Article

Hammill Institute on Disabilities

Learning Disability Quarterly
2017, Vol. 40(4) 211–224
© Hammill Institute on Disabilities 2017
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0731948717714715
journals.sagepub.com/home/ldq
SAGE

A Practitioner Implementation of a Tier 2 First-Grade Mathematics Intervention

Mari G. Strand Cary, PhD[1], Ben Clarke, PhD[1], Christian T. Doabler, PhD[2], Keith Smolkowski, PhD[3], Hank Fien, PhD[1], and Scott K. Baker, PhD[4]

Abstract

We report on a practitioner implementation of Fusion, a first-grade mathematics intervention. Studies such as this evaluation of a loose implementation under realistic conditions are important to curriculum developers' understanding of how evidence-based programs and tools work under a variety of implementation scenarios. In this quasi-experimental study, first-grade students (n = 253) in 10 schools were assigned to treatment (n = 154) or control (n = 99). Rather than randomly assigning students, schools assigned those students most at risk to treatment and, typically, those less at risk to control. School staff administered pre- and postassessments and led Fusion sessions approximately 30 min per day, 3 days per week. The intervention resulted in a significant positive effect on a researcher-developed first-grade math measure. The implementation of Fusion and feedback from school staff provided insights to guide the curriculum development process.

Keywords

at risk, program evaluation, content-area instruction, quantitative methods

Many parents, educators, researchers and policy makers recognize that mathematics is critical for 21st century jobs and full participation in everyday life. Yet, despite repeated calls for increased funding, rigor, and performance in the area of mathematics (Maltese & Hochbein, 2012), the overall performance of U.S. students continues to be below that of students in other countries (Program for International Student Assessment [PISA]; Organisation for Economic Co-Operation and Development [OECD], 2012), and the average score on the National Assessment of Educational Progress (NAEP) of both fourth and eighth graders falls into the below basic category (National Center for Education Statistics, 2013). Results for minority students and students from low SES are even more dismaying. Approximately 5% of the school-age population has identified learning disabilities (National Center for Learning Disabilities [NCLD]; Cortiella & Horowitz, 2014) that prevent them from learning at the same pace or with the same supports as their peers. Students with little or no mathematics or preschool experience enter elementary school academically behind their peers (Barnett, Jung, Frede, Hustedt, & Howes, 2011), and an increasing portion of students are learning both English and the academic subjects at the same time. These facts highlight the academic diversity of U.S. classrooms and the challenges teachers face in trying to ensure all their students productively engage with early mathematics concepts and use them as a foundation for more advanced learning. Unfortunately, most of the core mathematics curricula in use in the nation's schools offer little support. These curricula address a wide swath of content, assume rather than teach prerequisite knowledge needed to understand and utilize that content, and do not include strategies useful for both struggling and high achieving students (Doabler, Fien, Nelson, & Baker, 2012).

As early as kindergarten and first grade, achievement gaps are present between the average student and those who enter school with a poor understanding of mathematics or experience math learning difficulties. In subsequent years, as difficulties morph into—or are identified as—disabilities, disenfranchised students lose faith in their ability to learn, and achievement gaps tend to remain and even widen (Morgan, Farkas, & Wu, 2009). Struggling students will continue to face challenges in learning mathematics unless

[1]University of Oregon, Eugene, USA
[2]University of Texas at Austin, USA
[3]Oregon Research Institute, Eugene, USA
[4]Southern Methodist University, Dallas, TX, USA

Corresponding Author:
Mari G. Strand Cary, Center on Teaching and Learning, University of Oregon, Eugene, OR 97403-5292, USA.
Email: mscary@uoregon.edu

schools provide additional supports (Agodini & Harris, 2010; Clements, Agodini, & Harris, 2013), including differentiating instruction within whole-class settings or providing interventions for the individual needs of struggling students.

One potential solution to meet the needs of a wide range of students is a response to intervention (RTI) approach (National Association of State Directors of Special Education [NASDE], 2006). Educators successfully implementing RTI flexibly meet students' learning needs through a tiered system of support (Fuchs et al., 2005), with a goal of preventing mathematics difficulties before they are fully established (Cortiella & Horowitz, 2014; Seethaler & Fuchs, 2011). To be successful, educators implementing RTI must have access to effective universal screening measures to identify at-risk students, high quality core curricula and supplementary Tier 2 and Tier 3 interventions (Fuchs & Vaughn, 2012; Riccomini & Smith, 2011), means of formatively monitoring student response to instruction, and the resources and support to adjust instruction as needed. By definition, a student who is slated to receive Tier 2 services should have already received high quality Tier 1 (core) instruction and not made meaningful progress. The Tier 2 instruction will, ideally, give the student the boost he or she needs in knowledge and skills so that he or she can be successful in Tier 1 instruction once Tier 2 services are removed.

Interest in early elementary mathematics interventions suitable for use in an RTI framework is growing (Clarke, Baker, & Fien, 2009). Although a deep conceptual understanding of whole number provides a critical, strong foundation for later mathematics concepts (Jordan & Levine, 2009; National Mathematics Advisory Panel [NMAP], 2008; National Research Council [NRC], 2009), relatively few interventions (Clarke et al., 2014) prioritize instruction around whole number and deliver this content to students using proven principles of instruction (Baker, Gersten, & Lee, 2002; Bryant, Bryant, Gersten, Scammacca, Funk, et al., 2008; Clarke et al., 2009). However, a number of researchers have begun to develop Tier 2 interventions that demonstrate potentially promising outcomes for students (Bryant, Bryant, Gersten, Scammacca, & Chavez, 2008; Bryant, Pfannenstiel, & Bryant, 2014; Fuchs et al., 2005). Work in the development of Tier 2 interventions has often been guided by curriculum development and evaluation frameworks (Clements, 2007), and tends to include explicit and systematic instruction in the area of whole number (Gersten et al., 2009). A program that illustrates this dual focus on systematic and explicit instruction in the area of whole number is Number Rockets (Fuchs et al., 2005). This 63-lesson, Grade 1 program includes 17 units focused on whole number concepts and skills. It is delivered through small-group and computer-based instruction. A randomized controlled efficacy trial revealed effect sizes of 0.11 to 0.70 on computation, concepts and applications, and story problems, but not fact fluency. Work from this initial efficacy trial was followed by a larger scale effectiveness study (Gersten et al., 2015) that revealed the Number Rockets intervention to be effective as measured by student gains (i.e., effect size = 0.34) on the *Test of Early Mathematics Ability, Third Edition* (Ginsburg & Baroody, 2003). The early numeracy intervention by Bryant et al. (2011) provides another example. Its systematic and explicit instruction helps young, struggling students gain conceptual, strategic, and procedural knowledge related to number and operation concepts and skills. Interventionists follow systematic and explicit routines that include teacher modeling, use of concrete and visual representations, guided and independent practice with error correction procedures, and judicious review.

Developers and authoring teams with backgrounds in teaching, educational psychology, and educational research highly value educator feedback and testing in schools throughout development. Unfortunately, even university-based teams with federal funding tend not to have the access, resources, or length of time needed to do large-scale formative evaluations. Developers and researchers are constrained to a handful of authentic contexts and stakeholders until—and unless—they procure scarce funding for summative, large-scale evaluations and scale-up research (e.g., Goal 3 and Goal 4 studies funded by the Institute for Education Sciences [IES]). Even then, they face challenges to recruitment and fidelity to study design, especially when using traditional rigorous research designs that call for delaying or withholding services from some students. This is not just an issue for developers of new curricula or services. When seeking to evaluate existing practices or educational innovations, schools themselves face these same challenges related to resources and time, as well as challenges related to a lack of evaluation personnel and expertise. All this is to say that school-based evaluations are still relatively rare.

