Full Reference: Doabler, C. T., Clarke, B., Fien, H., Baker, S. K., Kosty, D. B., & Strand Cary, M. (2015). The science behind curriculum development and evaluation: Taking a design science approach in the production of a tier 2 mathematics curriculum. *Learning Disability Quarterly, 38*(2), 97–111. https://doi.org/10.1177/0731948713520555

The Science Behind Curriculum Development and Evaluation: Taking a Design Science Approach in the Production of a Tier 2 Mathematics Curriculum

Authors:

Christian T. Doabler, Ph.D.

Ben Clarke, Ph.D.

Hank Fien, Ph.D.

Scott K. Baker, Ph.D.

Mary Strand Cary, Ph.D.

Derek B. Kosty, B.S.

Funding Acknowledgements
The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education through grant R324A090341 to the Center on Teaching and Learning at the University of Oregon.

Abstract

The production of an effective mathematics curriculum begins with a scientific development, evaluation, and revision framework. The purpose of this study was to conduct an initial investigation of a recently developed Tier 2 mathematics curriculum designed to improve the outcomes of 1st grade students at-risk for mathematics difficulties. The curriculum, which is anchored to a scientific design framework and a well-specified theory of change, centers on the careful integration of foundational concepts of whole number and validated-design principles of explicit and systematic instruction. Four instructional groups, with each consisting of five 1st grade students with mathematics difficulties and one interventionist, participated in the study. Data related to the feasibility, usability, and promise of the curriculum to improve student mathematics achievement were collected using multiple methods. Results provide preliminary evidence in terms of these curriculum development and evaluation areas. Implications for instruction and design research are discussed.

The Science Behind Curriculum Development and Evaluation: Taking a Design Science Approach in the Production of a Tier 2 Mathematics Curriculum

Developing mathematical proficiency is absolutely critical for success in school and in postsecondary experiences (National Mathematics Advisory Panel [NMAP], 2008; National Research Council [NRC], 2001). For students who acquire early success in mathematics, the stage is set for learning more complex mathematical content. Conversely, for students who struggle early with mathematics, the probability that they will become mathematically proficient remains low given that the trajectories of mathematics achievement become established early and typically remain difficult to change (Bodovski & Farkas, 2007; Locuniak & Jordan, 2008; Morgan, Farkas, & Wu, 2009). Unfortunately, the long-term consequences of early mathematics difficulties (MD) are severe. Research now suggests that students who struggle to develop mathematical proficiency in the early grades are far more likely than other students to experience persistent difficulties in later mathematics (Bodovski & Farkas, 2007; Morgan et al., 2009).

One contributing factor in supporting all students, including those at risk for MD, in reaching their mathematical potential is the elementary mathematics curriculum. According to Wu (2009), “The main goal of the elementary mathematics curriculum is to provide children with a good foundation for mathematics” (p. 6). Toward that end, the elementary mathematics curriculum sets the curricular expectations of children’s mathematics learning. It influences when and how well children progress through content. Importantly, a coherent curriculum reflects the hierarchy of mathematics by forming a coherent connection between foundational concepts and skills both within and across grade levels (Schmidt, Houang, & Cogan, 2002).

A Scientific Approach to the Development of Mathematics Curricula

Critical to the role that core (Tier 1) and intervention (Tier 2) curricula have in supporting struggling students' in developing mathematics proficiency is the process in which they are developed. A recommendation by experts is for developers to take a scientific approach to curriculum development (Chard et al., 2008; Clements, 2007; Superfine, Kelso, & Beal, 2009). Despite this call, there is scant evidence in the field of curriculum development and evaluation of a scientific imprint. For example, in the area of early mathematics, the What Works Clearinghouse (WWC, 2013) has reviewed a total of 78 elementary school programs. Of these 78 programs, only seven (i.e., < 10% of elementary mathematics programs reviewed) had research studies that meet WWC evidence standards required for evaluating program effectiveness or ineffectiveness. Of the seven reviewable programs, only four demonstrated "potentially positive" effects on student mathematics outcomes.

Curriculum development research, like other design-based research (Diamond & Powell, 2011), centers on a systematic and empirically-based development process. Thus, the operational definition of a scientific design process applied in this study focuses on the accumulation of empirical evidence related to a curriculum's feasibility, usability, and promise to improve student mathematics achievement. A feasible and usable curriculum functions as a practical instructional tool for teachers in authentic educational settings. A curriculum that demonstrates promise has an initial evidentiary base suggesting that it preliminarily improves student achievement. In this study, we use these types of data (i.e., feasibility, usability, and promise) as a formative influence on the development and evaluation of a Tier 2 mathematics curriculum designed to target whole number concepts and skills.

We argue that a scientific design process is essential for addressing the complex challenges faced in the development of a Tier 2 mathematics curriculum. One inherent challenge in developing a coherent Tier 2 mathematics curriculum is being mindful of the instructional needs of students struggling with mathematics. In this case, developers must devote additional attention to the architectural features of the curriculum (Kame'enui & Simmons, 1999). This means embedding validated-design principles of instruction, such as explicit modeling of mathematical concepts and procedures (Coyne, Kame'enui & Carnine, 2011), to support teachers in delivering critical mathematics content. Another challenge is allowing for extensive flexibility during the development process. Immediate and ongoing refinements to lesson components are critical for the curriculum development process and a recommendation is for these adjustments to take place before, during, and after field tests (Anderson & Shattuck, 2012). Other challenges include ensuring that there is acceptability of the curriculum's goals, procedures and outcomes. As observed by Lindo and Elleman (2010), "An intervention's sustainability depends not only on how effective it is for students but also on how well it fits into the classroom context and how it is perceived by those involved" (p. 490).

Several frameworks of curriculum development have been proposed to guide developers in designing curriculum (Anderson & Shattuck, 2012; McKenney & Reeves, 2013). Clements' (2007) Curriculum Research Framework (CRF) suggests that curriculum development is a "design science." Clements observed: "As a science, knowledge created during curriculum development should be both generated and placed within a scientific research corpus, peer reviewed, and published" (p. 37). Against this backdrop, the CRF focuses on cycles of development, formative evaluation, and curriculum revision, and comprises three categories of development activities: a *priori foundations, learning models*, and *evaluation*. Embedded within the three categories, Clements proposes 10 phases of the curriculum development and evaluation process. In Phases 1-3 (a priori foundations), developers complete a comprehensive review of the supporting literature relevant to the instructional practices and mathematical content for improving the mathematics achievement of students. A goal of the literature review is to guide developers in specifying the instructional design attributes that underlie the curriculum and its theory of change (Simmons et al., 2007). For example, given the preponderance of empirical evidence for the use of explicit instruction when teaching students with MD (Gersten et al., 2009; NMAP, 2008), developers of a Tier 2 curriculum would likely include highly specified instructional procedures and teaching examples.

