multiple studies have reported on gains that “fade” without adequate attention to the follow-up—more frequently, the lack of follow-up—planned and implemented for these children. Even in our TRIAD Follow-Through treatment, multiple factors impeded implementation, including teacher’s views that district rules and “fidelity police” demanded following scripts and schedules exactly—and would not allow formative assessment or curriculum contraction.
Appendixes

References


Appendix B. Tables and Figures

Figure 1: Revised Network of Influences Theoretical Framework*

* Contextual variables in dotted ovals include the school (A-D), teacher (E), and child (F-H) factors. For example, child socioeconomic status, or SES (G), impacts children’s initial math knowledge (H), which influences children’s achievement (R)—an outcome variable indicated by the solid rectangle. Implementation variables in solid ovals are features that the project can encourage and support, but cannot control absolutely. For example, heavy arrows from professional development (J), to teacher knowledge (N), to implementation fidelity (O), to child achievement (R), indicate the strong effects in that path. Support from coaches (L) also has a strong effect on implementation fidelity, while other factors (J, K, M) are influential, but to a moderate degree (not all small effects are depicted). Relationships are further described in the following section.
**Building Blocks Learning Trajectories (BBLT) Web Application**

BBLT provides scalable access to the learning trajectories via descriptions, videos, and commentaries. Each aspect of the learning trajectories—developmental progressions of children’s thinking and connected instruction—are linked to the other. For example, teachers might choose the (curriculum) view and see the screen on the left, below. Clicking on a specific activity provides a description. Clicking on slides the screen over to reveal descriptions, several videos of the activity “in action,” notes on the video, and the level of thinking in the learning trajectory that activity is designed to develop, as shown below on the right. (See UBTRIAD.org for a demonstration.)

Clicking on the related developmental level, or child’s level of thinking, ringed above, switches to the view of that topic and that level of thinking. This likewise provides a description, video, and commentary on the developmental level—the video here is of a clinical interview task in which a child displays that level of thinking. Teachers can also study a development view, studying clinical interviews of children at each level of thinking, and, if desired, link back to activities.
Title: Early Results from a Randomized Trial of Two Algebra Sequences for Underprepared Freshmen

Author(s): Ruth Curran Neild (with Vaughan Byrnes and Robert Balfanz), Johns Hopkins University
Abstract

Background/context

One of the aims of the education standards movement is to make intellectually demanding course work in high school the norm. In the area of mathematics, a growing number of states and districts now require that students not only earn an Algebra 1 credit to graduate, but that they take Algebra 1 during their freshman year. In addition, some large school districts have made earning an Algebra 1 credit a requirement for promotion to 10th grade. There are compelling arguments for enrolling most – if not all – freshmen in Algebra 1. Successful completion of Algebra 1 in the first year of high school places all students on an equal footing, credit-wise, to reap the benefits of advanced mathematics courses in high school.

The painful reality associated with such mandates, however, is that Algebra 1 course failure rates among freshman are typically very high. When Los Angeles required all freshmen to take Algebra 1 in Fall 2004, only 39% of the students earned a C or better in the course and 44% failed (Helfand, 2006). Milwaukee’s “algebra for all” policy resulted in about half of the freshmen failing algebra, on average, over a 7 year period (Ham and Walker, 1999). These and other data suggest that mandates alone – without a more effective instructional approach - will not produce substantially higher percentages of disadvantaged students who are “college ready” and may actually accelerate the high school dropout rate as numerous students become discouraged over their prospects of earning the credits needed for graduation.

A large part of the reason for the dismal Algebra 1 results in cities like Los Angeles is that most first-time freshmen in public schools in these districts are seriously under-prepared to succeed in a rigorous Algebra 1 curriculum if additional supports are not provided. The traditional Algebra 1 course assumes that students have mastered basic and intermediate math skills, including multiplication and division of fractions, decimals, and signed numbers. Students who are below grade level in mathematics, however, struggle to perform operations with rational numbers and integers (National Research Council, 2001; National Mathematics Advisory Panel, 2008). A second area in which many students need extra help is with the transition from arithmetic to mathematics. One of the central TIMSS findings is that the mathematics curriculum in U.S. schools is highly repetitive and remains strongly focused on arithmetic between the 4th and 8th grades. Students in the US are taught fewer advanced mathematical topics in 7th and 8th grade than are students in higher achieving nations (Schmidt et al., 1999). As a result, the learning curve in high school mathematics can be very steep. A perusal of the leading algebra textbooks used in the 9th grade, for example, indicates that many of the texts begin with a rapid “review” of probability, proportional reasoning, measurement, data, and geometry - topics that large numbers of students, according to TIMSS data, have had limited opportunity to learn in middle school.