Building a Corpus of Evidence for Emerging Programs

Vested in bridging research and practice, developers and researchers are increasingly combining rigorous large and small, randomized control trial (RCT) studies with studies assessing effectiveness and ease of implementation in ways more palatable to educators (e.g., design-based research, regression discontinuity designs, truly collaborative university–district research partnerships, practitioner-initiated studies). If given some control and flexibility, schools are far more likely to buy in to the research endeavor. This combined approach extends the research of developers and researchers by effectively utilizing development and evaluation dollars while providing richer, more varied insights to the field. It allows effective and ineffective components to be identified earlier in the development process and affords

researchers more opportunities to identify patterns in implementation and efficacy that can illuminate the mechanism(s) of change (i.e., why a program works for all or subgroups of students). Having learned as much as possible about active ingredients, critical components, and learning outcomes during development phases, researchers are then well-positioned and justified in conducting rigorous large-scale evaluations characterizing the latter end of the research continuum. Ultimately, researchers, practitioners, and policy makers will benefit from a cumulative, realistic sense of program demands and payoffs.

This type of triangulation and replication builds a corpus of evidence involving a variety of techniques and designs and is a desirable, though oft-neglected, feature of education research (Simmons et al., 2011). As Smolkowski, Strycker, and Seeley (2013) note, "to enhance desired outcomes, programs should be installed and implemented with as much fidelity as possible to the original designs, but adaptations and innovations may be needed to fit the community, school, students, and staff." Ideally, through broad pilot work and initial larger-scale testing, researchers and developers would test or at least observe all possible implementation approaches, variables, and unexpected consequences so that they can factor all those in when planning for efficacy studies and so that educators can make more informed decisions about which interventions might fit their contexts and students. Researchers and funders (e.g., IES) alike recognize this, yet it is often practically difficult or impossible within development projects and even efficacy studies to fully investigate the many factors that can aid or sabotage implementation of an intervention (Smolkowski et al., 2013). This means researchers implement subsequent efficacy studies with some—but not necessarily enough—evidence behind their approach.

The Fusion Intervention Project

We took this multi-study approach to bridging the research-to-practice gap as we developed and evaluated a first-grade Tier 2 intervention, Fusion. Fusion is comprised of whole number instruction with explicit and systematic instructional design principles to improve student mathematics achievement (Clarke et al., 2014).

Development and Formal Pilot Study

Through a 3-year, federally funded IES grant, we first used curriculum development and evaluation frameworks to develop and formatively evaluate the Fusion program. Specifically, during Fusion's 2-year development phase, university faculty worked closely with teacher-researchers from local districts to iteratively develop and test components of the intervention. Teacher-researchers implemented small portions of lessons, then sets of lessons with their students as lessons were developed, then helped train other teachers to implement the full curriculum to test its feasibility as a whole. These increasingly complex and lengthy formative trials prepared the curriculum and the Fusion research team for the summative evaluation phase in the third year of the project. Our development process and the Fusion curriculum itself are described in further detail in Clarke et al. (2009). The development project concluded with a summative, formal pilot study (Clarke et al., 2014) during which 89 first-grade students determined to be at-risk in mathematics were randomly assigned to Fusion (n = 44) or a standard district practice control condition (n = 45). Treatment students participated in Fusion lessons led by research-team-trained district interventionists 30 min per day, 3 days per week for 20 weeks. Results from random-effects models revealed that gain scores of intervention students were significantly greater than their control peers on a proximal measure of conceptual understanding (estimate = 12.9, p = .015, Hedges' g = 0.82, corresponding to a large effect) and were positive, but not statistically significant, on a proximal measure of procedural fluency (estimate = 7.8, p = .667, Hedges' g = 0.14), and a distal measure of achievement (estimate = 1.1, p = .590, Hedges' g = 0.11).

Large-Scale Efficacy Study

That rigorous, but small-scale evaluation within the development grant provided "evidence of promise"; thus, we next pursued funding for a larger-scale IES-funded efficacy study. Such studies are funded to evaluate "fully developed education interventions . . . under ideal or routine conditions by the end user in authentic education settings" (IES, 2016b). The IES Goal 3 efficacy study (Clarke, Doabler, Fien, & Smolkowski, 2016) is designed to study the efficacy of Fusion across four cohorts, 120 classrooms, and 1,200 first-grade students identified as at risk for math learning disabilities. The project is currently in year 1. Within classroom, students are randomly assigned to a high intensity Fusion intervention (group size of two), low intensity Fusion intervention (group size of five), or control group (district business as usual). In-depth data collection includes observations of implementation fidelity, documentation of instructional practice through the use of instructional logs, and proximal and distal mathematics measures to determine impact on student outcomes.

District-Initiated Implementation and Evaluation

Concurrent with the formal pilot study (Clarke et al., 2014), we also had an opportunity to measure Fusion's fit and effectiveness under more realistic, less controlled conditions. The opportunity arose when educators in Hawaii expressed interest in using the Fusion intervention and receiving professional development (PD) from Fusion

authors. Although not available to the public at the time, Fusion was a good candidate for independent implementation by schools because it was designed to be led by instructional assistants with some, but not extensive, Fusion-focused PD (i.e., Fusion does not require interventionists to have extensive content knowledge about Grade 1 mathematics and pedagogy or to receive substantial training). We expected this independent implementation by practitioners to help us "better understand the impact of [this intervention] under authentic and replicable conditions that parallel the resources and personnel typically found in schools" (Simmons et al., 2011) and to provide a hint of the challenges and successes we might expect during more formal effectiveness and scale-up research studies (as noted by Smolkowski et al., 2013). Many interventions never reach that end of the research continuum, so it is even more useful to anticipate practitioner challenges during development and small-scale evaluation. Our purpose in including this study in our work was to provide an additional source of information to help guide refinement of the curriculum and to help prepare for the larger scale efficacy studies that would be conducted under an IES Goal 3 efficacy grant.

Coordinating District and Researcher Objectives

The district-initiated evaluation took place under authentic conditions (i.e., contexts, interventionists, resources, and students), not highly-controlled research conditions. The research team and the school district were primarily interested in the degree to which Fusion affected student mathematics achievement under realistic implementation conditions. Administrators wanted information on (a) whether school staff could successfully implement the program, (b) whether students would learn from the program, and (c) whether educators and students would enjoy using the program. University researchers hoped the evaluation would provide a realistic view of Fusion's implementation to complement findings from the pilot study. Given that this study was not conducted under ideal conditions (i.e., it did not have heavy institutional or research support), we did not expect gains to be as great as in our more formal pilot study (Clarke et al., 2014). We did expect to informally and qualitatively learn about the variety of ways schools independently choose to use Fusion in authentic contexts, intervention feasibility, and teacher perceptions of the intervention.

Because the district and the research team were both interested in the extent to which Fusion would improve student outcomes (our primary research question), district administrators agreed to provide student outcome data and to solicit feedback from educators about ease of implementation and general perceptions of the program. Administrators also agreed to provide data about students who did not receive Fusion instruction (e.g., a quasi-control group). Given distance and lack of research funds, it was not possible to do a rigorous RCT or regression discontinuity study. The research team had little to no control over sample size, participant selection (i.e., rigorous adherence to cut-scores or screening criteria), or program implementation. However, the team was able to provide training to the interventionists and those administering assessments, provide recommendations for student selection and intervention implementation, and solicit educator feedback that we could use to inform the design of and support offered during later, large-scale, highly controlled efficacy studies of the Fusion intervention (e.g., Clarke et al., 2014).