Phase 4 (learning models) of the CRF (Clements, 2007) has developers design prototype activities that actively engage students in learning critical concepts and skills of mathematics. Once the prototypes are designed, developers begin the formative evaluation cycle, which includes Phases 5-10 (evaluation) of the CRF. Here, developers test the prototype activities in classrooms and collect evidence on features such as task difficulty, the selection and sequencing of instructional examples, and the number of opportunities for practice and review. The CRF recommends that developers use this evidence to refine and revise the prototypes. Eventually, the curriculum is compiled into a complete package and field tested to produce critical, quantitative pilot data, such as evidence related to a curriculum's feasibility, usability, and promise to improve student outcomes. As with the current study, such initial data sources should position the curriculum for more rigorous evaluation in future research studies. In this context, the purpose of this study is to conduct a preliminary investigation of Fusion, a Tier 2 mathematics curriculum designed to support at-risk 1st graders in developing a fundamental understanding of whole number concepts and skills. A critical component of the Fusion intervention is a well-specified theory of change.

Theory of Change and Its Role in Curriculum Development

In recent years, primary funding sources for the development of rigorous curricula have been the U.S. National Science Foundation (NSF) and Development and Innovation Projects offered through the Institute of Education Sciences (IES, 2013), U.S. Department of Education (USDE). According to recent guidelines proposed by these two funding agencies (USDE, 2013), Design and Development projects should include: (a) the full development of an intervention or product, (b) demonstration of the intervention’s feasibility and usability for end-users, and (c) pilot data for the promise of the intervention to improve intended student outcomes. Another outcome recommended in these co-agency guidelines is to generate a “well-specified theory of action” (USDE, 2013; p. 48). A theory of action or change represents the underlying core elements of a mathematics curriculum that, when facilitated well, are hypothesized to produce important outcomes for students. That is, a theory of change is a theory-driven and empirically-based model of how an intervention is assumed to lead to improved student outcomes on proximal and distal measures of mathematics achievement. While Design and Development projects are not an appropriate venue for rigorously testing a theory of change, they are optimal for generating preliminary evidence that supports an intervention’s theoretical foundation.

Tenets of a Strong Theory of Change

A strong theory of change demonstrates several key tenets and collectively they form the architectural structure of an intervention. One tenet is its link to strong theory and the empirical knowledge base of effective instruction. For example, if researchers were developing a 1st grade Tier 2 curriculum, they would draw from several relevant literatures. These literatures would include findings from recent meta-analyses targeting interventions for students with or at-risk for MD and theoretical frameworks on how children learn mathematics.

Theory-driven models are also characterized by proximal and distal student outcomes that the intervention purports to impact. For instance, researchers might hypothesize that an intended outcome of an early mathematics intervention is to improve students' immediate and long-term mathematics achievement. Further, they might theorize that students' proximal outcomes will have a positive and direct causal relation with general measures of student achievement. To measure the intervention's proximal impact, they would need to identify or design a measure that is closely aligned to the learning objectives of the intervention curriculum. Additionally, the researchers would need to select a set of standardized, distal measures of student mathematics achievement to assess gains in student mathematics outcomes.

Other key tenets of strong theories of change are variables hypothesized to mediate and moderate the impact of an intervention. According to Rothman (2013), "Mediators and moderators are the building blocks of theory and, in turn, intervention design, specifying the connections between these two classes of constructs is at the heart of developing, testing, and refining theory" (p. 190). Mediating variables refer to the active ingredients that comprise and guide an intervention (Rothman, 2013). For example, researchers might hypothesize that the instructional interactions facilitated by a Tier 2 curriculum will mediate student outcomes on a measure of early number sense. Moderating variables refer to student and teacher factors that may change the relationship between an intervention and student outcomes. For example, the impact of an intervention mathematics curriculum may vary by students' risk status.

The Theory of Change Behind the Fusion Intervention

We believe that the theory of change supporting the Fusion intervention is characterized by these same key tenets. As shown in Figure 1, the Fusion theory of change contains: (a) the curriculum, (b) Mediator Variables, and (c) Proximal and Distal Student Outcomes. The curriculum is comprised of two key components: (a) critical whole number content and (b) validated-design principles of explicit and systematic instruction. These components, when integrated well, facilitate high-quality instructional interactions between teachers and students around critical content of whole numbers. Such instructional interactions are hypothesized to mediate the impact of Fusion on proximal and distal student outcomes. Proximal outcomes of the intervention include students' conceptual understanding and procedural fluency, whereas the distal outcome is students' overall mathematics achievement.

< Add Figure 1 >

There are several reasons why Fusion is hypothesized to produce promising outcomes for 1st grade students struggling with mathematics. One reason is that it was developed through a scientific design process similar to the one proposed by Clements (2007). Our curriculum team collected a variety of qualitative and quantitative data, such as professional feedback from teachers, to shape and refine Fusion. Another reason why Fusion may improve student mathematics achievement is that it focuses intensely on the critical aspects of whole number, and thus is responsive to the calls of experts for greater instructional focus on number (NRC, 2001). Finally, Fusion may be beneficial for at-risk learners because it is based on sound theoretical models of learning (i.e., Donovan & Bransford, 2005) and the growing knowledge base of effective mathematics instruction (e.g., Gersten et al., 2009; NMAP, 2008).

At its core, the curriculum uses an *explicit* instructional approach, which is considered to be one of the most effective methods for teaching mathematics to at-risk learners (Baker, Gersten, & Lee, 2002; Gersten et al., 2009; NMAP, 2008). Explicit instruction is a method for teaching "essential skills in the most effective and efficient manner possible" (Carnine, Silbert, Kame'enui, & Tarver, 2004, p. 5). Instruction in Fusion becomes explicit by orchestrating important instructional interactions between teachers and students around key mathematics content. For example, Fusion's explicit instructional approach allows teachers to (a) provide clear explanations and vivid demonstrations when introducing new mathematical concepts and skills, and (b) facilitate frequent opportunities for students to practice with such content.

In summary, the quality of a mathematics curriculum greatly depends on the processes in which guide its development and evaluation. A recommendation by the research community is for developers to take a scientifically-based approach. Conceptually, this approach is anticipated to generate preliminary evidence in support of the curriculum and its theory of change.

Purpose of the Study

The primary purpose of this study was to build preliminary support for Fusion and its theory of change by generating empirical evidence related to two specific areas of curriculum development: (a) the feasibility and usability of Fusion by intended end users in authentic education settings, and (b) Fusion's promise to improve the mathematics outcomes of students at risk for MD. Our study targeted these two areas because they (a) align with the recommendations of experts in the field of curriculum development (Clements, 2007) and (b) address research and development guidelines co-proposed by IES and NSF (USDE, 2013).