One of the strategies used by districts with large percentages of freshmen who are underprepared for Algebra 1 is “double dosing.” Double-dose classes allow for approximately 70-90 minutes of instruction in a subject per day, throughout the school year. The idea behind the double dose is that additional time spent on mathematics in the ninth grade provides an opportunity for students to catch up on intermediate skills and/or to proceed through algebra at a slower pace, allowing time for extra practice and the clearing up of misconceptions.

Additional class time allowed by the double dose opens the question of what to do with that extra time. There are two primary theories about the sequence of the curriculum. The first theory, sometimes called “Stretch Algebra” or “Algebra 1A/1B,” is that students should begin to
study algebra from the beginning of the course, stopping throughout the course to backfill on intermediate skills as when the teacher observes that review is needed. The second theory is that students should take a structured “catch up class” during the first semester, followed by Algebra 1 during the second semester.

**Research questions**

The research compared the impacts on student achievement and credit accumulation of the two curriculum sequences described above (“Stretch Algebra” v. “catch up/Algebra 1”), for freshmen who entered high school between 1 and 4 years below grade level. The “catch up/Algebra 1” strategy was operationalized as a sequence of the Transition to Advanced Mathematics (TAM) curriculum, developed by the Johns Hopkins University Center for Social Organization of Schools, followed by an Algebra 1 curriculum of the school’s choosing. The TAM curriculum is described in more detail in a subsequent section of this abstract. No specific curriculum was specified for the “Stretch Algebra” strategy; schools were allowed to choose their own scope and sequence as well as materials.

The key research questions were:

1) Are there end-of-course differences in Algebra 1 proficiency between a) those 9th graders who spend time building intermediate skills before taking an Algebra 1, in comparison to b) those who are placed directly into an Algebra 1 class that has additional instructional time built-in? If differences exist, are they large enough to be educationally significant?

2) Are there mid-year differences in growth in intermediate mathematics skills and mathematical reasoning between students in the two conditions? To what extent can end-of-course Algebra 1 proficiency be attributed to gains in intermediate skills and reasoning acquired during the first semester?

3) Are there differences between students in the two conditions in credit accumulation in Algebra 1 and in course grades (of C or better) that would indicate potential for success in future mathematics classes?

**Setting**

A multi-district randomized trial took place in eight school districts during the 2008-2009 school year and is taking place in five additional districts during the 2009-2010 school year. These districts are medium to large districts with at least two high schools serving a minimum of 75 first-time freshmen who are taking Algebra 1 but are underprepared in mathematics. The districts are located in the Northeast, South, Southwest, and West.

**Participants**

Forty-six schools in thirteen districts have participated in the study. Overall, 133 teachers and about 5,000 students participated in either 2008-2009 or 2009-2010.

**Intervention**

Within each district, schools were randomly assigned to implement either the Stretch Algebra or the TAM/Algebra 1 sequence. Districts with two participating schools had one
Stretch Algebra school and one TAM/Algebra 1 school; districts with four participating schools had two Stretch Algebra schools and two TAM/Algebra 1 schools; and so on. Each participating district had at least two schools in the study; the maximum number of schools was six. As part of the study, each school implemented their assigned condition for a single year.

Each district selected its own Algebra 1 text, materials, and scope and sequence for Stretch Algebra. The relative emphasis of the Algebra 1 courses varied in part as a result of state standards and the specificity of Algebra 1 courses of study.