In sum, educators in authentic contexts received Fusion training but then implemented the intervention without external funding, resources, or support. Schools administered a pre- and posttest battery to participating students, and received guidance regarding student identification, assignment, and implementation practices, but were not subjected to the extensive oversight, measurement, and timelines traditionally associated with research studies. The study was in line with Dewa et al.'s (2002) definition of educational scale-up studies:

> In a multisite study, participating sites may provide different services but share a common protocol. Operationally this translates into measuring the same outcomes with the same instruments using the same timeframe across different programs at multiple sites. The common protocol makes outcomes comparable.

This district-initiated evaluation gave us the opportunity to come closer to the idealistic approach (e.g., conducting studies along the entire research continuum) during our development and evaluation phase and allowed the district to evaluate the impact of a new practice with greater rigor than is typical for assessing changes to standard practice.

Method

Design

This practitioner implementation enabled the research team to compare outcomes for students participating in Fusion to students receiving the typical instruction and services offered by schools in the Hawaii Department of Education. Because students who most needed intervention were offered Fusion, and not randomly assigned to treatment or control, this study is best classified as a quasi-experimental study. Our primary research question was

> **Research Question:** What is the impact of the Fusion program on the mathematics achievement of at-risk students?

We hypothesized the following:

Hypothesis: Students who experienced Fusion would make meaningful performance gains on proximal measures (i.e., those with content directly linked to the Fusion curriculum) and would make small gains on distal measures (i.e., those with different formats or that encompass far more content than covered in Fusion).

Such results would narrow the gap between them and their higher performing (at pretest) peers. Secondarily, we sought information about the feasibility of Fusion.

Participants and Setting

With prior approval from the district and the research team's human subjects review board, nine schools in four distinct areas of the state participated. The district serves the majority of the state's 349,086 inhabitants under the age of 18. The population is ethnically diverse (35% Multi-ethnic, 23% Asian, 13% White, 12% Latino, and 11% Hawaiian Native; 10% of students qualify as English Language Learners) and economically diverse (31% qualify for free/reduced lunch). Beyond PD and assessment critical to the study (details below), schools received no compensation for participating.

Recommended identification of at-risk students and assignment to condition. The research team recommended that at-risk students participating in the study be those exhibiting the lowest math performance at each school. Specifically, the team encouraged schools to identify students through classroom teacher recommendations based on (a) a screening measure (i.e., students' fall scores on the district's standard first-grade progress monitoring measure, easyCBM; Alonzo, Tindal, Ulmer, & Glasgow, 2006) and (b) teacher knowledge of in-class performance. Teachers were asked not to select students whose math difficulties were likely due to behavioral issues or chronic absenteeism. The lowest performers in each school were to be identified as at-risk students who might benefit from the Fusion intervention. In the context of this study, it was not feasible to consider more nuanced or alternate definitions of at-risk. Up to double the number of students the schools were able to serve (i.e., provide Fusion to) were to be identified so that half could be assigned to Fusion and half to the comparison condition. To illustrate, schools offering one Fusion group were instructed to identify 10 students, whereas those offering five groups were to identify 50 students.

To determine which students actually participated in Fusion, schools were asked to administer Early Numeracy–Curriculum Based Measures (EN-CBM; Clarke & Shinn, 2004) to the identified at-risk students, then assign those students beneath the school median to intervention (Fusion instruction) and those above to the comparison condition (standard district practice). Thus, the intervention group was designed to be lower performing than their comparison group peers.

Actual identification of at-risk students and assignment to condition. In reality, schools approached identification and assignment to condition in many ways. For example, one school reported assigning all students performing below the 40th percentile on the easyCBM screening measure to the intervention group. At one school, teachers met as a group to assign students to condition based on all three screening and pretest measures (heavily weighting EN-CBM). A third school incorporated all screening and pretest measures as well as special education and general education teacher recommendations and "attendance, tenacity, and overall instructional behavior" into their assignment process. One educator reported that seemingly arbitrary changes to group assignment were made after an agreed-upon assignment process had taken place. Logistics, student behavior patterns, and scheduling constraints also played a role in student assignment at many schools.

In sum, although schools were asked to follow a step-by-step identification and assignment process that would have standardized inclusion to some degree across participating schools, educators utilized and differentially weighted a wide range of data as they made their decisions. Assignment processes were not fully documented by school staff or articulated to the research team and thus cannot be considered in our analyses.

The final sample included 253 first-grade students, 154 intervention and 99 comparison. According to district data, the sample was 63% Hawaiian or Pacific Islander, 32% Asian, 19% White, 6% Hispanic, and 5% American Indian or Alaskan. More detailed demographic data (e.g., gender, free and reduced lunch participation, English learner status) were requested for all participants; it was only received for subsets of students and thus cannot be usefully included in our analyses. No other demographic information was available for all students.

Coordinators and interventionists. Participating schools identified one to two staff (sometimes classroom teachers) to serve as coordinators. Coordinators (n = 11) arranged the Fusion intervention schedule and selected interventionists to teach the Fusion lessons. They also managed the screening, selection, and assessment of students, provided support to the interventionists, and served as liaisons to the research team. Interventionists (i.e., a mix of instructional aides, classroom teachers, and coordinators) prepared and taught the Fusion lessons and provided feedback on the intervention through conversations with coordinators and end-of year surveys administered online. Interventionists and coordinators attended four PD sessions led by the authoring and research team to help them understand the intervention curriculum and their roles in instruction, assessment, and, for

coordinators, instructional support. Coordinators and interventionists were not formal participants in the study and received no stipends for participation. They and participating students' classroom teachers were invited to complete surveys, but survey participation was voluntary (i.e., we entered them into drawings for classroom supplies if they completed surveys) and largely anonymous, and we were not able to obtain enough educator characteristics data to meaningfully report here.

Measures

Student assessments included the easyCBM (Alonzo et al., 2006), Early Numeracy Curriculum-Based Measurement (EN-CBM; Clarke & Shinn, 2004), ProFusion (researcher developed), and Stanford Achievement Test–Tenth Edition (SAT-10; Harcourt Educational Measurement, 2002). Educators were invited, but not required, to share information about themselves and their use and perception of Fusion through researcher-developed online surveys

easyCBM. Based on the National Council of Teachers of Mathematics (NCTM) Focal Point Standards, easyCBM (Anderson, Alonzo, & Tindal, 2010) tests emphasize conceptual understanding over basic computation. Students scoring below benchmark on these normed tests can be considered at-risk. Each individualized math assessment is computer-administered and contains 16 items, and there are 10 comparable forms. Estimated administration time is 18 to 30 min per assessment. For Grade 1, the measures exhibit strong internal consistency (Cronbach's α from .78-.89) and concurrent validity (correlation of .73 with the Terra Nova; Anderson et al., 2010). EasyCBM was included as a distal measure.

EN-CBM. The 1-min EN-CBMs (Clarke & Shinn, 2004) were used as distal measure of students' procedural fluency. *Oral Counting* requires students to rote count as high as possible without making an error. Concurrent and predictive validity range from .46 to .72. *Number Identification* is also an oral measure requiring students to identify numbers between 0 and 10 when presented with a set of printed number symbols. Concurrent and predictive validity range from .62 to .65. *Quantity Discrimination* requires students to name which of two visually presented numbers between 0 and 10 is greater. Concurrent and predictive validities range from .64 to .72. Students completing *Missing Number* name the missing number from a string of numbers (0-10). Concurrent and predictive validities range from .46 to .63. A total EN-CBM score, calculated by summing raw scores from the four subtests, was used in the analysis.