In the aggregate, this body of evidence was anticipated to advance the field of curriculum development forward in two ways. First, our curriculum development team used a "design science" (Clements, 2007) approach to systematically develop, and formatively refine and evaluate Fusion. As noted, findings from the WWC (2013) suggest that the development and evaluation of mathematics curricula, particularly in the area of elementary mathematics, is not occurring in a scientific manner. *The dearth of these findings reveal an urgent need to supply the field with soundly engineered, evidence-based mathematics curricula for students struggling with early mathematics*. We believe this study is an initial step in addressing that call.

The second way this study aims to add to the knowledge base is through its selected unit of analysis; i.e., the instructional group. Small group instruction is an integral component in multi-tiered approaches to instruction, particularly in Response to Intervention (RtI) models that employ standard protocol approaches (Fien et al., 2011; Fuchs, Fuchs, & Stecker, 2010). By selecting this unit of analysis, therefore, we have direct links to the practical considerations that arise in making instructional decisions in group-based interventions, such as Fusion. A recommendation in these RtI models is to take into consideration group level data when determining the instructional effectiveness of these types of interventions (Fien et al., 2011).

Two research questions guided the current study: (1) Does the Fusion curriculum demonstrate evidence of feasibility and usability for interventionists and students in authentic education settings? (2) Is there a functional relation between Fusion and improvements in mathematics outcomes for students at-risk for MD?

Method

Data for this study were collected within the second year of an IES-funded Development and Innovation Project. Such development projects are typically funded for three years, with the first two years devoted toward major development activities and the final year used for pilot testing the intervention's promise for improving student outcomes. This study, which took place during the 2010-2011 academic year, was anticipated to generate initial data that would guide future design research on Fusion and its theory of change. Specifically, the study targeted Fusion's feasibility, usability, and promise to improve student mathematical performances. To address our two research questions, data were derived from multiple methods, including surveys, direct observations, student performance assessments, and a single case research design (SCD).

Setting and Participants

The study took place in three elementary schools from two suburban school districts located in the Pacific Northwest. District-1 enrolls approximately 10,850 students: 17.4% receive special education services, 5.9% are English learners, 59.8% are eligible for free or reduced lunch, and 25.5% are minorities. District-2 enrolls approximately 5,800 students: 19.6% receive special education services, 3.1% are English learners, 57% are eligible for free or reduced lunch, and 24.1% are minorities. All three schools received Title 1 funding and provide, on average, three classrooms at the 1st grade level.

Four instructional groups, referred hereafter as Groups A, B, C, and D, participated in the study, and each group represented one 1st grade classroom. Groups A and B were located in the same school and instruction for these two groups took place in same educational setting, a special education classroom. Group D was located in a different school from Groups A and B, but was from the same school district, District-1. Instruction for Group D took place in a special education classroom. Group C took place in a special education classroom located in an elementary school from District-2. Participating interventionists included two instructional assistants (Groups A and B) and two special education teachers (Groups C and D). The four interventionists had an average of 2.5 years of experience in working with students with or at-risk for learning disabilities. One interventionist was male and three were White.

Twenty students from four 1st grade classrooms participated in the study. These 20 students comprised the four Fusion groups, with five students per group. Of the 20 participating students, 10 were females, 18 were White, and 2 were Native American. Students were between the ages of 6 and 7 years of age. Three of the participating students received special education instruction and fourteen qualified for free or reduced price lunch programs. Across the 20 participants, school attendance during the 2010-2011 academic year was relatively high at 92%.

In all, 16 students, representing 4 groups, successfully completed the 20-week Fusion intervention. Data from these 16 students were included for analysis. One student from Group A moved two weeks into the intervention. The Group C interventionist withdrew two students from the study during baseline due to scheduling conflicts with other school activities. A third student from Group C moved out of the study, approximately 11 weeks into the intervention. Data from these four students were excluded from the analysis of all outcome measures.

The selection of students involved a multistep process that began with universal screening. All students in participating 1st grade classrooms were administered two curriculum-based measures (CBM) of early mathematics: missing number (MN) and quantity discrimination (QD) measures. Each screening measure, described in greater detail below, focus on a discrete skill associated with early number proficiency for 1st grade students. From each classroom, five students with the lowest performances on the two CBMs were determined eligible for the intervention. Compared to the full sample of 1st grade students assessed in the screening process ($N = \sim 100$), the 20 Fusion students performed in the low average to well-below average range on both screening measures.

Names and scores of the selected students were then given to their respective general education teachers for verification. Teachers were responsible for providing input on the appropriateness of the student selections. They considered, for example, each student's response to core mathematics instruction and whether the student might benefit from a Tier 2 mathematics intervention. Students verified by classroom teachers were then considered eligible for the current study. No discrepancies between the screening data and teachers were found.

Feasibility and Usability

To assess feasibility and usability of the Fusion intervention, interventionists completed a five-item survey at the second curriculum workshop, which marked the halfway point of the intervention. The purpose of this survey was to determine interventionists' perceptions about implementing Fusion and the academic growth their students had made through the first half of the curriculum. At the conclusion of the intervention, interventionists completed a second survey, rating six aspects of implementation: (a) the level of difficulty implementing Fusion, (b) the amount of teacher scripting provided, (c) the amount of activities completed, (d) the usability of Fusion's materials, and (e) the degree to which Fusion students benefited both mathematically and behaviorally. Items included in both surveys had varying rating scales.

Student Outcome Measures

Proximal outcome measure. To assess proximal outcomes in mathematics, a researcher-developed instrument, ProFusion, was administered at pretest and posttest time periods. ProFusion is a group-administered assessment that targets three areas of number and numeration: (a) place value concepts, (b) basic number combinations, and (c) problems involving multi-digit addition and subtraction. In an untimed setting, ProFusion has students decompose two-digit numbers (3 items), and write numbers from dictation (4 items), numbers missing from a sequence (3 items), and numbers matching base-ten models (3 items). Students also complete one-minute, timed addition (32 items possible) and subtraction (24 items possible) number combinations, and multi-digit addition and subtraction problems (each with 4 items possible). Inter-correlations between the three ProFusion areas ranged from .40 to .82, with an average inter-correlation of .65. Clarke et al. (in press) report ProFusion to have adequate predictive validity with CBMs of early mathematics (r = .56) and the SAT-10 (r = .68). In this study, a paired samples t test was conducted to estimate growth on ProFusion across the 20-week intervention (i.e., pretest to posttest gains). Because two students moved just prior to the posttest administration of ProFusion, only 14 of the 16 students completed both the pretest and posttest assessments. Data from these 14 students are included in the *t*-test analysis.