Each TAM/Algebra 1 school received all curriculum and classroom materials necessary for implementing the TAM course (Algebra materials were of the district’s choosing). The TAM curriculum contains a sequence of lessons, and each lesson is tightly outlined, though not scripted. The 90 minute class period is broken into a consistent series of routines that include segments of teacher directed instruction, partner and small group interaction, and individual work and practice.

The curriculum was developed by teachers and others who have deep, firsthand knowledge of the conditions for teaching mathematics in urban high schools. The curriculum is tightly scripted because many experienced teachers do not have a good sense of how to make the best use of classes of 70-90 minutes, and because it is not uncommon for the most inexperienced teachers to be assigned to the ninth grade. In addition, TAM students receive consumable workbooks that contain all the problems and exercises and explorations they are asked to do, as well as explanatory text and examples. This is because copiers and copy paper are often in short supply in urban schools. Each TAM/Algebra 1 teacher receives a complete set of teaching materials, including a teacher manual, all of the materials needed for the course (ranging from paperclips to string to algebra tiles), and a class set of white boards for students, because teachers cannot assume that resources will be available to purchase classroom supplies.

The TAM curriculum was developed and field tested over a period of ten years prior to the start of this randomized field trial. The recognition for the need for such a “catch-up” curriculum came from the developers’ deep experience with struggling urban high schools and on-the-ground work with teachers at these schools who were charged with helping students to master algebra. Preliminary data, using less rigorous methods, had suggested that there might be a positive impact of the TAM curriculum on student achievement. For example, MDRC’s evaluation of the Talent Development program found impacts in 9th grade credit accumulation that were especially striking in Algebra. The impacts, in terms of differences in percent passing Algebra 1 between the Talent Development and demographically similar control schools, ranged between 17 and 30 percentage points. In Talent Development schools, the percentage of students passing Algebra 1 approximately doubled (to over 60%) between the baseline (pre-implementation) period and the implementation phase (Kemple, Herlihy, and Smith, 2005). A small-scale, non-randomized study in Baltimore found an effect size for TAM of .18; a similar comparison in Philadelphia showed an effect size of .52 (Balfanz, Legters, and Jordan, 2004).

Teachers implementing the TAM/Algebra 1 condition received several days of professional development on the curriculum, prior to the start of school. Teachers also met four additional times each semester to preview the upcoming course sub-units. The study provided a
classroom coach for the entire year, whose responsibility was to interact with each TAM/Algebra 1 teacher for the equivalent of two class periods per week.

Stretch Algebra teachers also received study-provided professional development (on curriculum mapping) at the beginning of the school year, as well as professional development on instructional strategies throughout the year, as requested by the district. The study did not support a coach for the Stretch Algebra condition.

Research Design
The overall research design is a multi-site cluster randomized trial. Within districts, schools were randomly assigned to implement either the Stretch Algebra or the TAM/Algebra 1 condition. Each district participates in the study for a single year, during which relevant data are collected.

Data Collection and Analysis
There are four key sources of data. First, student achievement data that provide outcomes for research questions #1 and #2, described above, were obtained from nationally normed achievement tests in mathematics. At the beginning of the school year, students took the CTBS Terra Nova test in mathematics (Level 19), as a pre-test of their knowledge of intermediate mathematics. The same test was given again in January, at the end of the first semester, to assess academic growth in intermediate math. To assess their Algebra knowledge, students take the CTB Algebra 1 assessment at the end of their freshman year. As a covariate control for prior Algebra knowledge, we use students’ scores on the Orleans-Hanna Algebra Prognosis Assessment, given at the beginning of the school year.

Second, teacher and student surveys given at the beginning and end of the school year provide important information about students’ motivation and experiences in math class, as well as teachers’ prior experience in teaching mathematics, their beliefs about mathematics, and opinions about the usability of the curricula.

Third, classroom observations conducted for each Stretch Algebra and TAM/Algebra 1 teacher twice during implementation (once during the fall semester and once in the spring). These observations provide quantitative and qualitative information about basic fidelity and instructional quality that allow us to conduct exploratory analysis of the context(s) in which one or the other of the curricula produced strong effects.

Fourth, student administrative records provide data on student demographics, prior achievement, and grades in ninth grade mathematics courses.