ProFusion. The Fusion research team developed the ProFusion measure to assess students' conceptual and procedural knowledge of number and numeration, place value concepts, basic number combinations, and problems involving multi-digit addition and subtraction. It is administered in an untimed, group setting. Students are asked to write numbers from dictation (four items) and numbers missing from a sequence (three items), write numbers matching base 10 block models (three items), and decompose double digit numbers (three items). Students also complete addition problems and subtraction problems (eight items), story problems (two items) and 1-min, timed addition (32 items possible) and subtraction (24 items possible) fluency measures. Finally, students work with proctors individually to identify numbers (eight items). Because its content is aligned with content presented in Fusion, it was used as a proximal measure.

SAT-10. The SAT-10 (Harcourt Educational Measurement, 2002) is appropriate for kindergarten through Grade 12 students. It is a group-administered, norm-referenced exam and contains two math subtests used in this study as distal measures of mathematics performance: *Math Problem Solving* assesses problem solving and mathematical reasoning; *Math Procedures* assesses computational fluency. A standardized achievement test, the SAT-10 has adequate and well-reported validity ($r = .67$) and reliability ($r = .93$). We considered it to be a far-transfer, distal measure in the context of this study.

Extant data and educator surveys. Student assessment measures were supplemented with requests for student assignment and demographic data, and phone and email conversations with coordinators. Through invitations passed on by coordinators, we invited classroom teachers, interventionists, and coordinators to complete role-specific versions of an optional, researcher-designed survey focused on educator characteristics, Fusion implementation details, perceptions of student learning attributable to Fusion, and educator perceptions of the Fusion curriculum and associated PD. Most questions were comprised of Likert scales followed by "please explain" prompts. In some cases (e.g., obstacles to implementing Fusion, length of lessons, district role of interventionists.), categorical formats were utilized; whenever appropriate, these too were followed by open-ended (e.g., "please explain") prompts. As described in the "Results" section, our requests went largely unheeded; response rates were low.

Procedures

Fusion intervention. Fusion is a 60-lesson, Grade 1 (Tier 2) mathematics intervention designed to promote students' mathematical proficiency with whole number concepts and skills. Specifically, Fusion targets content from two mathematical domains identified in the Common Core State

Standards Initiative (2010) for first-grade mathematics: (a) Operations and Algebraic Thinking, and (b) Number and Operations in Base Ten. Within these domains, Fusion prioritizes basic number combinations, place value concepts, multi-digit computation without regrouping, and word problem solving.

At its core, Fusion carefully integrates foundational concepts and skills of whole number, and validated-design principles of explicit and systematic instruction (Baker et al., 2002; Coyne, Kame'enui, & Carnine, 2011; Doabler et al., 2015; Doabler, Strand Cary, et al., 2012). Specifically, instruction, academic feedback (e.g., feedback including error correction), and review are carefully integrated into the curriculum (see Clarke et al., 2009 for details). Each lesson contains specifications for interventionists to (a) model what students are to learn, (b) assist students as they work as a group or individually through scaffolded instructional examples, (c) facilitate opportunities for students to engage in mathematical discourse, and (d) provide specific academic feedback (e.g., including error corrections and reasons for correct response) to students during the mathematics activities. For example, when a new mathematical concept is introduced, the lesson scripting offers guidelines for interventionists on how to model the concept and provide a clear explanation of its relevance to whole number understanding. Through individual and group practice opportunities, students' progress from scaffolded to unscaffolded independent practice and problem solving, all with immediate corrective feedback.

Other central features of Fusion include the facilitation of student mathematics verbalizations and incorporation of visual representations of mathematics ideas. Recent research supports the idea that mathematics verbalizations are critical for supporting students' mathematical proficiency (Doabler et al., 2015; Gersten et al., 2009). For students struggling with mathematics, teacher-facilitated mathematics verbalizations permit a structured opportunity to communicate mathematical understanding, thinking, and reasoning. With respect to visual representations of mathematics ideas, Fusion incorporates a variety of mathematics representations, including number lines, strip diagrams, and place value blocks, to build a deep understanding of whole numbers and operations. These representations are scaffolded across concepts and systematically withdrawn to promote abstract mathematical thinking. Detailed scripting supports interventionists' clear and systematic introduction of new and complex whole number concepts and skills, bolsters fidelity of implementation, and is intended to increase the quantity and quality of instructional interactions between teachers and students around whole number concepts and skills.

Fusion implementation guidelines. Implementation was expected to occur after PD and pretesting were complete (i.e., mid-October) and last until all 60 Fusion lessons were completed or mid-May, whichever came first. The research team requested that interventionists provide Fusion instruction to at-risk students assigned to Fusion once per day, three times per week. Lessons were expected to be approximately 30 min and to be delivered in small-group instructional formats, with four to five students per group to match standard recommendations for small-group math instruction (Gersten et al., 2009). Instruction was to occur at times that did not interfere with students' core mathematics instruction. Actual implementation timing and approaches were managed by each school. Project resources and educator involvement and commitment were not sufficient to formally document program implementation; thus, fidelity of implementation is not a variable in our analysis.

Control. Control students were expected to participate in standard core mathematics instruction. We requested that control students not receive Fusion instruction, though we did not discourage schools from offering them standard district intervention services. Schools reported no such services were provided to control students.

School curricula and interventions. School staff reported using five core curricula across participating schools. Everyday Math (McGraw Hill Education) was used by four schools as a core program and as a supplement to the core by an additional school, whereas Investigations (Pearson), Saxon Math (Houghton Mifflin Harcourt), Math Expressions (Houghton Mifflin Harcourt), and Go Math (Houghton Mifflin Harcourt) were used by the remaining schools. No schools offered first-grade math interventions.

PD. Four PD sessions were provided to Fusion interventionists and coordinators. PD primarily focused on (a) the research-based principles of mathematics instruction underlying the Fusion intervention; (b) the instructional design and delivery features of Fusion (with a particular emphasis on clear and consistent communication, immediate academic feedback, and the importance of many opportunities to respond); (c) an overview of Fusion lessons; and (d) small-group management techniques. Evaluation design and data collection were addressed as well. The first two hands-on sessions occurred in the district before the study began and were led by Fusion authors (also members of the research team). In August 2011, authors provided an overview of the Fusion intervention within the context of a larger discussion of RTI frameworks and evidence-based practices recommended in the IES practice guide (Gersten et al., 2009). Fusion excerpts and practices were highlighted during the discussion to make concepts concrete and applicable to the attendees. Attendees also learned about the evaluation's parameters (e.g., responsibilities, calendar, student identification procedures) and practiced administering

the EN-CBM screening assessment. During the September 2011 PD, authors reiterated the reason behind early intervention and provided a walk-through of the first half of the Fusion curriculum (Book 1: Lessons 1-30), along with instructional demonstrations. Attendees practiced delivering instruction, received training regarding the ProFusion assessment, discussed implementation guidelines and recommendations, and reviewed evaluation parameters. Pretesting and instruction began in October 2011. The third in-service took place in the district in December 2011 and was again led by the authoring team. Educators had an opportunity to provide feedback and ask questions about the Fusion intervention, further discuss guidelines and recommendations (especially those related to group management and troubleshooting absences and schedule changes), walk through Book 2 (Lessons 31-60), and practice providing Fusion instruction. Recommendations and tips for success were a focus of the training, and attendees were encouraged to follow up with authors and research staff with questions and support, as needed. In the spring, the research team held a fourth PD session—a webinar to train coordinators and interventionists in SAT-10 assessment procedures.

Data collection. All first-grade students completed the easyCBM as part of standard district practices. The beginning of year scores served as the initial screener, and participating students' end-of-year scores served as a posttest measure. Coordinators and interventionists administered the remaining measures. Participating students completed the EN-CBM and ProFusion prior to receiving Fusion instruction and again after Fusion instruction ended. Participants also completed the SAT-10 at posttest. Data were securely transferred to the University of Oregon research team for processing and scoring. Hand-scored protocols were double-scored and double entered by two staff members to ensure accuracy. Discrepancies were flagged and corrected. In all, 20% of machine-scored (i.e., Teleform) results were hand-scored for comparison to confirm scanning. Discrepancy rates were in acceptable ranges. An extensive list of de-identified student demographics (e.g., age, ethnicity, free and reduced lunch status, special education status) were requested from the district but could not be reliably obtained for all students.