Distal outcome measures. Two CBMs of early mathematics were used as outcome measures and included for visual analysis in the SCD: Missing Number (MN) and Quantity Discrimination (QD; Clarke & Shinn, 2004). Both MN and QD are 1-minute, fluency-based measures and considered as distal outcome measures to the Fusion intervention[1]. MN and QD assess, respectively, two important aspects of early numeracy development: strategic counting and magnitude comparison. In the current study, we adapt the QD and MN measures to be administered in small-group instructional settings. Three alternate forms of the QD and MN measures were developed and changes made to the measures were based exclusively on student response formats. Originally, the measures required students to verbally respond to each item. The measures employed in the current study, however, had students record their responses through written format. The modified QD measure, this study's primary outcome measure, requires students to circle the number in a pair (numbers 0 to 10) with the higher value. The modified MN measure requires students to write in the missing number among a string of three numbers (0-10), with the first, middle, or last number of the string missing (e.g., 5 ___ 7).

Test-retest reliability for the MN and QD measures was, respectively, .85 ($p < .001$) and .87 ($p = .003$) between the last two intervention-phase assessments. As a measure of concurrent

---

[1] Skills assessed by the MN and QD measures were indirectly and infrequently taught in Fusion. Therefore, we argue that these outcomes measures are not "overaligned" with the Fusion intervention. For example, Fusion addresses the skill of identifying a missing number in a string of numbers through an activity called the Missing Number Game. The game requires students to put a series of number cards in the correct numerical sequence (e.g., 11-15). Once the cards are placed in the correct order, the interventionist has the students close their eyes and then she removes one card from the sequence. Students then open their eyes and raise their hands when they recognize the missing number. This game appeared in just 12 lessons during the first half of the curriculum. For the skill of making magnitude comparisons, students select a number card between 0-9 and then identify a number that is more or less than the number selected. These magnitude comparison activities appeared in the first five lessons of Fusion only.

validity at the beginning of the baseline phase, ProFusion was moderately correlated with MN ($r = .57$, $p = .022$) and QD ($r = .58$, $p = .019$). Concurrent validity coefficients at the end of the intervention phase suggest ProFusion was moderately correlated with MN ($r = .42$, $p = .174$) and highly correlated with QD ($r = .80$, $p = .059$). We also computed predictive validity coefficients and found that ProFusion scores at posttest were correlated with pretest MN ($r = .45$, $p = .102$) and pretest QD scores ($r = .67$, $p = .007$).

Data Collection Procedures

Data collection of MN and QD during baseline varied between two to three times per week for participating groups. Across the four groups, the number of baseline probe administrations for QD and MN ranged from five to eight. During the intervention phase, Fusion interventionists administered the QD and MN measures once per week, immediately following the conclusion of the week's second lesson. Across the four groups, the number of probes for QD and MN ranged from 10 to 26. Two project members, blind to the study's hypotheses, separately scored all outcome measures (i.e., MN, QD, and ProFusion). The two staff members then separately entered the raw scores into an Excel spreadsheet for data analysis. In instances where scores were discrepant between the two scorers, an independent third scorer rescored the measure and provided the arbitration score. Scores for MN, QD, and ProFusion represented the total number of problems answered correctly. Group averages for MN and QD were calculated for each administration and depicted for visual analysis in the single-case design.

Experimental Design and Procedures

Among the different methods employed in this study was a multiple-probe-across-groups design (Gast, 2009) used to explore initial evidence of Fusion's impact on the mathematics achievement of students at-risk for MD. Multiple probe designs can help reduce testing effects, a major threat to a study's internal validity, by providing intermittent pre-intervention data collection (Gast, 2009). However, the benefits of a multiple probe design can be offset by the potential for carryover effects, sometimes referred to as contamination or treatment diffusion (Shadish, Cooke & Campbell, 2002). This threat was controlled for in the current study by having participating students remain in their general education classrooms until their group started Fusion. The unit of analysis for this study was instructional groups.

Baseline. During the baseline and intervention phases, all participating students continued to receive "business-as-usual" district-approved core mathematics instruction in their general education classroom. Core programs used in the participating 1st grade classrooms included *Saxon Math, Everyday Mathematics,* and *Math Expressions*. Interventionists administered baseline assessments (i.e., QD and MN) to groups of students outside of the general education classroom and at a time that was essentially scheduled for the Fusion intervention. Generally, for multiple probe designs, intervention introduction occurs with the presence of a stable baseline. Here, baseline stability is defined as three consecutive data points (i.e., group averages) on the primary outcome measure, QD, that lack significant variability or substantive trend towards the hypothesized direction of change for the Fusion intervention (i.e., increases in group averages). Introduction of Fusion, which was contingent on stability of the QD measure, was staggered across instructional groups, with Group A starting the intervention first. It is important to note that a scheduling conflict in Group C's school caused the introduction of Fusion to Groups B and C to be separated by only one week. However, because Groups B and C occurred in different schools threats to internal validity (i.e., diffusion) were not tenable.

Fusion Intervention. The Fusion curriculum is a Grade 1 (Tier 2) mathematics intervention designed to build students' early knowledge of whole number concepts and skills identified in the Common Core State Standards for 1st grade mathematics (CCSS, 2010). Specifically, the curriculum targets content standards from two mathematical domains in the CCSS: (a) Operations and Algebraic Thinking, and (b) Number and Operations in Base Ten. Fusion contains 60 scripted lessons, each 30 minutes in duration. Each lesson contains four to five brief mathematics activities with detailed scripting to systematically introduce mathematics content, bolster fidelity of implementation, and increase the consistency and quality of instructional interactions between teachers and students. Lessons in the first half of Fusion prioritize basic number combinations and place value concepts, whereas lessons in the second half focus on multi-digit computation without regrouping and word problem solving. Figure 2 shows the CCSS (2010) topics addressed by instructional week in Fusion.

< Figure 2 here >

A central feature of Fusion is the careful integration of foundational concepts and skills of whole number, and validated-design principles of explicit and systematic instruction (Coyne et al., 2011; Doabler & Fien, 2013; Doabler, Fien, Nelson-Walker, & Baker, 2012). Each lesson contains opportunities for teachers to (a) model what they want students to learn, (b) deliver scaffolded instructional examples, and (c) provide specific academic feedback to students during the learning activities. For example, if a student responds incorrectly, lesson scripting provides the teacher with procedures on how to explicitly address the error. Lessons also contain frequent opportunities for student practice and judicious review. These opportunities occur through multiple formats but particularly important are practice opportunities that involve verbal interactions between teachers and students. Mathematics verbalizations allow students to communicate their mathematical understanding and thinking, and explain and justify their methods for solving problems (Gersten et al., 2009). In Fusion, students are provided frequent opportunities to verbalize their mathematical understanding through unison, choral responses. For example, a teacher might have a group of five students state in unison the commutative property. To build students' conceptual understanding of number, the curriculum incorporates a variety of visual representations of mathematics, including number lines, strip diagrams, and place value blocks.