Findings / Results
We have conducted initial analyses of student academic growth in intermediate mathematics (fall to winter) for the first set of districts, which participated during 2008-2009. Our models indicate that there is a substantial effect (> .25 s.d.) of the TAM/Algebra 1 condition relative to the Stretch Algebra condition. While the size of the effect varies across districts, it is not driven by a single district. By spring 2010, when we present this paper, we expect to have the fall to winter growth results for all 12 districts in the study.
For the first set of districts, we fail to find a statistically significant difference in algebra achievement. This makes some sense, since students in the Stretch Algebra condition studied algebra over a period of eight months (theoretically interspersed with catch-up activities in intermediate mathematics), while those in the Transition Math course studied algebra for half that time. We have yet to conduct more analyses to understand when and where differences in algebra achievement existed between the conditions.

Conclusions

This study involved a large-scale randomized trial of a curriculum that had been developed and field tested over a period of ten years. Revisions were made based on teacher input about the organization of the curriculum and success of particular lessons and activities. The recognition for the need for such a “catch-up” curriculum came from the developers’ deep experience with struggling urban high schools and on-the-ground work with teachers at these schools who were charged with helping students to master algebra.

Our conclusions from initial analysis of the first year of data from the randomized trial are that students in the TAM/Algebra 1 sequence substantially outgained the Stretch Algebra students in intermediate mathematics and performed as well on a test of Algebra knowledge. Taking the perspective of whether any mathematics gains were made, one could argue that TAM/Algebra 1 students are in a better position than Stretch students to succeed in more advanced mathematics: their algebra knowledge is equivalent, and their foundational skills in math are on a firmer footing. The TAM/Algebra 1 students achieved as much in the domain of algebra after having four months of algebra instruction, in comparison to the eight or more months that the Stretch Algebra students had. However, if one takes the position that what is critical for future success in high school mathematics is solely student skill in Algebra 1, then the argument could be made that the Stretch Algebra and TAM/Algebra 1 sequences produce approximately equivalent results. We have yet to conduct more analyses to understand when and where differences in algebra achievement existed between the conditions.
Appendices

Appendix A. References


Title:

Evaluating Math Recovery: Implications for Policy and Practice

Author(s):

Thomas Smith (with Paul Cobb, Dale Farran, David Cordray, Charles Munter, Sarah Green, Annie Garrison, and Alfred Dunn), Vanderbilt University
Abstract Body

Background/context:

This presentation focuses on an initial evaluation study of Math Recovery (MR), a pullout, one-to-one tutoring program that has been designed to increase mathematics achievement among low-performing first graders, thereby closing the school-entry achievement gap and enabling participants to achieve at the level of their higher-performing peers in the regular mathematics classroom. Following Cordray and Morphy (2009), our goal was not merely to assess whether MR works, for whom, under what circumstances. We also attempted to understand how and why the program works to produce particular outcomes (cf. Clements, 2007). In addition, we illustrate that assessments of implementation fidelity can help identify aspects of an intervention that need improving. Assessments of implementation fidelity in turn require that the evaluation begins with a “well-stated set of expectations about how the intervention is supposed to work, its underlying logic, and rationales for how and why these actions will produce the desired enhancements in student learning, motivation, and achievement” (Hulleman & Cordray, 2009, p. 90).

The rationale for the evaluation study is grounded in the well-documented finding that children enter school at a wide range of mathematical abilities (Baroody, 1987; Dowker, 1995; Gray, 1997; Griffin & Case, 1999; Housasart, 2001; Wright, 1991, 1994a; Young-Loverage, 1989). A study conducted by Aunola, Leskinen, and Lerkkanen (2004) found that, in the absence of intervention, the initial gap in mathematics achievement continues to widen. Furthermore, Duncan, Claessens, and Engel’s (2004) analysis of ECLS-K indicated that pre-K mathematical ability is highly predictive of achievement at the end of first grade, and Princiotta, Flanagan, and Germino Hausken’s (2006) analysis of the same data set revealed that achievement gaps are still prevalent in fifth grade. They found that 67% of students who scored in the top third in their kindergarten year did so again six years later, and that those among the lowest third in 1998 generally scored low in 2004. Taken together, these findings emphasize that identifying effective methods for closing the pre-K gap is an pressing policy concern (McWayne, Fantuzzo, & McDermott, 2004). Using an experimental design, we assessed the effectiveness of MR in improving mathematics achievement of low performing first grade students and examined whether gains made in first grade were maintained through the end of second grade.