The research team attempted to collect implementation details and educator perceptions of Fusion through end-of-year surveys. The research team developed online Qualtrics surveys suitable for each educator role (e.g., coordinator, interventionist, classroom teacher) along with instructions regarding who should complete which survey (because some educators served more than one role). Because the research team was not in direct communication with anyone but coordinators, the coordinators distributed survey links to adult stakeholders (i.e., coordinators, interventionists, and classroom teachers). As noted above, educators were not formal research participants nor were they the focus of the evaluation. Thus, survey completion was not only optional but also anonymous. Response rates were low, and a review of survey responses strongly suggest that surveys were not always completed by the intended respondent-type. For these reasons, information gleaned from the surveys must be considered anecdotal, at best.

Statistical analysis. The analysis used two different approaches depending on the availability of pretest data. For measures available at pretest and posttest, we conducted a time-by-condition analysis, which tests the difference between conditions in the net gains for students. For assessments available only at posttest, we conducted an analysis of covariance (ANCOVA). Both analyses accounted for student membership within schools. The analysis nested students in an instructional group (Fusion instruction or regular instruction) within each school.

Because the lowest performing students were generally assigned to Fusion, the two experimental groups were different at pretest. Thus, a time-by-condition or gain-score analysis was the most appropriate analytical approach. We used this model to test condition differences with ProFusion, EN-CBM, and easyCBM. This analysis tests "whether the two groups differ in terms of their mean change over time" (Fitzmaurice, Laird, & Ware, 2004, p. 124). This model also includes students whether or not they had data at both time points, reducing bias associated with missingness (Graham, 2009). With 20 clusters (10 schools × 2 conditions), the analyses used 17 degrees of freedom.

The SAT-10 problem-solving and procedures measures were available only at posttest, so we analyzed these data with a mixed-model ANCOVA using ProFusion at pretest as the covariate. The ANCOVA contrasts residualized outcomes scores, nested within instructional groups, between the intervention and control conditions and "addresses the question of whether an individual belonging to one [condition] is expected to change more (or less) than an individual belonging to the other [condition], given that they have the same baseline response" (Fitzmaurice et al., 2004, p. 124, emphasis in original). Technically, because students cannot have the same baseline response due to their assignment to condition by pretest scores, this analysis answers an invalid question. Nonetheless, without pretest data, other options were unavailable. We therefore interpret the results of these analyses cautiously.

We fit models to our data with SAS PROC MIXED version 9.2 (SAS Institute Inc., 2009) using restricted maximum likelihood. The models assume independent and normally distributed observations. We addressed the first assumption (Van Belle, 2008) by explicitly modeling the multilevel nature of the data. Murray et al. (2006) showed that violations of normality at either or both the individual and group levels are not likely to bias results. To ease

Table 1. Descriptive Information for Mathematics Measures by Time and Condition.

	Pretest		Posttest	
Measure	Fusion	Comparison	Fusion	Comparison
ProFusion				
M	23.24	27.20	59.71	53.29
SD	7.34	10.15	16.11	13.36
N	139	70	143	80
EN-CBM				
M	162.15	172.21	224.02	222.13
SD	40.17	39.78	39.85	37.15
N	129	77	126	80
easyCBM				
M	19.27	22.88	27.09	31.79
SD	12.57	12.45	16.85	15.30
N	146	92	132	89
SAT-10 problem solving				
M			25.63	26.28
SD			7.33	6.91
N			134	96
SAT-10 procedures				
M			18.41	18.31
SD			5.95	5.76
N			130	94

Note. EN-CBM = Early Numeracy Curriculum Based Measures; SAT-10 = Stanford Achievement Test–Tenth Edition.

interpretation, we computed an effect size, Hedges' *g* (Hedges, 1981), for each fixed effect according to the What Works Clearinghouse (WWC, 2014) standards. Hedges' *g* is comparable to Cohen's *d* (Cohen, 1988). Both represent individual-level effect sizes.

Results

Baseline Equivalence and Attrition

Due to the recommended screening and assignment process, we expected students in the Fusion condition to score lower on the pretest measures than comparison students. This was indeed the case (Table 1), but there were no statistically significant differences at pretest by condition for ProFusion ($p = .0778$), EN-CBM ($p = .0789$), or easyCBM ($p = .5739$). See the Condition row of Table 2 for additional details. Given that each school was responsible for screening and assigning students to intervention, it may be that there was greater-than-expected heterogeneity within conditions and less between-condition-heterogeneity than the research team would have anticipated. If anything, though, this makes the test of Fusion's effects more rigorous.

Student attrition was defined as students with data at pretest but missing data at posttest. Approximately 28% of the 253 students were missing one or more posttest measures, but only 3% were missing all posttest measures. The rate of attrition did not differ between conditions for students missing one or more measures ($\chi^2 = 0.25$, $df = 1$, $p = .9604$) but did for those missing all measures ($\chi^2 = 5.31$, $df = 1$, $p = .0212$). Among the 99 comparison students, none were missing all measures, while eight students in the intervention condition had no data at posttest.

Although differential rates of attrition are undesirable, differential scores by condition present a greater threat to validity (Barry, 2005). To test whether student scores were differentially affected by attrition across conditions, we examined the effects of condition, attrition status, and the interaction between the two on pretest scores within a mixed-model analysis of variance (Murray, 1998) that nested students' pretest scores within schools and condition. We found no evidence of differential attrition for any of our dependent variables: $p > .50$ for all tests.

Intervention Effects for Fusion

Results of the time by condition analyses (Tables 2 and 3) suggest a positive, statistically significant effect for the proximal ProFusion measure ($p = .046$, $g = .61$). Positive nonsignificant effects were found on the distal EN-CBM ($p = .13$, $g = .34$) and SAT-10 procedures ($p = .71$, $g = .13$) measures. Negative nonsignificant results were found on the distal easyCBM ($p = .49$, $g = -.21$) and SAT-10 problem solving ($p = .51$, $g = -.27$) measures.

Enacted Fusion Intervention

Survey responses collected at the end of the year ($n = 14$) reveal that the district's implementation of Fusion generally matched the research team's recommendations, but that variation existed. Fusion instruction occurred between October and May and, with a few exceptions, involved 30-min lessons outside of regular mathematics and reading instruction. Most Fusion groups had five members (range: 4-10 students) and met 3 days per week (range: 2-5 days/week). With the exception of one site that offered Fusion after school, instruction generally took place during school hours. Most interventionists taught a single group, but there were exceptions (including two who taught four groups each). The majority of groups covered through at least lesson 50 (of 60 possible lessons). The variation in fidelity is unfortunate, though certainly not unexpected. As Harwell (2012) points out, "Assuring treatment fidelity often requires significant support during implementation" and that was not a part of this study.

Perceptions of the Fusion Curriculum

We sought coordinator, interventionist, and classroom teacher perceptions of the Fusion program. Because a more

Table 2. Results From Mixed-Model Time × Condition Analysis of Condition Effects on Fall-to-Spring Gains in Math Measures.