Members of our curriculum team developed Fusion through a series of iterative cycles of development, implementation field-testing, analysis, and revision. These iterations provided opportunities for the curriculum team to examine implementation data and make ongoing improvements to the curriculum. Revisions occurred during the first two years of the project and included edits such as shortening the length of lessons and determining the optimal configuration of Fusion's instructional design principles (e.g., sequencing of teaching examples). To help further shape Fusion, the curriculum team collected various forms of professional feedback throughout the development process, including interventionists' perceptions of the curriculum's goals, procedures, and outcomes. These data sources proved instrumental for making adjustments to the Fusion curriculum prior to the start of the current study.

Fusion Procedures. In this study, interventionists were responsible for delivering one lesson per day, three times per week for a total intervention duration of approximately 20 weeks. Lessons lasted approximately 30 minutes and were delivered in small-group instructional formats, with approximately 5 students per group. Instruction for all four groups occurred in classrooms outside of their general education settings and at times that did not interfere with students' core mathematics and reading time. As described above, instruction was led by explicit scripting within each lesson and focused on activities specific to whole number concepts and skills identified in the CCSS (2010). Because mathematical content is systematically reviewed and extended across the 60 Fusion lessons, interventionists did not apply a specific criterion of learning (e.g., ≥ 90%) for advancing in the curriculum. However, interventionists were trained to reteach an activity the following instructional day if they recognized that students struggled to master particular content and if that content was not present in the next lesson. Across all four groups, the total number of instructional sessions ranged from 60 to 65 days.

Professional Development. Prior to the current study, interventionists received four hours of professional development in early mathematics instruction. This initial curriculum workshop focused on four key elements: (a) the research-based principles of mathematics instruction, (b) the instructional design and delivery features of Fusion, (c) an overview of lessons 1-30, and (d) small-group management techniques, such as rewarding positive student behaviors. In the first workshop, participating teachers were provided opportunities to deliver sample lessons and receive feedback on their teaching from the project staff. Midway through the study, all interventionists participated in a four-hour follow-up workshop. A central focus of the final workshop was previewing lessons 31-60 of the curriculum. To enhance implementation fidelity, all four interventionists received on-site coaching from members of the curriculum team. Coaching visits included direct observation and post-observation feedback that focused on implementation fidelity and the quality of instructional interactions.

Fidelity of Implementation. Implementation fidelity of Fusion was assessed using two observation measures: (a) Fusion observation instrument and the (b) Ratings of Classroom Management and Instructional Support (RCMIS; Doabler & Nelson-Walker, 2009). Each intervention group was observed at least five times, and all fidelity observations were scheduled in advance and conducted by trained staff. After interventionists taught each activity (range 4-5 per lesson), observers used the Fusion observation instrument to rate implementation fidelity for each activity, using a 0-2 scale (0 = not taught, 1 = partial implementation, and 2 = full implementation). A fidelity score for each observation was calculated by averaging ratings across the activities. Overall fidelity scores were averaged across the five observation occasions.

The RCMIS is a holistic rating system comprised of 11 items that measures the quality of instructional interactions that take place between teachers and students around critical mathematics content (Cronbach's $\alpha = .92$). Each item is rated on a 4-point scale from low (1) to high (4). For each observation, a score was calculated by summing the ratings across the 11 items. For each group, an overall quality score was calculated as the mean across all observations. Interobserver reliability was conducted on 20% of all observation occasions and reported at 83% and 85% for the Fusion observation instrument and RCMIS, respectively.

Reported fidelity of implementation for all groups was 80% (M = 1.59, SD = .40). Analysis for each instructional group showed implementation fidelity was moderate to high for Group A (84%), Group B (95%), and Group D (93%). Fidelity of implementation for Group C was rated markedly lower (50%). Analysis of instructional quality data documented by the RCMIS showed between-group variability: Group A (M = 34.33, SD = 2.94), Group B (M = 36.00, SD = 5.14), Group C (M = 27.33, SD = 5.31), and Group D (M = 43.16, SD = 1.60). Groups A, B, and D appeared to facilitate high-quality interactions. As with implementation fidelity, instruction quality ratings for Group C were notably lower than the other three groups.

Results

We first summarize results of the feasibility and usability surveys. We then provide results related to Fusion's promise to improve student mathematics achievement. These results are summarized by distal and proximal outcome measures.

Feasibility and Usability

Overall, interventionists rated the Fusion intervention both as feasible and usable in authentic educational settings. On the first survey, for example, which was administered at the second curriculum workshop, interventionists' ratings indicated satisfaction with the progress their students had made through the first 30 lessons of Fusion. Also, all four interventionists agreed that students' progress in mathematics was greatly attributed to Fusion. Additionally, interventionists noted that students' attitudes toward core mathematics improved based on Fusion. Finally, two of the interventionists rated Fusion more favorably for improving mathematics outcomes than other mathematics interventions used in their school.

On the end-of-intervention survey, two of the interventionists reported that the intervention was somewhat easy to implement, whereas the remaining two rated the difficulty of implementation as "average" or similar to that of other mathematics programs. When asked to rate the amount of the teacher scripting provided, responses ranged from "not enough" to "sometimes too much." All four interventionists noted that they completed the curriculum as intended, reporting that the scripting was followed most, if not all, of the time. For usability of Fusion, on average, interventionists found the materials highly user-friendly. For example, interventionists reported the place value template and base-10 blocks as very user-friendly in small-group settings. Finally, on a 7-point scale, with 1 (greatly hindered) to 7 (greatly benefited), interventionists reported that, on average, students benefited from Fusion, both in mathematics ($M = 5.75$, $SD = .50$) and in terms of behavior ($M = 5.00$, $SD = .00$).

Distal Outcome Measures

Figure 3 shows group mean scores on the QD and MN measures, with group averages for the measures depicted as white and black circles, respectively. Introduction of Fusion (i.e., phase change from baseline to intervention) was contingent on baseline stability of the QD measure.