As we clarify below, MR tutoring is a demanding form of practice in which tutors are expected to adjust instruction to the current level of a student’s thinking at any given point in time. The teacher development literature suggests that teachers can learn in the context of their practice, often as they attempt to understand students’ reasoning and adjust instruction accordingly (Franke, Carpenter, Fennema, Ansell, & Behrend, 1998). In other words, the effectiveness of MR tutoring may improve as the tutors gain additional experience with the intervention. Thus, although it might be important to select tutors based on their initial knowledge and skills, it is also essential to consider the extent to which tutors can develop the necessary knowledge as they enact MR. This necessary knowledge includes mathematical knowledge for teaching (MKT) (Hill, Rowan, & Ball, 2005). MKT denotes a form of mathematical knowledge that is specific to problems and decisions that arise in the practice of teaching. A priori, MKT appeared to be central to MR because tutors are expected to assess and build on students’ current reasoning. A
second aspect of the necessary knowledge concerns tutors’ knowledge of the MR Learning and Instructional Frameworks in Number (LFIN and IFIN, respectively). A primary goal of MR training is to enable tutors to understand these frameworks, and to use them in their tutoring practice. The frameworks lay out developmental trajectories for students in early number learning and suggest instructional activities to support students at various points along those trajectories. A tutor’s ability to understand and use the frameworks is therefore integral to effective implementation of MR. Thus we examine both the impact of tutors initial MKT, LFIN, and IFN knowledge on their tutoring effectiveness and whether increases in tutors’ knowledge in these areas over the course of the study is associated with increasing effectiveness in their tutoring.

It is widely acknowledged that claims of treatment effectiveness may be unjustified and invalid unless the degree to which programs are implemented as intended is defined and assessed (Dusenbury, 2003; O’Donnell, 2008). However, little is known regarding the feasibility of assessing the implementation fidelity of unscripted interventions such as MR, where measuring fidelity requires the identification and operationalization of complex, often implicit facets of the intervention (Cordray & Pion, 2006). As part of the study, we measured the implementation fidelity of MR tutoring and will eventually link the measures to student outcomes.

**Purpose / objective / research question / focus of study:**

Our research questions were as follows:
1. Does participation in MR raise the mathematics achievement of low performing first-grade students?
2. If so, do participating students maintain the gains made in first grade through the end of second grade?
3. Do initial differences in tutor knowledge (both MKT and knowledge of MR frameworks) persist as tutors gain experience with MR and learn through practice?
4. To what extent does fidelity of implementation influence the effectiveness of MR?

**Setting:**

The two-year evaluation of Math Recovery was conducted in 20 elementary schools (five urban, ten suburban, and five rural) from five districts in two states. Each was a ‘fresh site’ in that the program was implemented for the first time for the purposes of the study.

**Population / Participants / Subjects:**

Students were selected for participation at the start of first grade based on their performance on MR’s screening interview and follow-up assessment interview. The screening is designed to select the lowest achieving first graders (25th percentile and below) in terms of math achievement. The number of students eligible for tutoring ranged from 17 to 36 across the 20 schools. The number of study participants before attrition totaled 517 in Year 1 and 510 in Year 2, of which 172 received tutoring in Year 1 and 171 received tutoring in Year 2.

We recruited 18 teachers to receive training and participate as MR tutors. Sixteen of the tutors received half-time teaching releases to serve one school each; two of the tutors received full-time
teaching releases to serve two schools each.

**Intervention / Program / Practice:**

MR consists of three components: 1) tutor training, 2) student identification and assessment, and 3) one-to-one tutoring. The first component of the MR program, tutor training, involves 60 hours of instruction provided by an MR leader. The goal of this training is to support tutors in learning new practices for clinical assessment and intervention teaching in which they use the MR Learning Framework and the Instructional Framework to adjust instruction based on cognitive evaluations of student responses. The tutors participating in the study were trained by MR personnel at two sites in different states.