Effect or statistic	ProFusion	EN-CBM	easyCBM
Fixed effects			
Intercept	30.86** (3.35)	174.44** (7.50)	23.66** (3.32)
Time	24.78** (3.30)	50.31** (6.12)	8.85* (3.55)
Condition	−8.24† (4.39)	−18.82† (10.07)	−2.61 (4.52)
Time × Condition	9.29* (4.32)	13.03 (8.23)	−3.39 (4.84)
Variances			
Cluster intercept	34.17 (25.26)	203.10 (131.39)	40.25 (27.43)
Cluster gains	31.31* (14.35)	90.01 (58.90)	41.04* (17.95)
Student	50.58** (8.14)	805.13** (109.65)	58.30** (11.47)
Residual	49.87** (5.20)	491.68** (52.16)	94.78** (9.56)
ICC	.386	.155	.302
Hedges' *g*			
Time × Condition	.612	.336	−.209
p-value			
Time × Condition	.0463	.1319	.4932

Note. Table entries show parameter estimates with standard errors in parentheses except for intraclass correlations (ICCs), Hedges' *g* values, and *p* values. Tests of fixed effects (first four rows) used 17 *df* to account for the instructional group as the unit of analysis. ICCs calculated as per Murray (1998, p. 301). EN-CBM = Early Numeracy Curriculum Based Measures; ICC = intraclass correlation.
†$p < .10$. *$p < .05$. **$p < .0001$.

Table 3. Results From Mixed-Model Analysis of Covariance for Condition Effects on Fall-to-Spring Gains in SAT-10 Math Measures.

Effect or statistic	SAT-10 problem solving	SAT-10 procedures
Fixed effects		
Intercept	19.52** (2.31)	12.41** (1.87)
Condition	−1.74 (2.59)	0.71 (1.87)
Pretest ProFusion	0.30** (0.04)	0.23** (0.04)
Variances		
Cluster intercept	23.25* (9.46)	10.94* (5.25)
Residual	19.11** (2.03)	19.03** (2.06)
ICC	.549	.365
Hedges' g		
Condition	−.267	.131
p value		
Condition	.5107	.7107

Note. Table entries show parameter estimates with standard errors in parentheses except for ICCs, Hedges' *g* values, and *p* values. Tests of fixed effects (first four rows) used 18 *df* to account for the instructional group as the unit of analysis. SAT-10 = Stanford Achievement Test–Tenth Edition; ICC = intraclass correlation.
†$p < .10$. *$p < .05$. **$p < .0001$.

refined analysis and presentation is precluded by our limited access to school staff; receipt of de-identified data; the dual roles played by coordinators, teachers, and interventionists; and respondents completing surveys that were not intended for their role, a qualitative snapshot is provided. First and foremost, schools differed in how they chose to implement Fusion and with whom. For example, one school conducted Fusion after school. Another assigned Fusion to all students who did not meet benchmark on the easyCBM. Some staff slated as Fusion coordinators served as interventionists as well, whereas in others, these roles were separate. Generally, students assigned to intervention were not those the research team would have recommended, given screening data alone.

Participants commented that Fusion's clear, concise, and detailed lesson plans mean that "Fusion can be taught by many different types of teachers," and "any education professional can deliver Fusion." "The variations and number of tasks per session kept the pace moving which was important to keep the students on task" but, at the same time, it is "sometimes difficult [to] catch a teachable moment when it occurs." As would be expected with a new curriculum, some respondents noted that, "Fusion is the kind of program that you get better at the more you do it," and "It became easier to follow as I became familiar with the routine."

Interventionists recommended some formatting changes to the instructor materials to make them more usable during the lesson, additional PD and planning time, and that more than 30 min be allocated to lessons. Interventionists gave mixed reviews to specific strategies. Some respondents noted students sometimes blended or confused strategies taught in Fusion with those taught in class. Others noted the strategies were a major reason the students became more successful in math. Logistical/resource concerns and consequences (e.g., lack of time, lack of personnel, Fusion participation coming at the expense of other important services) were the biggest obstacles noted by respondents. A number of respondents commented that students' attitude and engagement with mathematics in the core classroom had improved (e.g., confidence, participation, talkativeness,

enjoyment). Across all types of respondents, most noted they would be likely or very likely to recommend the intervention.

Discussion

Both this study and the concurrent formal pilot study (Clarke et al., 2014) revealed significant positive impacts on proximal math measures. This is important because it indicates the potential promise for Fusion as a Tier 2 intervention program whether implemented by school practitioners with limited Fusion training and ongoing support or by interventionists selected, trained, and highly supported by Fusion developers. Specific to this study, significant student performance gains were made on the ProFusion assessment, and nonsignificant, positive gains were made on the EN-CBM and the SAT-10. Educator survey responses indicate gains in student knowledge, confidence, and interest in math instruction during Fusion lessons, and during core instruction as well.

With moderate PD before and early on in implementation, but without ongoing onsite coaching and support from program developers, interventionists implemented Fusion lessons and had positive impressions of their experience as instructors, the intervention itself, and student growth. Feedback from coordinators and classroom teachers was equally positive. They believed lessons were comprehensive and implementable with little training, though some participants felt additional PD, additional instructional time, and some formatting changes would improve the program. These perceptions mirrored those in the concurrent formal pilot study (Clarke et al., 2014).

This practitioner-driven study provides a glimpse into the effects of a research-based first-grade math intervention, Fusion, when enacted under authentic conditions. It is an example of researchers maximizing grant dollars and practitioner interest in emerging curricula by supplementing planned iterative development work and formal studies with hands off studies in schools. Although less rigorous, such studies give developers the opportunity to see what works (or does not work) in a wider variety of contexts and with a wider variety of students than would otherwise be possible during a typical development grant. These studies offer practitioners opportunities to answer their own important questions about new programs, for example, (a) can school staff implement the program? (b) do students learn from the program? and (c) do educators and students enjoy the program? Such district-initiated evaluations—regardless of rigor or interest to the broader education community—are extremely rare.

Limitations

This study involved a relatively small sample in a unique geographical and sociological area, and thus may not generalize to other populations and settings. There was not a distinct effort to include (or exclude) students speaking English as a second language or students with disabilities, and thus, we cannot speak to subgroup effects of this implementation (though we plan to explore those questions as part of the current Fusion efficacy trial).

The nature of authentic implementations is that school, student, and educator needs and priorities play a large role in implementation and tend to differ from those of a research entity. Consequently, this informal, external evaluation conducted from afar brought with it a host of challenges and limitations. Thus, in this manuscript, we simply present what we—as program developers and evaluators—prescribed and what the schools reported. Even we would be hard-pressed to replicate this study exactly as conducted given our lack of control and knowledge of study details.

The most obvious limitations are that the study was not an RCT, and implementation varied from site to site. This was an opportunistic evaluation. Conducting an RCT or other rigorously designed study was not an option. Similarly, although the research team encouraged adherence to an implementation protocol, schools and individual educators determined where and how Fusion would be implemented. Relatedly, the research team had no control—and little knowledge—of implementation fidelity. Project and school resources were not sufficient to conduct observations of Fusion instruction or standard district practice, collect teaching experience data, or maintain direct communication with interventionists, maintain instructional logs, or provide coaching. Despite efforts to track implementation specifics, the unknowns remain unknown. Similarly, we cannot attest to the procedural integrity of assessments. Instead, we must assume that any variations over time or across assessors similarly affect intervention and control students' scores.

Despite our lack of control and statistical power, we hoped to meaningfully explore effects for subgroups of students and effects of particular implementation practices (e.g., intensity, duration, group size, interventionist characteristics) once we had a complete data set including student performance data, student demographic data, and educator self-reports. Given that the expected complete data set did not materialize, we instead extracted what meaning we could from the data we had (as reported in the "Results" section). By quantitatively exploring Fusion's effects on student outcomes, we found that Fusion showed potential promise.