< Add Figure 3 >

Baseline. As shown in Figure 3 between five and eight data points per outcome were collected during baseline for each instructional group. Also, to attend to the WWC Standards for multiple probe designs (Kratochwill et al., 2010), for both outcome measures, each instructional group had at least one data point in the first three sessions and at least three consecutive data points just prior to the introduction of Fusion. For QD, significant between-group variability can be found in Figure 3, with mean scores ranging from 8.20 to 27.93. Visual analysis within the baseline phase of Group A indicates no evidence of variation or upward trend for QD mean scores, with group averages ranging from 8.2 to 11.2. Mean QD baseline scores for Groups B and C show a slight upward trend following the initial data points. For Group B, the final three baseline data points clearly demonstrate a downward trend. Scores for Group C stabilized after the third data point and continued to demonstrate within phase stability across the final three administrations leading up to the introduction of Fusion. QD baseline averages for Group D showed an increase between the third and fourth administrations (23.8 to 27.4). However, the QD averages for Group D appeared to stabilize for the remaining four probes (28.6 to 30.0) and any indication of trend in those data was negligible.

With respect to the MN measure, mean baseline scores for Group A were quite stable, with group averages ranging from 2.6 to 4.2. Mean MN baseline averages for Group B indicate a slight increase between the second and third probes, however averages appeared to stabilize across the final four data points. Group C baseline averages also showed an increase from the second to the third probes. This change is level is likely attributed to the loss of one student from Group C at the time of the third probe. Group B averages ranged from 7.6 to 13.8, whereas for Group C averages ranged from 5.0 to 9.5. For Group D, mean MN baseline scores demonstrate stability in the middle of the phase with a slight change in level in the final probes. Averages for Group D during these final probes ranged from 13.75 to 16.40.

Between-group differences on QD and MN at baseline were tested using ANOVA and Scheffe's test of all pairwise comparisons ($p < .05$). Results indicate that Group A was significantly lower than Group C on QD, and Groups B and D on the QD and MN measures.

Fusion Outcomes for Instructional Groups. In the intervention phase for QD, changes in level and trend were demonstrated for Groups A and D, respectively (see Figure 3). The influence of Fusion on Group B mean QD scores is noted after the fourth intervention phase data point, with mean scores showing a marked change in level of performance during the remaining 20 intervention probe administrations. Group C demonstrated significant variability in QD mean scores during the first half of the curriculum. Despite this variability, a positive, moderate change in level for Group C is noted in final eight weeks of the intervention.

On the MN measure, the mean number of problems answered correctly by Groups A, B, and D increased from baseline to intervention, with notable changes in level and trend for all three groups, particularly during the final weeks of the intervention. Similar to the QD means scores, significant variability in MN mean scores can be found for Group C during the first few weeks of the curriculum along with a downward trend in the middle of the intervention (see Figure 3). However, a somewhat robust change in level for MN means scores can be seen during the last seven probe administrations for Group C.

As an index of effect size, this study calculated Tau-U, using the Tau-U calculator developed by Vannest, Parker, and Gonen (2011). Tau-U is a metric that describes the non-overlap between measurements belonging to the baseline and intervention phases, and is an extension of Tau-nonoverlap in that it controls for baseline trends (Parker, Vannest, Davis &

Sauber, 2011). For this study, Tau-U was computed for the between-phase contrasts for each group and combined the contrasts into a single weighted average to describe the overall functional relation. Results indicated that, for each group, the majority of data pairs between the two phases showed improvement over time, even after controlling for baseline trend ($p < .01$ for all contrasts). The weighted average Tau-U indicated that 80% of all data pairs between the two phases showed improvement over time (Tau-U = .80, $p < .001$), suggesting a functional relation between Fusion and mathematics performance.

Fusion Outcomes for Individual Students. Average baseline and intervention scores, and gains between phases on QD and MN for all students are reported in Table 1. Mathematics gains on each measure were calculated by subtracting the baseline mean from the intervention mean. Individually, results indicate that most students made robust learning gains on the QD and MN measures. One student in Group B, Mark, demonstrated no growth across the 20 weeks on the MN measure. Notable increases of variability in student performance on both measures reveal significant group heterogeneity.

< Table 1 here >

Proximal Outcome Measure

A paired samples *t* test revealed a statistically significant difference between the mean score of student ProFusion pretest (M = 24.71, SD = 16.16) and posttest (M = 50.93, SD = 19.35) performances, $t(13) = 6.50$, p < .001). A pretest-posttest effect size was calculated by dividing the mean gain by the gain standard deviation (M = 26.21, SD = 15.09). The impact of Fusion on the proximal measure produced a pre-posttest effect size gain of 1.73.

Discussion

The purpose of this study was to investigate initial data (i.e., feasibility and usability, and impact on student outcomes) related to the development of Fusion, a Tier 2 mathematics curriculum designed to improve the mathematics outcomes of students with or at-risk for MD. All data were collected in the second year of an IES-funded Development and Innovation project using multiple methods, including surveys, direct observations, and a multiple-probe-across-groups design. While preliminary, several important findings emerged from this study.

Survey data reported positive interventionist feedback in terms of the feasibility and usability of Fusion. Because instructional time is a precious resource in schools, particularly in the context of interventions considered supplemental to core instruction, it is therefore critical to provide interventionists with a curriculum that can be implemented in the expected timeframe. Our development team used interventionist feedback obtained during the first professional development session to better shape Fusion for a 30-minute implementation window. For example, interventionists indicated that some of the lesson activities contained too many instructional examples to teach in 30 minutes. This feedback was used to reduce the number of instructional examples prescribed in the curriculum.

Although not a primary research aim, another important finding of the study was related to implementation fidelity. Observations of intervention implementation indicate that most of the interventionists were able to implement Fusion with acceptable levels of implementation fidelity and facilitate high-quality instructional interactions around whole number concepts. In Fusion’s theory of change, instructional interactions are hypothesized to mediate students’ proximal outcomes. While this study did not conduct mediational analyses, the observational data did provide initial support for our theoretical principles and the potential capacity for Fusion to initiate meaningful teacher-student interactions around critical mathematics content. We base this conclusion on the observation data documented by the RCMIS, which suggest that Fusion deeply engaged students in whole number concepts and skills.

Perhaps the most important finding was related to student mathematics outcomes. We hypothesized that Fusion, when implemented with high levels of fidelity, would improve student mathematics achievement. We based this final hypothesis on a well-specified theory of change for Fusion. Our findings, while preliminary, suggest that all students made promising gains on a measure proximal to the Fusion intervention. On the two distal outcome measures (i.e., QD and MN), visual analysis revealed that experimental control was documented for Groups A, B, and D. Experimental control, however, was not demonstrated with Group C on these measures.

Limitations

While the findings from the study provide initial evidence for Fusion and its theory of change, the data should be interpreted with caution in light of several research limitations. The fact that visual analysis did not reveal immediate effects for all four groups should be noted. A lack of immediacy of the effect, however, may be a function of group-based averages, which can be sensitive to outlying scores and missing data. Student attrition can also influence group-based averages as evidenced in Group C, where three of the five students moved before the end of the 20-week intervention. Despite this, more research is needed on the utility aggregate student data affords small-group mathematics interventions and, moreover, how such data can be used in tandem with the outcomes of individual students to inform instructional decision-making.