In the second component of the program, the tutor conducts an extensive video-recorded assessment interview with each child identified as eligible for the program. The tutor analyzes these video-recordings to develop a detailed profile of each child’s knowledge of the central aspects of arithmetic using the MR Learning Framework, which provides information about student responses in terms of levels of sophistication.

The third component of the program, one-to-one tutoring, is diagnostic in nature and focuses instruction at the current limits of each child’s arithmetical reasoning. Each selected child receives 4 or 5 one-to-one tutoring sessions of 30 minutes each week for approximately 11 weeks. The tutor’s selection of tasks for sessions with a particular child is initially informed by the assessment interview and then by ongoing assessments based on the student’s responses to prior instructional tasks. The Learning Framework that the tutor uses to analyze student performance is linked to the MR Instructional Framework that describes a range of instructional tasks organized by the level of sophistication of the students' reasoning together with detailed guidance for the tutor.

**Research Design:**

The structure of the MR program allowed us to use the fact that two thirds of the participating students have their treatment delayed by either 11 or 22 weeks to establish an experimentally assigned control group for each cohort of participants consisting of both students whose treatment has not yet begun and a small number of students who are on a “wait list” for treatment. By randomly assigning the students selected for participation in the study each year to one of the three treatment cohorts or the wait list, we could establish the essential characteristics of an experimental design.

To study teacher knowledge and learning, we assessed tutors at three time points with a measure of their MKT developed by Ball, Hill and colleagues (e.g., Hill, Ball, & Schilling, 2008), and a measure of their knowledge of the MR Learning and Instructional Frameworks constructed in collaboration with the developers of the MR program.

As part of their standard practice, MR tutors video-record all tutoring sessions in order to plan for subsequent sessions. In the presentation, we will describe both the iterative process by which we developed an instrument for assessing the fidelity of implementation of MR tutoring, and the
process by which we validated the instrument by comparing ratings on a subset of the video-recorded tutoring sessions with the assessments of 30 MR experts. We will also describe how we trained coders until agreement reached an adequate level (80%). The coders are currently coding a randomly selected 20% of the video-recorded tutoring sessions.

**Data Collection and Analysis:**

Each of the students participating in the study were assessed using alternating forms of the Applied Problems, Quantitative Concepts, and Fluency subtests of the Woodcock Johnson III Achievement (WJ III) subtests, as well as the MR proximal instrument, an assessment based on the learning framework that we designed in consultation with the program developers, at the start of the study and when each cohort entered or exited tutoring in December, March, and May. Wait list students took the Fluency subtest of the WJ III at the same time as each cohort entering treatment, as well as the full battery of other WJ III and MR proximal assessments at the start and end of the school year.

Our research design allowed us to describe and compare the growth trajectories of treatment and control cohorts across the whole school year, punctuated at the end of each 11-week period by the students completing MR tutoring. To estimate these growth trajectories, we used 3-level hierarchical linear growth models (Raudenbush and Bryk, 2002; Singer and Willett, 2002) with repeated observations of WJ III scores or MR proximal scores indexed by time, time since starting MR, and time since completing MR at level 1, student level demographics at level 2 (e.g., gender, minority status), and school characteristics at level 3. To assess whether gains made in MR tutoring are maintained after the tutoring is completed, a time varying covariate that counts the number of days after a student completes MR. Although the results presented here are only for the first year cohort in this study, the paper presented at SREE will include end of second grade data for Cohort 1 and end of first grade data for Cohort 2. We are particularly interested in testing the hypothesis that the gains made from participation in MR are maintained through the end of second grade.

The tutors were assessed using the externally validated LMT assessment to measure their MKT and an internally designed Tutor Knowledge Assessment (TKA) to assess their knowledge of the MR frameworks. The assessments were given at end of MR training, end of year 1, and end of year 2. The analysis of these data used a one-way analysis of variance (ANOVA) where the predictor was training site (as noted above, there were two training sites). ANOVA was used to test for a difference between means of the two groups on both the TKA and the LMT at time 1 and at time 3.