This litany of limitations is presented for full disclosure, to emphasize what other researchers have noted about the specific reasons why evaluation in schools is difficult (e.g., see Harwell, 2012 for a review), and to remind the reader that such studies still have something to offer within a broader spectrum of research. The study occurred as the field more explicitly recognized that to bridge the research to practice gap, greater engagement with practitioners was needed (Whitehurst, 2010) and that there is value in

cumulatively gathering evidence for practices or programs by conducting studies ranging from small-scale nonexperimental to large-scale evaluation trials. Indeed, as more developers and researchers try to close the research-to-practice gap, we anticipate more studies of this nature will appear—not in isolation but as pieces of larger research programs centered on particular curricula or programs. They will provide valuable data for meta-analyses or serve as miniature scale-up studies informative in and of themselves. Taking all the evidence together and looking for patterns and areas for further investigation will maximize the usefulness of the entire research endeavor.

Evidence of this direction includes attempts to codify researcher district partnerships. IES has begun a grant program focused on conducting low-cost RCT studies to enable districts to evaluate potential practices (IES, 2016a). The guidelines call for attending to five key attributes when conducting a low-cost RCT, including selecting an important practice, partnering with a local or state agency, the use of a rigorous design (including RCT studies, regression discontinuity, or single case), the use of administrative or secondary data sources, and a timely evaluation and dissemination of results with the participating districts. The attributes could serve as an organizing source for researchers to partner with districts and operationalize expectations prior to entering into partnerships. For example, in the research presented, greater clarity on roles and expectations for each partner would have enabled a more rigorous estimate of Fusion's impact on student mathematics outcomes.

Implications for Practice

Not every evaluation undertaken by a district will be with a formal research partner and meet design criteria that enable a rigorous evaluation. However, by engaging in various forms of evaluations, districts and practitioners can develop an understanding of the levels of evidence offered by different evaluations. The study presented here represents work to "determine whether they [interventions] produce a beneficial impact on student education outcomes relative to a counterfactual when they are implemented by the end user under routine conditions in authentic education settings" (IES, 2016b). Specific to this study, there is a growing body of research that early math intervention programs designed with a focus on whole number concepts and attending to key instructional design principles can positively impact student achievement (e.g., Dyson, Jordan, & Glutting, 2013; Fuchs et al., 2005; Sood & Jitendra, 2013) and that educators can deploy within multi-tier systems of support. More generally, there is a movement toward an environment in which policy makers and districts can evaluate whether innovations and practices work within varied contexts. Doing so will help ensure that districts can make informed decisions about whether and how to implement programs with their students as they work to meet the learning needs of all of their students.

Acknowledgments

The authors would like to acknowledge the educators and students at participating schools and the research team at the University of Oregon's Center on Teaching and Learning who helped make this research possible. Thanks to the reviewers and editors of this journal for helpful feedback during the review process.

Declaration of Conflicting Interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: Ben Clarke, Christian T. Doabler, Hank Fien, and Scott K. Baker are eligible to receive a portion of royalties from the University of Oregon's distribution and licensing of certain Fusion-based works. Potential conflicts of interest are managed through the University of Oregon's Research Compliance Services. Keith Smolkowski served as independent external evaluator and completed the research analysis described in the article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: 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.

References

Agodini, R., & Harris, B. (2010). An experimental evaluation of four elementary school math curricula. *Journal of Research on Educational Effectiveness*, *3*, 199–253. doi:10.1080/19345741003770693

Alonzo, J., Tindal, G., Ulmer, K., & Glasgow, A. (2006). *easyCBM online progress monitoring assessment system*. Eugene: University of Oregon. Available from http://easycbm.com

Anderson, D., Alonzo, J., & Tindal, G. (2010). easyCBM® mathematics criterion related validity evidence: Oregon state test (Technical Report No. 1011). Eugene: Behavioral Research and Training, University of Oregon. Retrieved from http://www.brtprojects.org/publications/dl/1014

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

Barnett, W. S., Jung, K., Frede, E. C., Hustedt, J., & Howes, C. (2011). *Effects of eight state prekindergarten programs on early learning*. New Brunswick, NJ: National Institute for Early Education Research, Rutgers University.

Barry, A. E. (2005). How attrition impacts the internal and external validity of longitudinal research. *Journal of School Health*, *75*, 267–270. doi:10.1111/j.1746-1561.2005.tb06687.x

Bryant, D. P., Bryant, B. R., Gersten, R. M., Scammacca, N. N., & Chavez, M. M. (2008). Mathematics intervention for first- and second-grade students with mathematics difficulties: The effects of Tier 2 intervention delivered as

booster lessons. *Remedial and Special Education*, *29*, 20–32. doi:10.1177/0741932507309712

Bryant, D. P., Bryant, B. R., Gersten, R. M., Scammacca, N. N., Funk, C., Winter, A., . . . Pool, C. (2008). The effects of Tier 2 intervention on the mathematics performance of first-grade students who are at risk for mathematics difficulties. *Learning Disability Quarterly*, *31*, 47–63. doi:10.2307/20528817

Bryant, D. P., Bryant, B. R., Roberts, G., Vaughn, S., Pfannenstiel, K. H., Porterfield, J., & Gersten, R. M. (2011). Early numeracy intervention program for first-grade students with mathematics difficulties. *Exceptional Children*, *78*, 7–23. doi:10.1177/001440291107800101

Bryant, D. P., Pfannenstiel, K. H., & Bryant, B. R. (2014). *Early numeracy intervention program*. Austin, TX: Psycho-Educational Services.

Clarke, B., Baker, S. K., & Fien, H. (2009). *Foundations of mathematical understanding: Developing a strategic intervention on whole number concepts* (FUSION). (Institute of Education Sciences, Mathematics and Science Education: Special Education Research, CFDA Num: 84.324A, 2009-2012, Funding Number: R324A090341, awarded $1,455,851). Retrieved from https://ies.ed.gov/funding/grantsearch/details.asp?ID=770

Clarke, B., Doabler, C. T., Fien, H., & Smolkowski, K. (2016). *A randomized control trial of a Tier 2 first grade mathematics intervention* (Institute of Education Sciences [IES]: Special Education Research. NCSER-Mathematics, Efficacy and Replication, Goal 3, CFDA Num: 84.324, 2016-2020, Funding Number: R324A160046, awarded $3,498,258). Retrieved from https://ies.ed.gov/funding/grantsearch/details.asp?ID=1815

Clarke, B., Doabler, C. T., Strand Cary, M., Kosty, D. B., Baker, S. K., Fien, H., & Smolkowski, K. (2014). Preliminary evaluation of a Tier-2 mathematics intervention for first grade students: Utilizing a theory of change to guide formative evaluation activities. *School Psychology Review*, *43*, 160–177.

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. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=pbh&AN=13907070&site=ehost-live

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

Clements, D. H., Agodini, R., & Harris, B. (2013, September). *Instructional practices and student math achievement: Correlations from a study of math curricula* (NCEE Evaluation Brief). Washington, DC: National Center for Educational Evaluation and Regional Assistance, Institute of Education Sciences. Retrieved from http://ies.ed.gov/ncee/pubs/20134020/pdf/20134020.pdf

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York, NY: Academic Press.