It is also important to highlight the difficulty for interventions that target academic outcomes to demonstrate immediate effects or a magnitude of change that applies to visual analysis in SCD. Unlike behavioral outcomes, such as the reduction of self-injurious behavior, acquisition of academic knowledge is incremental, systematically building upon prior understandings in a step by-step fashion. In mathematics, as Griffin (2005) indicates, understanding is gradually constructed across time, with instruction allowing children to “deepen and consolidate each new understanding before moving on to the next” (p. 281). For example, to acquire proficiency in strategic counting, a targeted outcome in this study, a child will first have to have prior understanding of making one-to-one correspondences, cardinality, and counting all. Such skills were incrementally taught across Fusion rather than in one set of lessons.

Another limiting factor was the study’s sample size, which substantially constrains the generalizability of our findings. We also recognize heterogeneity of instructional groups as a potential limitation. Notable increases of variability in student performance across time were documented. We took this as an indication that our sample had different subgroups of student risk types. The impact of Fusion on student mathematics achievement may depend on student risk status. Additionally, because each instructional group had its own interventionist, the study design did not control for interventionist effects (i.e., the same interventionist for all groups).

Also, the CBM measures may lack sensitivity to student growth on whole number concepts and skills. Unlike phonemic awareness and the alphabetic principle in early literacy, mathematics intervention studies have yet to casually link the skills associated with QD and MN to students’ mathematics achievement. Moreover, while the modified CBMs demonstrated adequate technical characteristics, we are not fully certain how the adaptations may have impacted the screening utility of the QD and MN measures. It is important to note that a primary aim of using these modified CBMs was to minimize the disruption of classroom operations and maximize students’ opportunity to access core instruction. By our calculations, we estimate these group-based CBMs saved approximately 2-3 hours of assessment time per child across the project. These savings may have implications for the resources that future research studies must provide to investigate interventions on RtI and mathematics. For example, if research projects can concurrently assess multiple students on group-administered CBMs and the data garnered is both reliable and valid, then researchers can save precious resources and remain in good standing with schools because of the minimal loss of students' opportunity to learn core content. Finally, it is likely that the QD and MN measures were too distal to the content of Fusion to adequately capture its impact on student mathematics achievement. Future studies, therefore, are needed to identify other measures of whole number proficiency for demonstrating the promise of Fusion.

Implications for Instruction and Design Research

Our results indicate that impact of the Fusion intervention for Group C was not as evident. From a curriculum development perspective, this finding can be thought of as bringing to light limiting factors within our design research. Curriculum research, like other empirical sciences, moves forward when hypotheses and theories are tested and then subjected to self-correction and professional scrutiny (Feuer, Towne, & Shavelson, 2002; Popper, 1959). Thus, researchers must be prepared to examine, and in some case modify, existing theories based on scientific results.

Take for example, the current study where the Group C interventionist demonstrated low levels of implementation fidelity and instruction quality. One possible reason for these findings is that the Group C interventionist was uncomfortable with the lesson scripting provided in Fusion. Interestingly, findings from studies involving scripted curricula typically point to high levels of implementation fidelity (Simmons et al., 2007). In our research, we have found, albeit anecdotally, that teachers and instructional assistants are relieved to receive and implement a scripted program because it removes the onus of having to design instruction for students struggling in reading and mathematics. Regardless, the Group C interventionist may have found the Fusion scripting restrictive or misaligned with his approach to instruction. Future research should examine how varying levels of scripting associates with teachers' pedagogical formats and knowledge for teaching early mathematics.

Another implication generated from the fidelity of implementation results of Group C pertains to the type of observation protocols used in the study. Our observation system focused strictly on the quality of instructional interactions and implementation fidelity of Fusion. However, to further unpack the active ingredients of Fusion and better understand the professional development needs of teachers, researchers may want to consider other observation instruments, such as those that capture the quantity of instructional interactions (Doabler et al., in press; Smolkowksi & Gunn, 2012). For example, data collected from a frequency-based instrument might suggest that a group's inadequate response to instruction, such as the case of Group C, is a result of too few practice opportunities for groups of students. Researchers could use such observation data to provide professional development to teachers on facilitating and managing student practice opportunities, such as choral responses.

Conclusion

In summary, Clements (2007) proposed that researchers engineer mathematics curricula through a design science approach. Linking science and curriculum development has researchers collect a variety of empirical data to make iterative refinements to the curriculum, project instrumentation, and the professional development materials for teachers. Another expectation of curriculum research is to make known research findings to the scientific community. This study was a first attempt to examine and disclose how a recently developed Tier 2 mathematics curriculum functioned in authentic education settings. Preliminary data suggest both the potential of the Fusion intervention for students and teachers along with the need for future refinement and further evaluation of the curriculum through increasingly rigorous research methods.

References

Anderson T., Shattuck J. (2012). Design-based research a decade of progress in education research? *Educational Researcher*, 41, 16–25. Crossref

Baker S., Gersten R., Lee D. S. (2002). A synthesis of empirical research on teaching mathematics to low-achieving students. *The Elementary School Journal*, 103, 51–73. Crossref. ISI.

Bodovski K., Farkas G. (2007). Mathematics growth in early elementary school: The roles of beginning knowledge, student engagement, and instruction. *Elementary School Journal*, 108, 115–130. Crossref. ISI.

Carnine D. W., Silbert J., Kame'enui E. J., Tarver S. G. (2004). *Direct instruction reading* (4th ed.). Upper Saddle River, NJ: Pearson.

Chard D. J., Baker S., Clarke B., Jungjohann K., Davis K., Smolkowski K. (2008). Preventing early mathematics difficulties: The feasibility of a rigorous kindergarten mathematics curriculum. *Learning Disability Quarterly*, 31, 11–20. Crossref. ISI.

Clarke B., Doabler C. T., Strand Cary M., Kosty D., Baker S. K., Fien H., Smolkowski K. (in press). Development and preliminary evaluation of a tier-2 intervention targeting whole number concepts in first grade mathematics: Utilizing a theory of change to guide formative evaluation activities. *School Psychology Review*.

Clarke B., Shinn M. R. (2004). A preliminary investigation into the identification and development of early mathematics curriculum-based measurement. *School Psychology Review*, 33, 234–248. Crossref. ISI.

Clements D. H. (2007). Curriculum research: Toward a framework for "research-based curricula." *Journal for Research in Mathematics Education*, 38, 35–70. ISI.