**Findings / Results:**

The first year results show a small to moderate effect of participation in MR on WJ III scores and moderate to large effects on the MR proximal assessments. Differences in the end of first grade mean scores on the WJ III subtests between students selected for tutoring and those on the waitlist ranged in effect size from .21 on the quantitative concepts scale to .28 on the applied problems scale (all differences statistically significant at the p<.05 level). Effect sizes on the MR proximal measures ranged from .34 on the forward number sequence scale to .92 on the
arithmetic strategies measure. These results compare favorably to those reviewed recently by Slavin and Lake (2006), including several cooperative learning programs that had median effect sizes of at least +0.30). However, a meta-analysis of 52 studies on the relationship between tutoring and student achievement (Cohen, Kulik, and Kulik, 1982) found average effect sizes greater than .40—higher than MR effects on the WJ III measures but lower than effects on some of the more proximal assessments. Results from the growth models show increases in mathematics achievement for MR participants across all assessments during the tutoring period (with p<.05 in each case), although this growth rate tends not to be maintained after completion of MR.

With regard to the tutors, there is a significant difference between the two training sites at time 1 on the LMT (F=7.81, p = 0.013) and on the TKA (F=15.18, p=0.0013), indicating that the tutors at one site were initially more knowledgeable in their MKT and also learned more about the MR frameworks from the initial MR training. However, at the end of the study there were no significant differences between these groups on either measure (LMT: F=3.55, p=0.08 & TKA: F= 1.36, p=0.26). This lack of difference between groups at the end of the study was due to a steeper increase in knowledge at the second site. We will present data on the relationship between tutor LMT and TKA scores during the presentation.

Conclusions:

The positive causal effect of MR tutoring demonstrates that programs that are diagnostic rather than scripted in nature can overcome fidelity concerns and have an impact on student early mathematics performance. Our findings therefore indicate that investing in tutors' knowledge of student reasoning and pedagogical content knowledge can pay off in terms of improvement in student's mathematical learning, particularly if tutors use carefully designed tools such as the MR Learning and Instructional Frameworks. With regard to policy, our finding that the MR program can reduce some of the pre-K mathematics achievement gap provides an initial indication that the cost of the program per student might be justified, although further work is needed to understand why initial gains made by participants appear to diminish after tutoring ends. It is possible that the forms of arithmetic reasoning that MR develops needs to be further supported in the regular classroom to see the full benefit of this form of tutoring.

The results concerning the tutors’ mathematical knowledge and knowledge of the MR Frameworks have two implications for policy and for future studies of this intervention. First, tutors who had higher MKT at the outset also had higher scores on the TKA, suggesting that tutors with more math knowledge for teaching may learn more from the initial MR training, potentially making them better choices for tutoring early on. Second, the initial differences did not persist between groups after two years of tutoring experience, indicating that tutors can and do grow in their understanding of the MR frameworks and also in their math knowledge for teaching through their MR tutoring practice. As a consequence, any limitations in the MR program indicated by the second-year findings cannot be attributed to tutors’ lack of knowledge of the Frameworks. An implication for policy and adoption of MR is that while initially tutors might struggle to learn the MR Frameworks, their knowledge of the both the Frameworks and their MKT will likely improve with experience with the program.
The findings from our fidelity study suggest that it is possible to create a reliable instrument to measure implementation fidelity for differentiated interventions, an endeavor that has typically been largely avoided in evaluations of educational interventions. Many potentially high-quality interventions are un-scripted and instead rely on teacher knowledge and professional development. Because the fidelity measures are reliable and true to program components, we will be able to link measures of treatment integrity to outcomes, further clarifying how and why MR to produce particular outcomes (Cordray & Pion, 2006). Critical aspects of the process included 1) the identification of the core implementation components of the intervention (Fixsen et al., 2005); 2) operationalization of those components; 3) training of coders in both the program itself and the coding schemes; and 4) collaborating with the coding team to further refine coding decisions.
Appendices
Not included in page count.

Appendix A. References
References are to be in APA version 6 format.


Appendix B. Tables and Figures

Not included in page count.