Common Core State Standards Initiative. (2010). *Common core standards for mathematics*. Retrieved from http://www.corestandards.org/the-standards/mathematics

Cortiella, C., & Horowitz, S. H. (2014). *The state of learning disabilities. Facts, trends and emerging issues*. New York, NY: National Center for Learning Disabilities. Retrieved from http://www.ncld.org/wp-content/uploads/2014/11/2014-State-of-LD.pdf

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

Dewa, C. S., Durbin, J., Wasylenki, D., Ochocka, J., Eastabrook, S., Boydell, K. M., & Goering, P. (2002). Considering a multisite study? How to take the leap and have a soft landing. *Journal of Community Psychology*, *30*, 173–187. doi:10.1002/jcop.10001

Doabler, C. T., Baker, S. K., Kosty, D. B., Smolkowski, K., Clarke, B., Miller, S. J., & Fien, H. (2015). Examining the association between explicit mathematics instruction and student mathematics achievement. *The Elementary School Journal*, *115*, 303–333. doi:10.1086/679969

Doabler, C. T., Fien, H., Nelson, 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. doi:10.1177/0731948712438557

Doabler, C. T., Strand Cary, M., Jungjohann, K., Clarke, B., Fien, H., Baker, S. K., . . . Chard, D. J. (2012). Enhancing core mathematics instruction for students at risk for mathematics disabilities. *Teaching Exceptional Children*, *44*(4), 48–57. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=aph&AN=72081580&site=ehost-live&scope=site

Dyson, N. I., Jordan, N. C., & Glutting, J. (2013). A number sense intervention for low-income kindergartners at risk for mathematics difficulties. *Journal of Learning Disabilities*, *46*, 166–181. doi:10.1177/0022219411410233

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2004). *Applied longitudinal analysis*. Hoboken, NJ: Wiley. Retrieved from http://www.loc.gov/catdir/toc/wiley041/2004040891.html

Fuchs, L. S., Compton, D. L., Fuchs, D., Paulsen, K., Bryant, J. D., & Hamlett, C. L. (2005). The prevention, identification, and cognitive determinants of math difficulty. *Journal of Educational Psychology*, *97*, 493–513. doi:10.1037/0022-0663.97.3.493

Fuchs, L. S., & Vaughn, S. (2012). Responsiveness-to-intervention: A decade later. *Journal of Learning Disabilities*, *45*, 195–203. doi:10.1177/0022219412442150

Gersten, R. M., Beckmann, S., Clarke, B., Foegen, A., Marsh, L., Star, J. R., & Witzel, B. (2009). *Assisting students struggling with mathematics: Response to intervention (RtI) for elementary and middle schools* (Report No. NCEE 2009-4060). Princeton, NJ: What Works Clearinghouse. Retrieved from http://www.eric.ed.gov/ERICWebPortal/detail?accno=ED504995

Gersten, R. M., Rolfhus, E., Clarke, B., Decker, L. E., Wilkins, C., & Dimino, J. (2015). Intervention for first graders with limited number knowledge: Large-scale replication of a randomized controlled trial. *American Educational Research Journal*, *52*, 516–546. doi:10.3102/0002831214565787

Ginsburg, H. P., & Baroody, A. J. (2003). *Test of Early Mathematics Ability–Third Edition (TEMA-3)*. Austin, TX: Pro-Ed.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, *60*, 549–576. doi:10.1146/annurev.psych.58.110405.085530

Harcourt Educational Measurement. (2002). *Stanford Achievement Test–Tenth Edition (SAT-10)*. San Antonio, TX: Author.

Harwell, M. (2012). Multisite studies and scaling up in educational research. *Educational Research Quarterly*, *36*(2), 20–41.

Hedges, L. V. (1981). Distribution theory for glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*, 107–128. doi:10.3102/10769986006002107

Institute of Education Sciences. (2016a). *Program announcement: Low-cost, short duration evaluation of education interventions* (CFDA 84.305L). Washington, DC: Institute of Education Sciences, National Center for Education Sciences, U.S. Department of Education. Retrieved from http://ies.ed.gov/funding/ncer_rfas/ncer_lcsd.asp

Institute of Education Sciences. (2016b). *Request for applications: Education research grants* (CFDA No. 84.305A). Washington, DC: U.S. Department of Education.

Jordan, N. C., & Levine, S. C. (2009). Socioeconomic variation, number competence, and mathematics learning difficulties in young children. *Developmental Disabilities Research Reviews*, *15*, 60–68. doi:10.1002/ddrr.46

Maltese, A. V., & Hochbein, C. D. (2012). The consequences of "school improvement": Examining the association between two standardized assessments measuring school improvement and student science achievement. *Journal of Research in Science Teaching*, *49*, 804–830. doi:10.1002/tea.21027

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. doi:10.1177/0022219408331037

Murray, D. M. (1998). *Design and analysis of group-randomized trials*. New York, NY: Oxford University Press.

Murray, D. M., Hannan, P. J., Pals, S. P., McCowen, R. G., Baker, W. L., & Blitstein, J. L. (2006). A comparison of permutation and mixed-model regression methods for the analysis of simulated data in the context of a group-randomized trial. *Statistics in Medicine*, *25*, 375–388.

National Association of State Directors of Special Education. (2006). *Response to intervention: A joint paper by the National Association of State Directors of Special Education and the Council of Administrators of Special Education* (Joint paper). Washington, DC: National Association of State Directors of Special Education, Council of Administrators of Special Education.

National Center for Education Statistics. (2013). *The nation's record card: A first look: 2013 mathematics and reading* (No. NCES 2014-451). Washington, DC: National Center for Education Statistics, Institute of Education Sciences.

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. (2009). *Mathematics learning in early childhood: Paths toward excellence and equity*. Washington, DC: National Academies Press. Retrieved from http://www.nap.edu/catalog.php?record_id=12519

Organisation for Economic Co-Operation and Development. (2012). *PISA 2012: What students know and can do—Student performance in mathematics, reading and science. Snapshot of performance in mathematics, reading and science*. Paris, France: Author. Retrieved from http://www.oecd.org/pisa/keyfindings/PISA-2012-results-snapshot-Volume-I-ENG.pdf

Riccomini, P. J., & Smith, G. W. (2011). Introduction of response to intervention in mathematics. In R. M. Gersten & R. Newman-Gonchar (Eds.), *Understanding RTI in mathematics: Proven methods and applications* (pp. 1-16). Baltimore, MD: Brookes Publishing.

SAS Institute Inc. (2009). *SAS/STAT ®9.2 user's guide*. Cary, NC: Author.

Seethaler, P. M., & Fuchs, L. S. (2011). Using curriculum-based measurement to monitor kindergarteners' mathematics development. *Assessment for Effective Intervention*, *36*, 219–229. doi:10.1177/1534508411413566

Simmons, D. C., Coyne, M. D., Hagan-Burke, S., Kwok, O.-M., Johnson, C., Zuo, Y., . . . Crevecoeur, Y. (2011). Effects of supplemental reading interventions in authentic contexts: A comparison of kindergarteners' response. *Exceptional Children*, *77*, 207–228. Retrieved from http://journals.sagepub.com/doi/abs/10.1177/001440291107700204

Smolkowski, K., Strycker, L., & Seeley, J. R. (2013). The role of research in evaluation of interventions for school-related behavioral disorders. In H. M. Walker & F. M. Gresham (Eds.), *Handbook of evidence-based practices for emotional and behavioral disorders: Applications in schools* (1st ed., pp. 552-566). New York, NY: Guilford Press.

Sood, S., & Jitendra, A. K. (2013). An exploratory study of a number sense program to develop kindergarten students' number proficiency. *Journal of Learning Disabilities*, *46*, 328–346. doi:10.1177/0022219411422380

Van Belle, G. (2008). *Statistical rules of thumb* (2nd ed.). Hoboken, NJ: Jon Wiley.

What Works Clearinghouse. (2014, June 13). *Procedures and standards handbook* (Version 3.0). Washington, DC: Institute of Education Sciences. Retrieved from http://ies.ed.gov/ncee/wwc/DocumentSum.aspx?sid=19

Whitehurst, G. J. (2010). *Education research: Past, present, and future*. Los Alamitos, CA: WestEd. Retrieved from https://www.wested.org/online_pubs/pp-10-01.pdf