Common Core State Standards Initiative. (2010). *Common core standards for English language arts & literacy in history/social studies, science, and technical subjects*. Retrieved from http://www.corestandards.org/assets/CCSSI_ELA%20Standards.pdf

Coyne M. D., Kame'enui E. J., Carnine D. W. (2011). *Effective teaching strategies that accommodate diverse learners*(4th ed.). Upper Saddle River, NJ: Prentice Hall.

Diamond K. E., Powell D. R. (2011). An iterative approach to the development of a professional development intervention for head start teachers. *Journal of Early Intervention*, 33, 75–93. Crossref ISI.

Doabler C. T., Baker S. K., Kosty D., Smolkowski K., Clarke B., Miller S. J., Fien H. (in press). Examining the association between explicit mathematics instruction and student mathematics achievement. *Elementary School Journal*. ISI.

Doabler C. T., Fien H. (2013). Explicit mathematics instruction: What teachers can do for teaching students with mathematics difficulties. *Intervention in School and Clinic*, 48, 276–285. Crossref. ISI.

Doabler C. T., Fien H., Nelson-Walker N. J., Baker S. K. (2012). Evaluating three elementary mathematics programs for presence of eight research-based instructional design principles. *Learning Disability Quarterly*, 35, 200–211.Crossref. ISI.

Doabler C. T., Nelson-Walker N. J. (2009). Ratings of Classroom Management and Instructional Support (Unpublished observation instrument). Center on Teaching and Learning, College of Education, University of Oregon, Eugene.

Donovan M. S., Bransford J. D. (Eds.). (2005). *How students learn: History, mathematics and science in the classroom*. Washington, DC: National Academies Press.

Feuer M., Towne L., Shavelson R. J. (2002). Scientific culture and educational research. *Educational Researcher*, 31(8), 4–14. Crossref

Fien H., Smith J. L., Baker S. K., Chaparro E., Baker D. L., Preciado J. A. (2011). Including English learners in a multitiered approach to early reading instruction and intervention. *Assessment for Effective Intervention*, 36, 143–157. Crossref.

Fuchs D., Fuchs L. S., Stecker P. M. (2010). The "blurring" of special education in a new continuum of general education placements and services. *Exceptional Children*, 76, 301–323. Crossref. ISI.

Gast D. L. (2009). *Single subject research methodology in behavioral sciences*. Routledge: New York, NY.

Gersten R. M., Chard D., Jayanthi M., Baker S. K., Morphy P., Flojo J. (2009). Mathematics instruction for students with learning disabilities: A meta-analysis of instructional components. *Review of Educational Research*, 79, 1202–1242. Crossref ISI.

Griffin S. (2005). Fostering the development of whole-number sense: Teaching mathematics in the primary grades. In National Research Council (Ed.), *How students learn: History, mathematics and science in the classroom* (pp. 257–308). Washington, DC: National Academies Press.

Institute of Education Sciences. (2013). *RFA: Request for applications: Education research grants* (CFDA Number: 84.305A). Washington, DC: Institutes of Education Sciences, U.S. Department of Education.

Kame'enui E. J., Simmons D. C. (1999). *Towards successful inclusion of students with disabilities: The architecture of instruction: An overview of curricular adaptations*. Reston, VA: Council for Exceptional Children.

Kratochwill T. R., Hitchcock J., Horner R. H., Levin J. R., Odom S. L., Rindskopf D. M., Shadish W. R. (2010). *Single-case designs technical documentation*. Retrieved from http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_scd.pdf

Lindo E. J., Elleman A. M. (2010). Social validity's presence in field-based reading intervention research. *Remedial and Special Education*, 31, 489–499. Crossref

Locuniak M. N., Jordan N. C. (2008). Using kindergarten number sense to predict calculation fluency in second grade. *Journal of Learning Disabilities*, 41, 451–459. Crossref PubMed. ISI.

McKenney S., Reeves T. C. (2013). Systematic review of design-based research progress is a little knowledge a dangerous thing? *Educational Researcher*, 42, 97–100. Crossref ISI.

Morgan P. L., Farkas G., Wu Q. (2009). Five-year growth trajectories of kindergarten children with learning difficulties in mathematics. *Journal of Learning Disabilities*, 42, 306–321. Crossref PubMed. ISI.

National Mathematics Advisory Panel. (2008). *Foundations for success: The final report of the national mathematics advisory panel*. Washington, DC: U.S. Department of Education.

National Research Council. (2001). *Adding it up: Helping children learn mathematics*. Washington, DC: Mathematics Learning Study Committee.

Parker R. I., Vannest K. J., Davis J. L., Sauber S. B. (2011). Combining non-overlap and trend for single case research: Tau-U. *Behavior Therapy*, 42, 284–299. Crossref. PubMed. ISI.

Popper K. R. (1959). *The logic of scientific discovery*. London, England: Hutchinson.

Rothman A. (2013). Exploring connections between moderators and mediators: Commentary on subgroup analyses in intervention research. *Prevention Science*, 14, 189–192. Crossref PubMed.

Schmidt W., Houang R., Cogan L. (2002). A coherent curriculum: The case of mathematics. *American Educator*, 26, 1–18.

Shadish W. R., Cook T. D., Campbell D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton-Mifflin.

Simmons D. C., Kame'enui E. J., Harn B., Coyne M. D., Stoolmiller M., Santoro L. E., . . . Kaufman N. K. (2007). Attributes of effective and efficient kindergarten reading intervention: An examination of instructional time and design specificity. *Journal of Learning Disabilities*, 40, 331–347. Crossref PubMed. ISI.

Smolkowski K., Gunn B. (2012). Reliability and validity of the Classroom Observations of Student-Teacher Interactions (COSTI) for kindergarten reading instruction. *Early Childhood Research Quarterly*, 44, 48–57.

Superfine A. C., Kelso C. R., Beal S. (2010). Examining the process of developing a research-based mathematics curriculum and its policy implications. *Educational Policy*, 24, 908–934. Crossref

U.S. Department of Education, Institute of Education Sciences and National Science Foundation. (2013). *Common guidelines for education research*. Retrieved from http://www.nsf.gov/pubs/2013/nsf13126/nsf13126.pdf?WT.mc_id=USNSF_179

Vannest K. J., Parker R. I., Gonen O. (2011). *Single case research: Web based calculators for SCR analysis* (Version 1.0). College Station: Texas A&M University. Available from singlecaseresearch.org

What Works Clearinghouse (2013). *Math publications and reviews*. Retrieved from http://ies.ed.gov/ncee/wwc/Topic.aspx?sid=9

Wu H. H. (2009). What's sophisticated about elementary mathematics? *American Educator*, 33(3), 4–14.